Neurocomputing 562 (2023) 126867

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

FABSA: An aspect-based sentiment analysis dataset of user reviews

Georgios Kontonatsios *, Jordan Clive, Georgia Harrison, Thomas Metcalfe, Patrycja Sliwiak, Hassan Tahir, Aji Ghose

Chattermill, London, United Kingdom

ARTICLE INFO

Communicated by D. Cavaliere

Keywords:
ABSA
Multi-domain dataset
Deep learning

ABSTRACT

Aspect-based sentiment analysis (ABSA) aims at automatically extracting aspects of entities and classifying the polarity of each extracted aspect. The majority of available ABSA systems heavily rely on manually annotated datasets to train supervised machine learning models. However, the development of such manually curated datasets is a labour-intensive process and therefore existing ABSA datasets cover only a few domains and they are limited in size. In response, we present FABSA (Feedback ABSA), a new large-scale and multi-domain ABSA dataset of feedback reviews. FABSA consists of approximately 10,500 reviews which span across 10 domains. We conduct a number of experiments to evaluate the performance of state-of-the-art deep learning models when applied to the FABSA dataset. Our results demonstrate that ABSA models can generalise across different domains when trained on our FABSA dataset while the performance of the models is enhanced when using a larger training dataset. Our FABSA dataset is publicly available.[1]

1. Introduction

Opinion mining methods have become increasingly popular in both academia and industry considering that these methods are able to extract useful insights by analysing vast amounts of user-generated data (e.g. feedback reviews) [1–3]. Early opinion mining methods [4–6] focused on a binary sentiment classification task wherein a given feedback review was classified as being either positive or negative. However, binary sentiment classification models fail to capture multiple and opposing sentiments that can be expressed within the same feedback review [7–9]. Moreover, a single feedback review may contain multiple opinion polarities towards different themes or aspects of a product. For example, the review below contains a positive sentiment towards the *price* of the product but a negative towards the *refund/return* process.

> *Price is great but it is not clear how to process a return during the quarantine period.*

To address the above-mentioned shortcomings of binary sentiment classification models, Aspect-based Sentiment Analysis (ABSA) aims at extracting more fine-grained information by identifying both the aspects and the sentiment of each aspect [7,10]. Existing ABSA systems can be broadly classified into sequence tagging [11–13] and document classification methods [14]. Sequence tagging methods identify ABSA labels at the word-level while document classification methods assign ABSA labels at the document-level. More specifically, sequence tagging approaches extract a word or a sequence of words within a review that refers to an aspect. In our example review above, a sequence tagging method extracts two aspect terms, namely *price* and *return*, and assigns a polarity label to each aspect term. Sequence tagging methods can be further classified into Aspect Opinion Pair Extraction [15], Aspect Sentiment Triplet Extraction [16] and Target Aspect Sentiment Detection [17] methods. Aspect Opinion Pair Extraction models identify tuples that consist of an aspect term and an opinion term. An aspect term refers to an attribute of a product while an opinion term is a word or sequence of words that modify the aspect term (<*price, great*>). Aspect Sentiment Triplet Extraction is an extension of Opinion Pair Extraction in that it extracts pairs of aspect and opinion terms but it also assigns a polarity label to each extracted pair (<*price, great, positive*>). Target Aspect Sentiment Detection aims at extracting aspects and sentiment labels towards multiple target entities. As an example, there are two target entities in the review below, namely *iPhone* and *Android*. Based on this, a Target Aspect Sentiment Detection method extracts two aspect/sentiment pairs: (a) <*price, negative*> which is

* Corresponding author.
E-mail address: georgios@chattermill.io (G. Kontonatsios).

[1] https://github.com/kontonag86/fabsa-dataset

https://doi.org/10.1016/j.neucom.2023.126867
Received 5 January 2023; Received in revised form 28 July 2023; Accepted 30 September 2023
Available online 6 October 2023

assigned to the *iPhone* entity and <price, positive> which is assigned to the *Android* target entity.

> *The price of an iPhone is high while the price of an Android is relatively low*

A limitation of sequence tagging approaches is that synonymous aspect terms (e.g. *price, value, fee*) need to be subsequently clustered into parent concept categories (e.g. purchase and booking) [14] to help more efficient indexing of documents (e.g. users can retrieve documents that contain not only the exact query term but also a synonym to the input query).

Document classification ABSA methods, including the work presented in this paper, follow a text classification approach to ABSA by classifying a feedback review against a predefined list of aspect labels. A potential challenge of document classification models is that the underlying list of aspect labels is usually associated with a specific domain and therefore multi-domain datasets are needed in order to allow the development of more generalised models [18].

Recently, the use of large pre-trained language models has achieved a new state-of-the-art performance on ABSA benchmark datasets [19–23]. Moreover, prior work has demonstrated that the performance of such models continuously improves when jointly trained on a mixture of different domains [18,24,25]. However, fine-tuning large pre-trained language models on downstream tasks such as ABSA requires a substantial amount of manually labelled data [20,26] which is expensive to produce for multiple domains. For this reason, existing ABSA datasets [7,17,27,28] are relatively small in size ($8,000 - 11,000$ sentences) and only cover a few domains. In this work we aim at developing effective document classification ABSA models that can generalise across different contexts and domains. To this end, we introduce a new large-scale and multi-domain ABSA dataset called FABSA (Feedback ABSA). To the best of our knowledge, FABSA is the largest manually curated ABSA dataset, consisting of approximately 10,500 feedback reviews (20,0000 sentences). FABSA is manually labelled against a predefined scheme of 12 aspect and 3 polarity labels (positive, negative and neutral). The development of FABSA enables users to train machine learning models that do not overfit to one narrow domain, and allows for the evaluation of these models across a wide range of distinct domains. The contributions that we make in this paper can be summarised as follows:

- We release a manually annotated ABSA dataset. Our aim is to foster research on aspect-based sentiment analysis by establishing a new large-scale and multi-domain benchmark dataset.
- We provide strong baseline ABSA models using large pre-trained language models.
- We report experiments that evaluate different dimensions of the ABSA models (e.g. performance over increasing number of training instances, comparison of different model architectures).

2. Literature review

Currently, there exists a number of manually labelled ABSA datasets. Table 1 shows different information about 7 existing ABSA datasets, including: (a) the publication source that introduced the dataset, (b) the name of the dataset, (c) the size (number of sentences), (d) the number of aspects and (e) the domains that it covers.

Ganu et al. (2011) [29] presented one of the earliest ABSA datasets. Their dataset consists of restaurant reviews which are collected from the Citysearch New York platform. In their work, they used an annotation scheme of 6 aspect categories to manually annotate 3,400 sentences. Moreover, they trained a regression model on the aspect and sentiment label of a review in order to predict the star rating of that review. Their results demonstrated that certain aspect categories (e.g. *Food*) show a stronger correlation to the star rating of the review when compared to other aspects (e.g. *Service*).

