Using Aspect-Based Sentiment Analysis to Classify ATTITUDE-bearing Words

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, David Eyers
Department of Computer Science
University of Otago
New Zealand
[pradeesh,andrew,veronica,dme]@cs.otago.ac.nz

Abstract

APPRAISAL is widely used by linguists to study how people judge things or people. Automating APPRAISAL could be beneficial for use cases such as moderating online comments. In 2020, the Australasian Language Technology Association (ALTA) organised a shared task to classify a branch APPRAISAL, which involves how humans judge other humans (JUDGEMENT). It proved to be a difficult task as the best performing system obtained an $F_1 = 0.155$. In this work, we hypothesise that JUDGEMENT and APPRECIATION branches in APPRAISAL are similar to opinion in Aspect-Based Sentiment Analysis (ABSA) tasks, as such we can leverage on ABSA opinion extraction techniques to further improve the performance of automated approaches for identifying JUDGEMENT and APPRECIATION. We evaluated the performance of six different ABSA models on two publicly available APPRAISAL data sets (biographies and psychological evaluation) by training them on existing ABSA SemEval data sets. Our results show that there is an overlap between opinion-extraction and APPRAISAL task, as we obtained $F_1 = 0.623$ on biographies data set and 0.414 on psychological evaluation data set. However, we cannot be certain if our findings can be extended across other APPRAISAL data sets due to the challenges in annotating and the availability of these data sets.

1 Introduction

In 2020, ALTA organised a shared task challenge aimed to classify how humans judge other humans using a well-known linguistic taxonomy known as APPRAISAL (Martin and White, 2005) automatically. APPRAISAL allows linguists to evaluate language in a social context such as identifying and understanding how people make judgements about people and objects (ATTITUDE) (Martin and White, 2005). The taxonomy is commonly used by Systemic Functional linguists to analyse the language choices and attitudes used by writers and speakers (Chen, 2022) in various mediums (Starfield et al., 2015; Ross and Caldwell, 2020; Su and Hunston, 2019).

Identifying ATTITUDE-bearing words can help to reduce the workload of moderators such as in online forums and Facebook by analysing the language that is being used and flagging it to moderators to be reviewed if there are any legal implications based on the APPRAISAL taxonomy (Steiger et al., 2021).

Although, there were two winners declared for the ALTA 2020 Shared Task challenge, the task proved to be difficult as the best-performing team only obtained an F_1 score of 0.155 (Mollá, 2020). The main reason for poor scores was the size of data set ($N = 300$): too small for automated methods to generalise from properly. A lot of the larger APPRAISAL data set is not publicly released, thus making it difficult for automated approaches to be be built.

However, there might be a solution for us to tackle this problem without the need of a large data set. Recently, Su and Hunston (2019), proposed that JUDGEMENT and APPRECIATION should be treated as opinions and AFFECT as emotions. Su and Hunston (2019), then provided qualitative examples to illustrate how JUDGEMENT and APPRECIATION can be viewed as opinions. Inspired by the findings of Su and Hunston (2019), we are interested in investigating this area, particularly if we can apply existing aspect-based opinion techniques to tackle this problem.

We argue that if the combination of the JUDGEMENT and APPRECIATION branches is the same as opinion, then the current ABSA opinion extraction techniques and models are applicable and therefore can be applied. BARTABSA is the current state of the art for ABSA's triplet extraction task.

BARTABSA has achieved the highest F_1 score in the triplet extraction task, which is 0.7246 on the laptop data set.[1] Thus, we are interested if we can use these existing models to identify JUDGEMENT and APPRECIATION-bearing words.

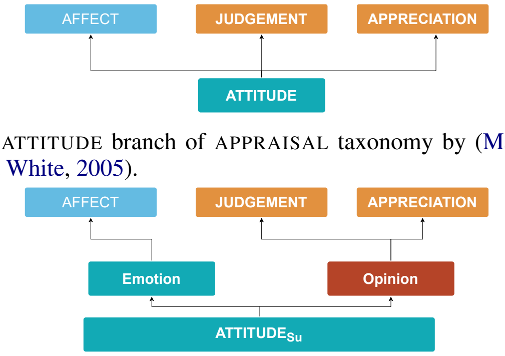
Our experimental results suggest that there is an overlap between JUDGEMENT and APPRECIATION words with the ABSA task. Existing ABSA models, that were trained on SemEval data sets, does perform reasonably well on JUDGEMENT data sets (F_1 = 0.623 on biographies, F_1 = 0.414 on psychological evaluation).