The SemEval-2014 Task 4 (Pontiki et al. (2014) [30]) proved to be a popular shared task on ABSA attracting a large number of teams and submissions. The organisers published an ABSA dataset which is annotated with both term-level (Aspect Term Extraction) and document-level (Aspect Concept Extraction) aspect labels. Their dataset includes the same Restaurant reviews which were previously presented in [29] but augments the dataset with additional reviews concerning Laptops. The SemEval-2016 Task 5 (Pontiki et al. (2016) [7]) was a follow-up shared task to SemEval-2014 which released several ABSA datasets in 8 different languages. The English dataset consists of 900 reviews (5,801 sentences) relevant to Restaurants and Laptops. SemEval-2016 introduced text-level aspect annotations which assign aspect labels to the whole review rather than to individual sentences as in SemEval-2014. In our work, we adopt the same text-level annotation scheme to manually label the FABSA dataset.

Sentihood (Saeidi et al. (2016) [17]) is another ABSA dataset which includes reviews about different neighbourhoods of the city of London. Sentihood was constructed from the Yahoo! question-and-answering platform. It is unique to other existing ABSA datasets in that it assumes that a review may refer to multiple distinct entities (neighbourhoods) and that each entity is tagged with a different aspect category. Consider the following example: "*location_1 is more expensive than location_2*". Here, the review contains two entities: *location_1*, which is associated with a negative sentiment towards the *price* aspect category (<*negative, price*>) and *location_2*, which is associated with a <*positive, price*> label.

The MAMS dataset (Jiang et al. (2019) [27]) was specifically designed to include challenging cases for automatic ABSA systems. The dataset collects 8,879 sentences which are annotated with 8 aspect categories. Each sentence in MAMS contains at least two different aspects that have opposing sentiment labels. Experimental results demonstrated that while state-of-the-art ABSA models exhibit strong performance on the SemEval-2014 dataset, they yield a substantially lower performance on the challenging MAMS dataset.

Lie et al. (2014) [31] released a manually curated Twitter dataset for target-dependent sentiment classification. The Twitter dataset contains 6,940 tweets which are annotated with sentiment labels towards 118 different target entities (product names, companies, celebrities etc.). The authors reported a high inter-annotator agreement of 82% in terms of accuracy. Moreover, they introduced an Adaptive Recursive Neural Network which takes into consideration syntactic relationships between words to more accurately classify the polarity of target entities.

AWARE (Alturaief et al. (2021) [32]) is a more recently created ABSA dataset which is based on smartphone app reviews. AWARE covers 3 different domains and uses an annotation scheme of 12 aspect categories to annotate a total number of 11,323 sentences.

Our proposed FABSA dataset is novel when compared to previously introduced ABSA datasets according to the following points. Firstly, FABSA is constructed from disparate data sources (e.g. Trustpilot and Google Play), whilst the majority of existing datasets are collected from only a single data source. Secondly, FABSA spans 10 domains, whereas previous ABSA datasets cover at most 3 domains. In practise this means that the reviews in the FABSA dataset cover more diverse contexts and topics. Thirdly, FABSA consists of approximately 20,000 sentences which is 1.7 times the size of AWARE and 2.6 times the size of the SemEval-2014 dataset.

3. Dataset construction

In this section, we describe the process that we followed to collect and manually annotate the FABSA dataset. Moreover, we compute the inter-annotator agreement (F1-score) between the three annotators that we used to develop our dataset and we analyse their disagreements across the different aspect labels. Finally, we report various dataset statistics such as the distribution of aspect labels, the distribution of sentiment labels and pairwise correlations among the aspect labels.

Table 1
Previously introduced ABSA datasets. The table includes the publication source, the size, number of aspects and domains for each dataset.

Publication source	Dataset name	Size	# aspects	Domains
Ganu et al. (2011) [29]	New York Restaurants	3400 sentences	6	Restaurants
Pontiki et al. (2014) [30]	SemEval-2014	7686 sentences	6	Restaurants, Laptops
Pontiki et al. (2016) [7]	SemEval-2016	5801 sentences	6	Restaurants, Laptops
Saeidi et al. (2016) [17]	SentiHood	5215 sentences	11	Neighbourhoods
Jiang et al. (2019) [27]	MAMS	8879 sentences	8	Restaurants
Lie et al. (2014) [31]	Twitter dataset	6940 tweets	118 (entities)	General domain
Alturaief et al. (2021) [32]	AWARE	11,323 sentences	12	Productivity, Social Networking, Games

Table 2
Size of each data source in the FABSA dataset in terms of number of reviews, sentences, unique orgs and uniqu industries that they contain.

	# reviews	# sentences	# unique orgs.	# unique industries
Trustpilot	1920	5174	3	3
Google Play	6130	9656	14	10
Apple App Store	2524	5716	12	10
Total	10,574	20,546	14	10

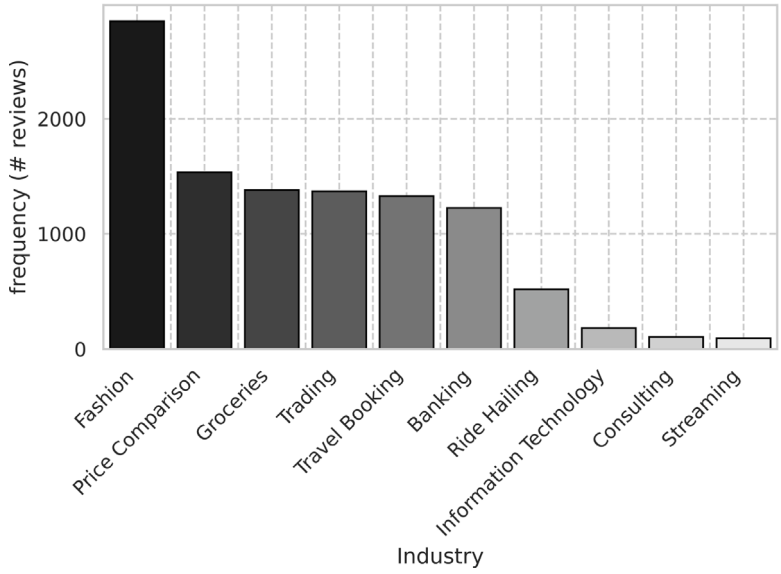


Fig. 1. Size of industries in terms of the number of reviews that they contain.

3.1. Data collection and pre-processing

The FABSA dataset is constructed from three different public data sources: Trustpilot,[2] Google Play[3] and Apple App Store.[4] Table 2 shows (a) the number of reviews, (b) the number of sentences, (c) the unique number of organisations, and (d) the unique number of industries per data source. Overall, we collected 10,574 feedback reviews from 14 organisations that span 10 different industries (e.g. banking, travel booking, fashion). The majority of the reviews are crawled from Google Play ($\sim 6.1K$ reviews) while Trustpilot and the Apple App Store cover $1.9 - 2.5K$ reviews, respectively.

Fig. 1 shows the size of each industry in terms of the number of reviews that they contain. Fashion (retail clothing) is the largest industry in our FABSA dataset, consisting of approximately 2.8K reviews, while Streaming is the smallest industry (94 reviews).

Regarding pre-processing, we mask organisation names with unique identifiers to avoid the identification of specific organisation names in our dataset. As an example, the review "*I love the Nike app!*" is converted into "*I love the ORGXX app!*" wherein the organisation name (i.e. *Nike*) is replaced with its corresponding unique identifier (*ORGXX*).

3.2. Annotation scheme

The FABSA dataset is manually labelled against a hierarchical annotation scheme which consists of 7 parent and 12 child aspect categories (Fig. 2). Each aspect category is associated with a sentiment label (positive, negative and neutral). This creates a total of 36 (12×3) target classification categories.

Following previous work, we adopt a multi-label classification scheme wherein each review is labelled with one or more aspect+sentiment label. Accordingly, a single review may contain multiple different aspects and express different (and in some cases contrasting) polarities. Table 3 shows an example of a review which is associated

[2] uk.trustpilot.com
[3] play.google.com
[4] www.apple.com/uk/app-store/

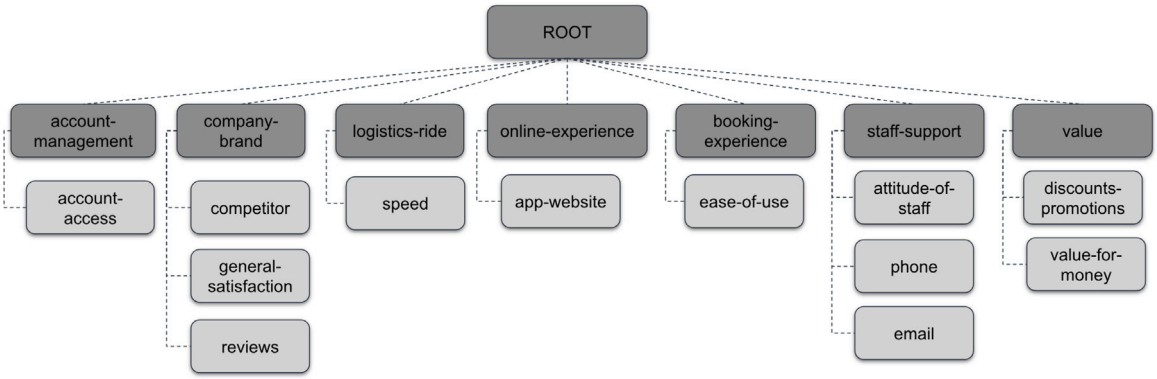


Fig. 2. Annotation scheme of the FABSA dataset consisting of 7 parent and 12 child aspect categories.

Table 3
An example of a review that contains multiple aspects with contrasting polarities.

Review	Labels
"product is very good but customer service is really bad, they never respond"	(general-satisfaction, Positive) (attitude-of-staff, Negative)

with two different aspect categories and two different and conflicting polarity labels (positive and negative, respectively).

The FABSA annotation scheme was refined iteratively over multiple annotation rounds. During each annotation round, we revised both the annotation scheme and annotation guidelines to resolve ambiguities and to increase the coverage of our scheme.

3.3. Annotation process

We invited three annotators to initially annotate a sample of 2,000 reviews (20% of the FABSA dataset). The annotators were provided with guidelines and accompanying examples that demonstrate how reviews should be labelled against our annotation scheme. The annotation guidelines are made publicly available together with the FABSA dataset. Two annotators have extensive experience in developing manually labelled ABSA datasets for a commercial company, but do not have a formal background/education related to linguistics. The third annotator has a PhD in computational linguistics and is assumed to be an expert tagger.

We firstly compute the inter-annotator agreement between all three annotators. We then select the annotator that has the highest agreement with the expert tagger to annotate the complete FABSA dataset. Similar annotation strategies are reported elsewhere in the literature. Saeidi et al. (2016) [17] employed three annotators to manually label a sample of their SentiHood dataset. The annotator with the best inter-annotator agreement was then used to label the complete dataset.

Table 4 shows pairwise inter-annotator agreements (F1 score) between tagger 1, tagger 2 and the expert tagger. We compute F1 agreement scores across the different levels of our hierarchical annotation scheme, namely *sentiment+aspect*, *aspect (child)* and *aspect (parent)* categories.

It can be observed that tagger 2 has a relatively low agreement score of 73.7% with the expert tagger. A high number of disagreements between tagger 2 and the expert tagger occurred on the *online-experience* aspect category. More specifically, we noted that tagger 2 failed to identify the *online-experience* aspect when the use of an app or software was implied but not explicitly stated within a review. As an example, the review "*very easy to input info*" was assigned the label (*company-brand.general-satisfaction, Positive*) by tagger 2 rather than the correct label (*online-experience, Positive*).

Tagger 1 obtained a reasonably high F1 agreement score of 82% with the expert tagger, which is consistent with inter-annotator agreement scores reported by previous work [30]. Based upon this, tagger 1 was selected to manually label the remaining 8,000 reviews of the FABSA dataset.

Table 4
Inter-annotator agreement (F1 score) across sentiment+child-aspect, child-aspect and parent-aspect categories.

	Sentiment+aspect	Aspect (child)	Aspect (parent)
Tagger 1 vs. expert tagger	**.820**	**.831**	**.840**
Tagger 2 vs. expert tagger	.737	.749	.757
Tagger 1 vs. tagger 2	.726	.734	.753

3.4. Dataset statistics

Fig. 3 shows the number of reviews that contain at least one positive, negative or neutral sentiment label. It can be noted that the sentiment labels are imbalanced in that more than 7,000 reviews contain a positive label while 3,318 have a negative label. Moreover, the neutral class is present in only 657 (6.2%) reviews . The neutral class includes reviews without a sentiment orientation ("*I use ORG160 for travel*").

Fig. 4 shows the distribution of aspect labels. The category *app-website* is the most frequent aspect category which can be explained by the fact that the majority of the reviews are sampled from Google Play and the Apple App store. The *general-satisfaction* label is the second most frequent aspect category. This label is used in diverse contexts and it includes mentions of general satisfaction towards a brand or product.

We further perform a correlation analysis of the aspect labels in order to identify labels that tend to co-occur together in the same reviews. Fig. 5 reports pairwise Pearson correlation coefficients between the 12 aspect labels of our dataset. A moderate positive correlation coefficient of 0.26 is observed between the *phone* and *attitude-of-staff* labels. This stems from the fact that both aspect labels appear in similar contexts relevant to communication between staff and customers. It can also be noted that the *price-value-for-money* aspect label yields a positive correlation with the *competitor* label. These two aspect labels usually appear together when the underlying review compares the price of the brand's product against the price of a competitor's product. Table 5 shows an example of a review that contains both a *competitor* and a *price-value-for-money* label.