2 Related Work

The APPRAISAL taxonomy consists of three main branches: ATTITUDE, ENGAGEMENT and GRADUATION. ATTITUDE expresses the current state of the person who wrote the text or uttered it—it consists of three subcategories: AFFECT (which represents the feeling of the author), JUDGEMENT (which describes the author's opinion of another person or object) and APPRECIATION (which represents the author's opinion on the quality of an object). ENGAGEMENT reflects probability or possibility (i.e., *perhaps*, *seems*). GRADUATION expresses the meaning of a term gradated by an adjective. These APPRAISAL attributes are often expressed with polarity and orientation. Polarity describes the tone of the sentence (i.e., negative, positive or neutral) whereas orientation explores how a sentence is weakened or strengthened (i.e., *very*/*few*/*a lot*).

To illustrate, consider the appraisal analysis of the sentence 'Robin Hood gave a sly grin'. It describes the appraiser (i.e., the person who wrote it), their attitude, what it is being appraised, and their polarity.

Appraiser	: *Writer*
Appraised	: *Robin Hood*
Attitude	: *sly* (JUDGEMENT)
Polarity	: *Negative*



(a) ATTITUDE branch of APPRAISAL taxonomy by (Martin and White, 2005).

(b) Proposed change in ATTITUDE branch by (Su and Hunston, 2019).

Figure 1: The proposed change in ATTITUDE branch of APPRAISAL taxonomy (ATTITUDE$_{\text{Su}}$) by (Su and Hunston, 2019) and its comparison with the original ATTITUDE by (Martin and White, 2005).

Extracting this detail of information can be challenging, but some tasks (such as polarity extraction) have already been tackled in sentiment analysis (Kanayama and Nasukawa, 2006). Here, we narrow our focus to extracting ATTITUDE-bearing words, as we are interested in quantifying the changes proposed by Su and Hunston (2019), to determine if we could use opinion-extraction from ABSA to extract the opinion. From Figure 1, we can see that JUDGEMENT and APPRECIATION are seen as opinions and AFFECT is seen as emotions. This is the key difference from the original taxonomy of Martin and White (2005). ATTITUDE$_{\text{Su}}$ will be used to represent the new change proposed by Su and Hunston (2019) and we are narrowing our focus to the opinion branch of ATTITUDE$_{\text{Su}}$. Although numerous works have attempted to automatically categorise APPRAISAL (Argamon et al., 2007; Bloom and Argamon, 2010; Whitelaw et al., 2005; Neviarouskaya et al., 2010; Taboada et al., 2011), including the 2020 ALTA Shared Task, most of the previous work has focused on identification at the sentence level rather than at the word level (Argamon et al., 2007; Bloom and Argamon, 2010).

As ABSA is used to identify aspects, opinions, and polarity. It would be interesting to explore if ABSA can be used in our case. We hypothesise that it may be possible to use triplet extraction for JUDGEMENT and APPRECIATION. However, the current sets of publicly available APPRAISAL data sets (Su and Hunston, 2019; Mollá, 2021) only label the ATTITUDE and not the APPRAISED (aspect). Annotating APPRAISAL is not straightforward as experts with a linguistic background are likely to be needed to do so (Parameswaran et al., 2022)—and so crowdsourcing (Standing and Standing, 2018) is likely to yield unusable results. Doing this is beyond the scope of this paper.

For ABSA tasks, transformers are the current state-of-the-art (Do et al., 2019). Transformers such as BERT (Devlin et al., 2019), BART

[1] https://paperswithcode.com/paper/a-unified-generative-framework-for-aspect

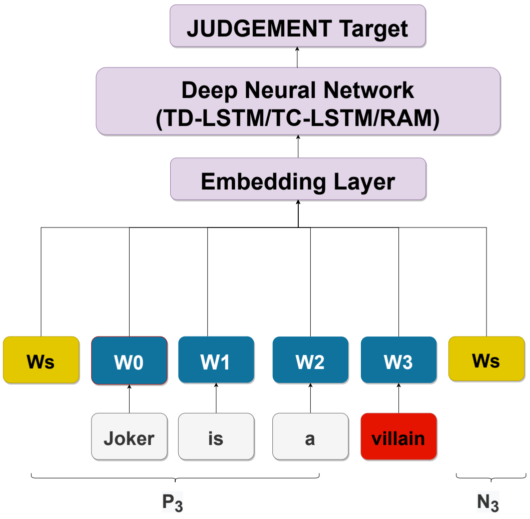


Figure 2: System architecture for LSTM-based models. Here, w_3 is the word to be classified as JUDGEMENT or not.

(Lewis et al., 2020), and GPT-2 (Radford et al., 2019) have consistently shown promising results in many NLP tasks (Adhikari et al., 2020). The current best for ABSA, BARTABSA (Yan et al., 2021), uses a sequence-to-sequence model to solve the triplet extraction problem. BARTABSA achieved an average F_1 score of 0.85 on opinion extraction and 0.58 on triplet extraction using the SemEval ABSA data set.