4. Methods

This section provides implementation details of different ABSA models that we evaluate on our FABSA dataset. As baseline methods, we implement a logistic regression classifier that uses bag-of-words features

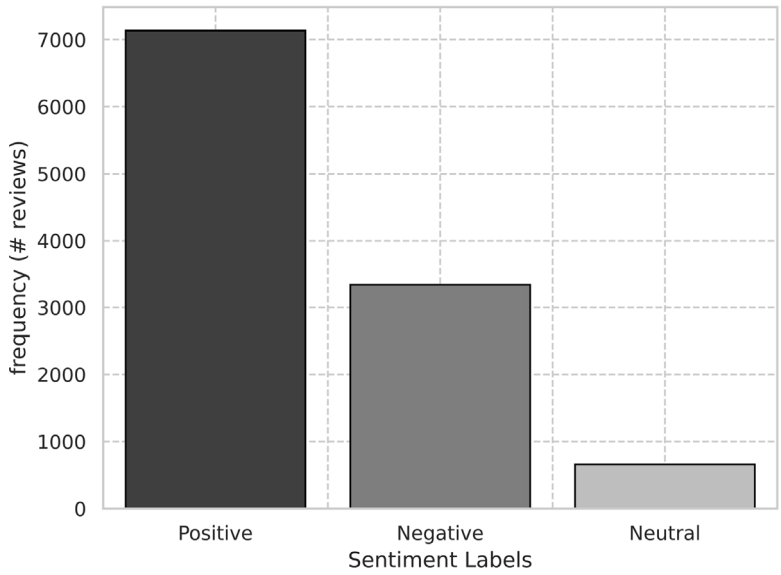


Fig. 3. Distribution of sentiment labels in the FABSA dataset.

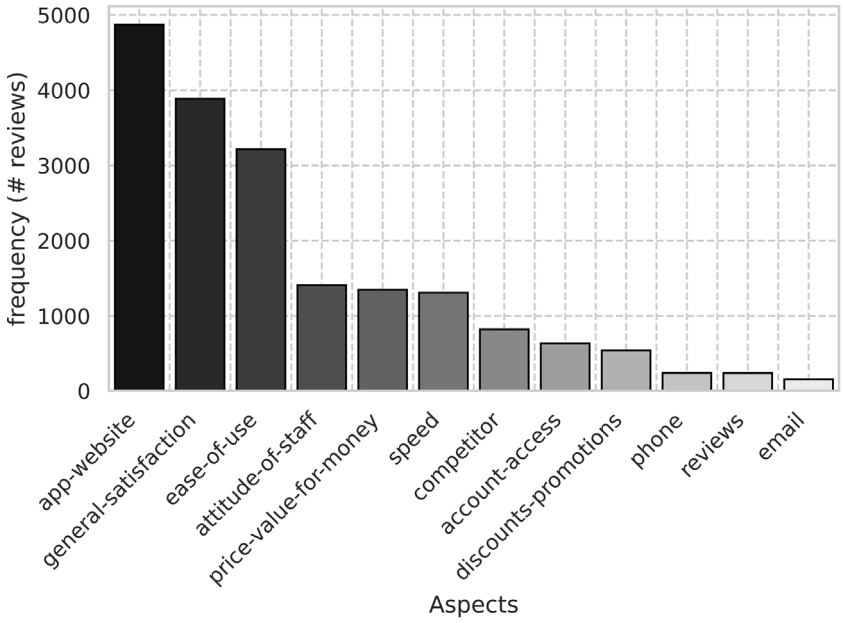


Fig. 4. Distribution of aspect labels.

Table 5
Example of a review that contains a *competitor* label and a *price-value-for-money* label.

Review	Labels
"*the price of hotel bookings is* ***cheaper than some other similar apps***"	(competitor, Positive) (price-value-for-money, Positive)

(LogReg-BoW) and a Convolutional Neural Network (CNN) classifier which is built on top of a Gated Recurrent Unit (GRU-GNN). Moreover, we fine-tune 6 large language models, namely BERT-single-base/BERT-single-large [33], RoBERTa-single-base/RoBERTa-single-large [34] and DeBERTa-single-base/DeBERTa-single-large [35], on a single text classification task (i.e. input is a review while output is the aspect+sentiment label of that review). In addition to the single-sentence text classification models, we fine-tune the RoBERTa (RoBERTa-pair-base/RoBERTa-pair-large) and DeBERTa (DeBERTa-pair-base/DeBERTa-pair-large) language models on a sentence-pair classification task (i.e. input is a pair of a review and a candidate aspect while output is the sentiment label of the candidate aspect in the given review). Finally, we adopt two previously introduced ABSA methods, namely BERT-PT [36] and GAS [21], which we fine-tune on the FABSA dataset. BERT-PT post-trains the BERT-base model on Amazon and Yelp reviews in order to improve the performance of the model when analysing review data. GAS (Generative Aspect-based Sentiment Analysis) follows a sequence generation approach (text-to-text) to ABSA. GAS uses T5-base [37], an encoder–decoder large language model, to learn to generate ABSA annotations when trained on a relevant ABSA dataset. Table 6 shows two examples of an input and output sequence of the GAS model.

With regard to the LogReg-BoW classifier, we apply basic pre-processing steps by removing stop words which are found in NLTK's stop word list.[5] Moreover, we use NLTK's lemmatizer to convert the

[5] www.nltk.org

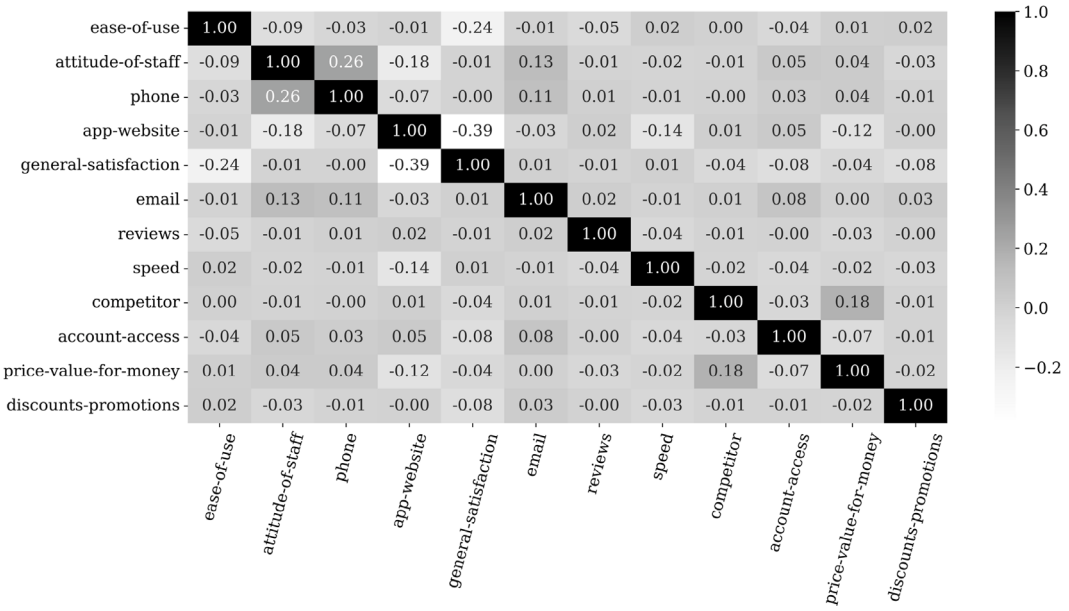


Fig. 5. Pairwise correlation coefficients between aspect labels.