Prior neural approaches to ABSA such as TC-LSTM (Tang et al., 2016), TD-LSTM (Tang et al., 2016), and BERT-AEN (Song et al., 2019) have been used outside of ABSA. Their use in tasks such as the prediction of the sea temperature (Liu et al., 2018), the optimisation of virtual network demand optimisation (Kim et al., 2019), and sarcasm target identification and extraction of sarcasm targets (Patro et al., 2019) leads us to explore these models for JUDGEMENT extraction.

3 Data Sets Used in this Research

We use three data sets to evaluate our approaches, as summarised below. Two are already publicly available, and the third is a subset of the second, constructed in order to perform a like-to-like comparison with the first.[2]

Bio This is the data used by (Su and Hunston, 2019). It comprises 360 sentences taken from snippets of 100 biographies. The data set contains four fields: the sentence, the words that bear APPRECIATION, JUDGEMENT, and AFFECT in each sentence.

There are 80 sentences in the AFFECT category, 125 in the JUDGEMENT category, and 161 sentences in APPRECIATION. There are overlaps in these sentences because a sentence can contain AFFECT, JUDGEMENT, and APPRECIATION. Only adjectives are annotated in this data set, so nonadjective JUDGEMENT words are not known.

Psyc We crawled the psychological evaluation texts from the APPRAISAL website[3]. Although this data has not been used in the literature on APPRAISAL for analysis, the intended purpose of this data set was to train linguistic students on how to perform APPRAISAL analysis.

This data set contains 50 sentences along with the words that imply AFFECT, JUDGEMENT, and APPRECIATION. Of the 50 sentences, 38 sentences belong to the JUDGEMENT category, 42 in the APPRECIATION category, and 34 in the AFFECT category. Unlike *Bio*, all words (including *adverbs* and *adjectives*) were classified as JUDGEMENT or non-JUDGEMENT.

Psyc$_a$ The previous two data sets differ in their coverage of parts of speech. To make it possible to compare the performance of our models in *Bio* and *Psyc*, we created *Psyc$_a$* from *Psyc* by removing all non-adjectives.

In our experiments using *Bio*, *Psyc* and *Psyc$_a$* we perform a three-fold cross-validation because there is not a sufficiently large amount of data to divide into training, validation, and test sets.

4 Methodology

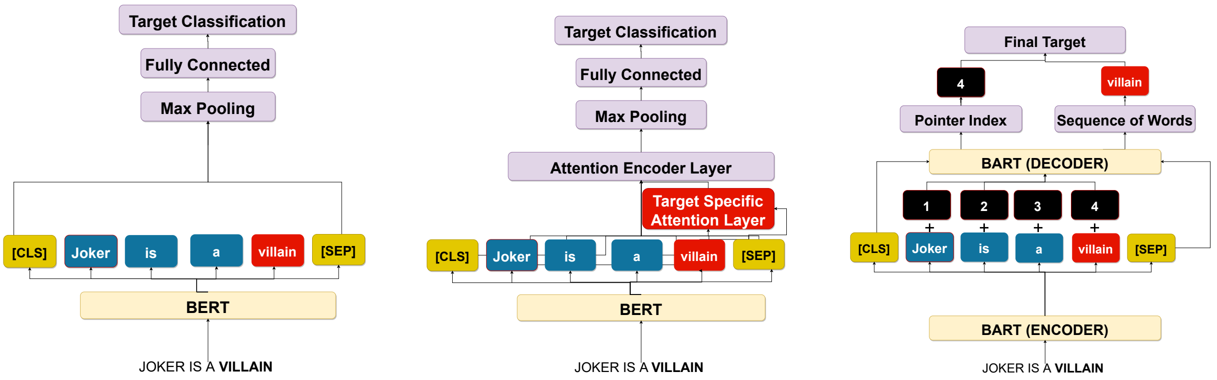
We briefly describe our methodology for carrying out our experimentation. We employ LSTM-based and transformer-based approaches.

4.1 Task Definition

We formulated our task as a sequence labelling problem similar to the way it was used for the opinion extraction task (Wang et al., 2016). A sentence S is defined as a sequence of words, $[w_1, w_2, w_3, \ldots, w_n]$. Our aim is to extract a set of phrases $X = \{o_1, o_2, o_3, \ldots, o_m\}$, where each $o \in X$ is either an opinion-ATTITUDE$_{\text{Su}}$ word or

[2] We will share the link to the data sets after the peer-review process

[3] http://www.grammatics.com/appraisal/pangesti/pangesti-psy-texts.pdf



(a) TD-BERT (b) BERT-AEN (c) BARTABSA

Figure 3: System architectures for the transformer-based models.

not and $|X| \leq n$. When a sentence does not contain any ATTITUDE$_{\text{Su}}$ word $|X| = 0$.