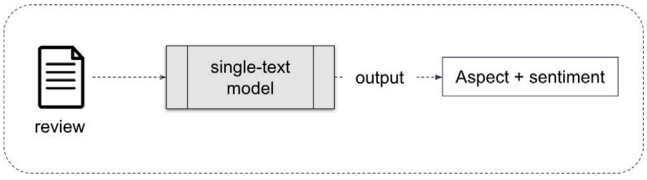


Fig. 6. Architecture of a single-text classification model.

Table 6
Example of an input and output sequence of the GAS sequence generation model.

Input sequence	Output sequence
Professional and friendly advice and service	(staff-support.attitude-of-staff, positive, general-satisfaction, positive)
App always has connection problems for me	(app-website, negative)

surface forms of words (e.g. *navigation, navigating*) into their corresponding base forms (*navigate*). We then convert the reviews into sparse document vectors consisting of the 20,000 most frequent word-lemmas in our dataset. As feature values, we use the tf-idf weighting scheme which normalises the frequencies of words by their inverse document frequencies. Finally, the sparse document vectors are used to train a Logistic Regression classifier.

The GRU-CNN classifier was previously introduced in [38] for classifying sentiment labels found in short texts. The model consists of the following 4 layers: (a) a word embedding layer, (b) a GRU, (c) a CNN and (d) a fully connected layer which generates the final output of the network. The word embedding layer encodes each word of the input review into a dense vector representation. In our experiments, we initialise the word embedding layer with 300-dimensional GloVe vectors which are pre-trained on a large web corpus consisting of 840 billion tokens [39]. The GRU and the CNN layers are both used as feature extractors. The GRU layer is able to identify long-range contextual features, while the CNN layer is more efficient at extracting local contextual features. Note that optimal hyperparameters for the GRU-CNN classifier are selected using the development subset; the GRU-GNN classifier is trained for 10 epochs using the Adam optimiser [40] with a learning rate of 0.00003.

The transformer models (e.g. BERT-single-base, RoBERTa-single-base, DeBERTa-single-large) fine-tune an encoder-only language model, such as BERT [33], RoBERTa [34] or DeBERTa [35], on the FABSA dataset. They take as input a sequence of text and generate an embedding representation for each token of the input sequence. A special meta-token ([CLS]) is added at the beginning of each sequence and it is used as a representation of the entire sequence. The output embedding of the CLS token can be subsequently processed by a classification head for text classification tasks.

The six single-sentence classification models, BERT-single-base, BERT-single-large, RoBERTa-single-base, RoBERTa-single-base, DeBERTa-single-base and DeBERTa-single-large adopt the same input and output representation (Fig. 6). Here, the input is a single review while the output is a 36-dimensional vector. Each dimension of the output vector corresponds to an aspect+sentiment label of our dataset. Considering that a review may contain multiple aspect+sentiment labels (multilabel classification task), we set a threshold to select multiple relevant labels to the input review.

The sentence-pair classification models (i.e. RoBERTa-pair-base, RoBERTa-pair-large, DeBERTa-pair-base and DeBERTa-pair-large) follow a sentence-pair classification approach to aspect concept extraction. Fig. 7 shows the input and output representation of a sentence-pair classification model. The input to the sentence-pair model consists of the review and the target aspect separated by a special meta-token ([SEP]) while the output has 4 possible values: *absent*, *positive*, *negative* and *neutral*. The *absent* label indicates that the target aspect is not present in the given review.

A sentence-pair classification model can potentially predict aspects that are not present in the training dataset and therefore, it can be used in zero-shot classification scenarios [20]. However, a limitation of a sentence-pair model is that its inference time grows linearly with the number of target aspects (for every review we create 12 instances at inference time, one instance for each aspect of the dataset).

We use the FABSA dev dataset for identifying the best hyperparameter values for the transformer models. We train all transformer models for 10 epochs with the exception of the GAS model which we train for 20 epochs. We pad the input sequences to 100 tokens, set the dropout probability to 0.1 and use a batch size of 16. As our optimiser, we use AdamW [41] with a learning rate of 0.00003 and a linear learning rate decay. We perform all our experiments on a single TITAN RTX GPU with 24 GB of memory.

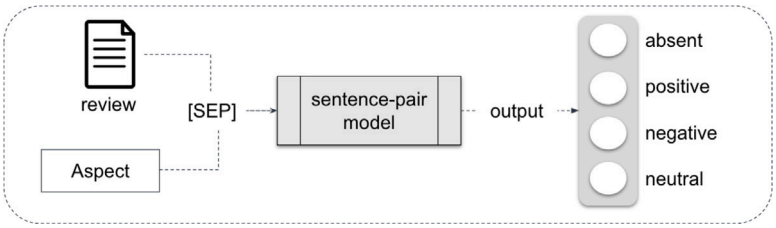


Fig. 7. Architecture of a sentence-pair classification model.

5. Experiments

In this section, we report the results of the ABSA models when they are trained and evaluated on the FABSA dataset.

5.1. Evaluation settings

We randomly partition the FABSA dataset into training, dev and test subsets consisting of 70%, 10% and 20% of the data, respectively. We then perform cross-validation by computing the average performance of the baseline models across 5 validation rounds where each validation round uses different partitions of training/dev/test subsets. All experimental results are computed over the 12 child aspect categories.

As evaluation metrics, we use the standard micro average precision, recall and F1 score metrics which are computed over the $n = 36$ (12 child aspect categories $\times$ 3 sentiment categories) labels of the dataset:

$$P = \sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i}, \quad R = \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i}, \quad F1 = \frac{2PR}{P + R} \tag{1}$$

where $n = 36$ is the number of aspect+sentiment labels while TP, FP and FN refer to the number of true positives, false positives and false negatives, respectively.

5.2. Results

Table 7 shows the precision, recall and F1 score performance of 14 ABSA models. The DeBERTa-pair-large model obtains the best F1 score performance while the DeBERTa-single-large yields the best recall. However, performance differences between the DeBERTa-single-large and the DeBERTa-pair-large are statistically insignificant.[6]

The DeBERTa-pair-large shows a statistically significant performance gain over the two RoBERTa-large models (RoBERTa-single-large and RoBERTa-pair-large). Moreover, the DeBERTa-pair-large outperforms the BERT-single-large model by 2.1% (F1 score). Previous studies have also demonstrated a superior performance of DeBERTa over RoBERTa and BERT across different NLP tasks [42]. This is explained by the fact that the DeBERTa model is pre-trained in a more efficient way (using a disentangled attention mechanism and an enhanced masked decoder) when compared to BERT and RoBERTa.

The BERT-PT model uses the same model architecture as BERT-single-base but it is further pre-trained on additional review data. BERT-PT improves upon the performance of BERT-single-base by 2.3% (F1 score) which shows the effectiveness of these models when they are pre-trained on in-domain datasets. GAS [21] yields a competitive F1 score performance of 0.782 which is on par with the performance obtained by BERT-PT.

The large RoBERTa models (RoBERTa-singe-large, RoBERTa-pair-large), which consist of 24 transformer blocks instead of 12 blocks found in the base models, obtain a statistically significant performance increase over their corresponding RoBERTa-base architectures (RoBERTa-single-base, RoBERTa-pair-base). Finally, the bag-of-words (LogReg-BoW) and GRU-CNN model show a substantially lower performance when compared to the transformer models.

Table 7
Precision, recall and F1 score performance of 10 different ABSA models.

	P	R	F1
LogReg-BoW	.784	.491	.604
GRU-GNN	.702	.667	.684
BERT-single-base	.785	.745	.765
BERT-single-large	.785	.792	.788
BERT-PT [36]	.785	.792	.788
GAS [21]	.785	.779	.782
RoBERTa-single-base	.781	.787	.784
RoBERTa-single-large	.792	.816	.804
RoBERTa-pair-base	.796	.764	.779
RoBERTa-pair-large	**.807**	.792	.800
DeBERTa-single-base	.777	.787	.782
DeBERTa-single-large	.791	**.820**	.805
DeBERTa-pair-base	.793	.804	.798
DeBERTa-pair-large	.806	.812	**.809**

We further compute the performance of DeBERTa-pair-large, DeBERTa-single-large and RoBERTa-pair-large over an increasing number of training instances (Fig. 8). The F1 score performance of all three models continuously improves as the size of the training dataset increases. We observe large performance improvements of 6% (DeBERTa-pair-large), 5.4% (RoBERTa-pair-large) and 8.4% (DeBERTa-single-large) when increasing the size of the training dataset from 1,000 to 3,000 instances. The models converge to a high F1 score performance of 80% when using 6,000 to 7,000 training instances. This indicates that large manually annotated training datasets are needed in order to develop robust ABSA models. It should also be noted that the DeBERTa-pair-large yields substantial performance improvements over the DeBERTa-single-large (+4.1%) when using a small training sample of 1,000 instances.

5.3. Error analysis and discussion

We perform an error analysis to identify cases that are misclassified by the DeBERTa-pair-large model. Firstly, we report the F1 score obtained by the model across the different aspect categories of the FABSA dataset. Secondly, we compute the performance of the model on short and long reviews and we analyse classification errors.

Fig. 9 shows the F1 score of the DeBERTa-pair-large model over positive, negative and neutral sentiment labels for each aspect of the dataset. Overall, there is a strong and significant correlation ($R^2 = 0.76, p < 0.05$) between the performance of the model on a given label and the size of that label in terms of the number of training instances that it contains. As an example, we observe that in most cases the model achieves the best F1 score performance on Positive labels which is explained by the large number of training instances that belong to this class.

Evidently the model failed to predict the Neutral sentiment label for several aspect categories of the FABSA dataset. This can be attributed to the fact that the neutral label is sparsely distributed across the different aspects. For example, the training dataset contains 6 instances for the

[6] We use the two-tailed paired t-test ($p < 0.05$) to perform statistical significance analysis.

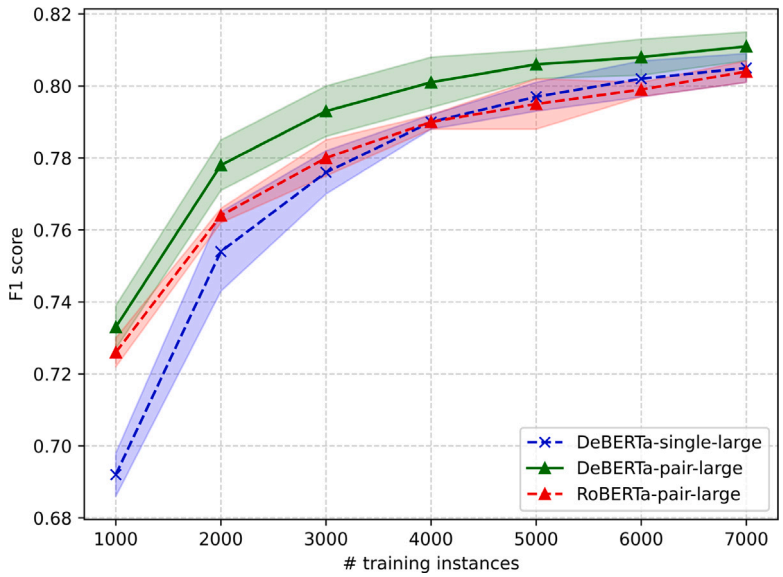


Fig. 8. F1 score performance of the RoBERTa-single-large and RoBERTa-pair-large model over an increasing number of training instances. The bands surrounding the thick lines represent the 95% confidence interval of the mean F1 score across 5 validation rounds.

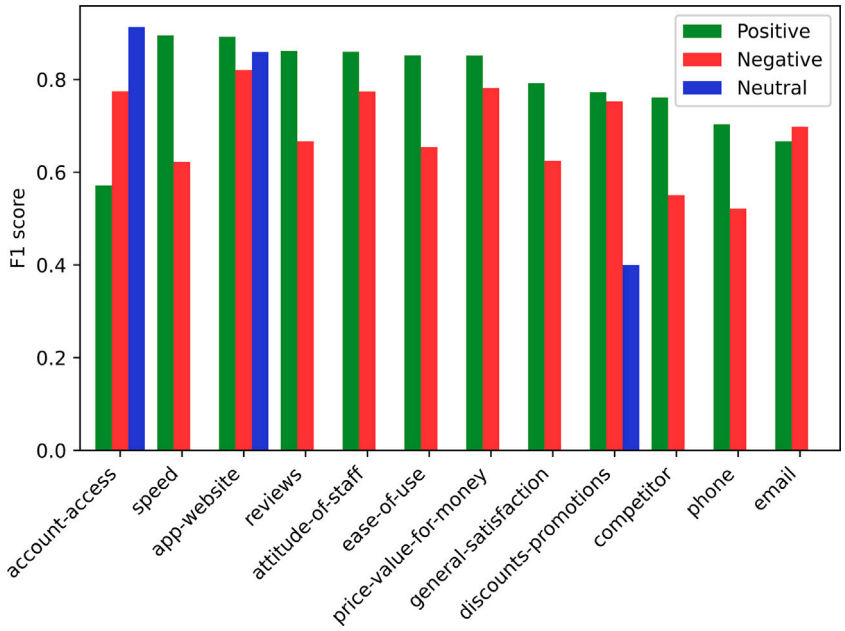


Fig. 9. F1 scores of the DeBERTa-pair-large model across Positive and Negative sentiment labels for each aspect of the FABSA dataset.

<*competitor, Neutral*> label. However, the DeBERTa-pair-large model yields a robust performance on the <*account-access, neutral*> and <*app-website, neutral*> labels, which consist of 202 and 360 neutral training instances, respectively.

The last experiment (Fig. 10) investigates the performance of the DeBERTa-pair-large model when applied to: (a) short reviews that contain no more than 10 tokens, (b) medium reviews (10 to 50 tokens) and (c) long reviews (more than 50 tokens). Here, we note a small decrease in the performance of the model on medium reviews when compared to short reviews. However, the performance of the model substantially decreases on long reviews (−8.6% to −12% compared to medium and short reviews, respectively). This shows that long reviews, which contain multiple clauses and/or sentences, are difficult to classify by the ABSA model.

6. Conclusions

In this paper, we have presented FABSA, a large-scale multi-domain dataset for aspect concept extraction. FABSA is manually labelled against a hierarchical annotation scheme consisting of 7 parent and 12 child aspect categories while each aspect category is associated with a positive or negative sentiment label. FABSA is the largest manually curated ABSA dataset to date consisting of more than 10,000 reviews covering 10 different domains. The development of FABSA enabled us to train and evaluate different machine learning based ABSA models. The results that we obtained demonstrate that large pre-trained transformer models obtain a superior classification performance when compared to other baseline ABSA models (e.g. bag-of-words). However, the performance of the models decreased on rare aspect categories for

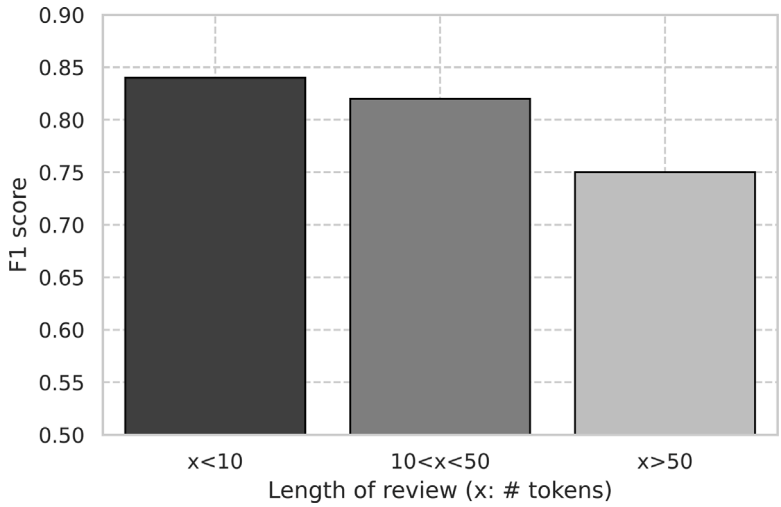


Fig. 10. F1 score performance of the DeBERTa-pair-large when applied to short (no more than 10 tokens), medium (between 10 and 50 tokens) and long reviews (more than 50 tokens).

which few instances are available for training. By releasing FABSA, we hope to establish a new large-scale and challenging benchmark dataset for evaluating future work on aspect concept extraction.

CRediT authorship contribution statement

Georgios Kontonatsios: Conceptualization, Methodology, Software, Writing – original draft, Formal analysis, Writing – review & editing. **Jordan Clive:** Conceptualization, Methodology, Software, Writing – review & editing. **Georgia Harrison:** Conceptualization, Methodology, Software, Writing – review & editing. **Thomas Metcalfe:** Conceptualization, Methodology, Software, Writing – review & editing. **Patrycja Sliwiak:** Conceptualization, Methodology, Software, Writing – review & editing. **Hassan Tahir:** Conceptualization, Methodology, Software, Writing – review & editing. **Aji Ghose:** Supervision, Conceptualization, Methodology, Software, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data availability

We have shared public link to the dataset.

References

[1] S. Sun, C. Luo, J. Chen, A review of natural language processing techniques for opinion mining systems, Inf. Fus. 36 (2017) 10–25.

[2] Y.A. Solangi, Z.A. Solangi, S. Aarain, A. Abro, G.A. Mallah, A. Shah, Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis, in: 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences, ICETAS, IEEE, 2018, pp. 1–4.

[3] P. Mehta, S. Pandya, A review on sentiment analysis methodologies, practices and applications, Int. J. Sci. Technol. Res. 9 (2) (2020) 601–609.

[4] S. Das, M. Chen, Yahoo! for Amazon: Extracting market sentiment from stock message boards, in: Proceedings of the Asia Pacific Finance Association Annual Conference, Vol. 35, APFA, Bangkok, Thailand, 2001, p. 43.

[5] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, 2002, arXiv preprint cs/0205070.

[6] S. Morinaga, K. Yamanishi, K. Tateishi, T. Fukushima, Mining product reputations on the web, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 341–349.

[7] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: International Workshop on Semantic Evaluation, 2016, pp. 19–30.

[8] A. Nazir, Y. Rao, L. Wu, L. Sun, Issues and challenges of aspect-based sentiment analysis: A comprehensive survey, IEEE Trans. Affect. Comput. (2020).

[9] B. Liang, H. Su, L. Gui, E. Cambria, R. Xu, Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks, Knowl.-Based Syst. 235 (2022) 107643.

[10] S. Ruder, P. Ghaffari, J.G. Breslin, A hierarchical model of reviews for aspect-based sentiment analysis, 2016, arXiv preprint arXiv:1609.02745.

[11] K. Liu, L. Xu, J. Zhao, Opinion target extraction using word-based translation model, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1346–1356.

[12] D. Ma, S. Li, F. Wu, X. Xie, H. Wang, Exploring sequence-to-sequence learning in aspect term extraction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3538–3547.

[13] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 452–461.

[14] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, 2018, arXiv preprint arXiv:1805.07043.

[15] W. Wang, S.J. Pan, D. Dahlmeier, X. Xiao, Coupled multi-layer attentions for co-extraction of aspect and opinion terms, in: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1, 2017.

[16] H. Peng, L. Xu, L. Bing, F. Huang, W. Lu, L. Si, Knowing what, how and why: A near complete solution for aspect-based sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05, 2020, pp. 8600–8607.

[17] M. Saeidi, G. Bouchard, M. Liakata, S. Riedel, SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1546–1556.

[18] J. Yu, C. Gong, R. Xia, Cross-domain review generation for aspect-based sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4767–4777.

[19] Y. Song, J. Wang, T. Jiang, Z. Liu, Y. Rao, Attentional encoder network for targeted sentiment classification, 2019, arXiv preprint arXiv:1902.09314.

[20] C. Sun, L. Huang, X. Qiu, 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, Vol. 1, Long and Short Papers, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 380–385, http://dx.doi.org/10.18653/v1/N19-1035.

[21] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, Towards generative aspect-based sentiment analysis, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 504–510.

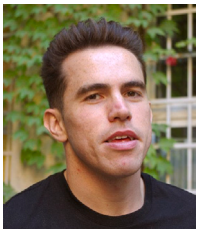
[22] J. Su, S. Yu, D. Luo, Enhancing aspect-based sentiment analysis with capsule network, IEEE Access 8 (2020) 100551–100561.

[23] X. Li, L. Bing, W. Zhang, W. Lam, Exploiting BERT for end-to-end aspect-based sentiment analysis, 2019, arXiv preprint arXiv:1910.00883.
[24] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4933–4941.
[25] C. Gong, J. Yu, R. Xia, Unified feature and instance based domain adaptation for aspect-based sentiment analysis, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 7035–7045.
[26] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, M. Carbin, The lottery ticket hypothesis for pre-trained bert networks, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 15834–15846.
[27] Q. Jiang, L. Chen, R. Xu, X. Ao, M. Yang, A challenge dataset and effective models for aspect-based sentiment analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6280–6285.
[28] Y. Yiran, S. Srivastava, Aspect-based sentiment analysis on mobile phone reviews with LDA, in: Proceedings of the 2019 4th International Conference on Machine Learning Technologies, in: ICMLT 2019, Association for Computing Machinery, New York, NY, USA, 2019, pp. 101–105.
[29] G. Ganu, N. Elhadad, A. Marian, Beyond the stars: Improving rating predictions using review text content, in: WebDB, Vol. 9, 2009, pp. 1–6.
[30] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, SemEVal 2014, Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35.
[31] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, Adaptive recursive neural network for target-dependent twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, 2014, pp. 49–54.
[32] N. Alturaief, H. Aljamaan, M. Baslyman, AWARE: Aspect-based sentiment analysis dataset of apps reviews for requirements elicitation, in: 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW, IEEE, 2021, pp. 211–218.
[33] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Vol. 1, Long and Short Papers, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[35] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2020, arXiv preprint arXiv:2006.03654.
[36] H. Xu, B. Liu, L. Shu, P.S. Yu, BERT post-training for review reading comprehension and aspect-based sentiment analysis, 2019, arXiv preprint arXiv:1904.02232.
[37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (1) (2020) 5485–5551.
[38] L.-x. Luo, Network text sentiment analysis method combining LDA text representation and GRU-CNN, Pers. Ubiquitous Comput. 23 (3) (2019) 405–412.
[39] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.
[40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[41] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, la, USA, May 6-9, 2019, OpenReview.net, 2019.
[42] B. Rodrawangpai, W. Daungjaiboon, Improving text classification with transformers and layer normalization, Mach. Learn. Appl. 10 (2022) 100403.

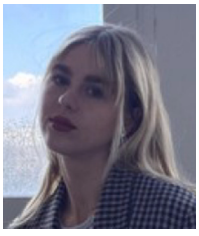
Georgios Kontonatsios is a Principal Data Scientist at Chattermill working on Deep Learning and Natural Language processing (NLP). Georgios holds a Ph.D. and an M.Sc. in Artificial Intelligence from the University of Manchester while his academic work includes more than 20 research publications in the field of NLP. Prior to joining Chattermill, Georgios worked as a Senior Lecturer at the University of Edge Hill where he led undergraduate and postgraduate NLP courses. His research focuses on a wide range of different NLP tasks including aspect-based sentiment analysis, named entity recognition, entity linking, relation extraction and zero-shot text classification.

Jordan Clive received the M.Sc. degree in Machine Learning and AI from Imperial College London in 2021. He worked on parameter-efficient text generation advised by Marek Rei and Kris Cao of DeepMind. He holds a Bachelor's degree in Mathematical Physics from Durham University. Jordan is currently a Deep Learning Engineer at Chattermill and also volunteers for the LAION AI (Stability AI) research group, conducting ML research on their Ezra-1 AI ultracluster. His research interests include long context summarisation and parameter-efficient adaptation of large language and vision models.

Georgia Harrison received an M.Sc. in Data Analytics from the University of Sheffield in 2018, with a specific focus on machine learning and natural language processing. She also holds a bachelor's degree in Mathematics, specialising in statistical methods and modelling. Currently, she serves as a Lead Data Scientist at Chattermill where she leverages state-of-the-art NLP methods on customer feedback data. She is a part-time lecturer at the Market Research Society and a founding member of the NLP Study Group meetup group, which has over 1000 members, where she regularly shares the latest advancements in AI with the community.



Thomas Metcalfe received the MPhys degree in Physics and Cosmology from Durham University in 2015. He is currently Chief Data Scientist at Twain where he works on Natural Language Generation products and is an alumni of Sequoia Capital's debut Arc cohort. He is an OS contributor of pytorch and IBM's quantum-computing framework Qiskit. His research interests include natural language generation, learning to rank and active sampling.



Patrycja Śliwiak is an AI researcher and Senior Data Scientist at Chattermill. Her focus is in Natural Language Processing, with emphasis on generative models. She holds a B.Sc. in Systems Engineering from Wroclaw University of Science and Technology and an M.Sc. in Artificial Intelligence from the University of Bath - a perfect blend of her interests in Software Engineering and AI. She's always committed to building practical solutions that can have a meaningful impact, with a strong foundation in well-designed systems.

Hassan Tahir is currently a Lead Deep Learning Engineer at Chattermill and also a Maker-in-Residence for the Ben's Bites AI publication with over 80,000 subscribers. He holds a bachelor's degree in Mathematics from University College London (UCL) where he explored areas such as group and ring theory including polynomial rings, topology of the complex plane and abstract measure theory leading to rigorous probability theory. After graduating he joined Chattermill as an early employee where he laid the foundations for the initial Data Operations and Data Science functions. With more than 5 years of experience in building NLP powered products, he has become a specialist in developing ML ideas end-to-end from research to proof of concepts through to production. His research interests include natural language processing, computer vision and deep reinforcement learning.

Aji Ghose serves as the Vice President of Data & Research at Chattermill, overseeing Data Science, MLOps, and MLEng, focused on unearthing insights from customer feedback. Previously, he was the Head of Research & Analytics at Sky, managing a large group of data scientists, researchers, and analysts, working on product and marketing optimisation, content recommendation, and robotic process automation. Aji earned his Ph.D. in Computational Cognitive Science, specialising in multimodal deep learning, from Birkbeck, University of London, supervised by Prof Rick Cooper. Following his Ph.D., Aji continued as a visiting postdoctoral fellow at Birkbeck. He holds an M.Sc. in Computer Science with Artificial Intelligence from the University of York and an Elite MA in Cognitive Semiotics (Symbolic Systems) from Aarhus University, Denmark. Aji also serves as a chair and lecturer for Data Science, Statistics, and AI courses at the Market Research Society. His research interests span NeuroAI, natural language processing, robotics, and cybersecurity.