4.2 LSTM-based Models

Figure 2 shows the overall architecture of our LSTM-based models (TC-LSTM, TD-LSTM and RAM). For brevity, we summarise the mechanism of our LSTM-based models as shown in Figure 2 below:

- **TD-LSTM** (Tang et al., 2016)—The idea of this model is to use the preceding and the following context surrounding the target word as a feature. Two LSTM networks are used for this; the left LSTM neural network consists of the preceding sentence along with the potential opinion-ATTITUDE$_{\text{Su}}$ word, and the right LSTM neural network consists of the remaining context along with the potential target. The left LSTM network runs from left to right, and the right LSTM network runs from right to left. These LSTM networks are capable of learning the semantics of the sentence (Tang et al., 2016).

- **TC-LSTM** (Tang et al., 2016)—This is a modification of TD-LSTM. The key difference between TC-LSTM and TD-LSTM is that, in TD-LSTM the input at each position includes the embedding of the current word, whereas TC-LSTM contains the concatenation of the set of words preceding and following the opinion-ATTITUDE$_{\text{Su}}$ word. We expect that the concatenation of the words will result in a higher accuracy than that of TD-LSTM.

- **RAM** (Chen et al., 2017)—This uses a bidirectional LSTM to produce a memory slice. The memory slice is used to address the shortcomings of the TC-LSTM model (not being able to capture the target word if it is far away from the target). These memory slices are weighted according to the position of the target. The input of RAM is the entire sentence and the distance of potential opinion-ATTITUDE$_{\text{Su}}$. Then, to classify the target of the results are combined non-linearly with a Gated Recurrent Unit (GRU).

Given the sentence '*Joker is a villain*' and '*villain'* as the current potential opinion-ATTITUDE$_{\text{Su}}$ word, we start by computing the embedding of each word of each sequence. We use the BERT embedding to perform a fair comparison between all LSTM-based models. Once the embeddings are computed, they are then averaged and passed to the deep neural network layer to determine the probability that villain is an ATTITUDE$_{\text{Su}}$-bearing word.

4.3 Transformer-based Models

Figure 3 shows the architecture of the transformers that we used for our experimentation, which are TD-BERT (Gao et al., 2019), BERT-AEN (Yan et al., 2021), and BARTABSA (Yan et al., 2021). We briefly describe the functionality below:

- **TD-BERT** (Gao et al., 2019)—TD-BERT's architecture closely resembles that of BERT. The key difference is that TD-BERT incorporates the potential target information into its classification input, as described above.

- **BERT-AEN** (Song et al., 2019)—This model uses an attention encoder network to model the semantic interaction between the whole sentence and the potential opinion-ATTITUDE$_{\text{Su}}$ word. The Target Specific Attention Layer is introduced so that it can compute the hidden states of the input embedding. Moreover, BERT-AEN uses label smoothing regularisation (LSR) in the loss function. LSR reduces overfitting by replacing the 0 and 1 targets for the classifier with smoothed values (such as 0.1 and 0.9, respectively). This works well in our situation, where we have a limited amount of data.

- **BARTABSA** (Yan et al., 2021)— (Yan et al., 2021) formulate ABSA as a sequence-to-sequence generation task. Specifically, they use a pre-trained BART model (Lewis et al., 2020) to extract a sentence's opinion, aspect, and polarity. BART brings together the strength of the GPT-2 model (decoder) and BERT (encoder) for text understanding and generation. Therefore, the researchers were able to exploit the '*student-teacher*' (Malik et al., 2021) concept, in which the network consists of an encoder (the teacher) and a decoder (the student). We are only interested in the opinion phrase, so we modify the model so that the decoder extracts only opinion-ATTITUDE$_{\text{Su}}$ words.

First, we feed a sentence S to our transformer-based models, which is a sequence of words $[w_1, w_2, \ldots, w_N]$. We then transform the given sentence (S) into `[CLS]` $+\!\!+$ S $+\!\!+$ `[SEP]` and `[CLS]` $+\!\!+$ w_k $+\!\!+$ `[SEP]` together with the label w_k, where $k \in \{1 \ldots N\}$. Here within is where all the similarities of all the transformer models stop; for BARTABSA—we include positional input which are $P = (p_s, p_1, ...p_k)$, where p_k is the positional encoding for w_k. Positional encoding is introduced to keep in mind the sequence of words that appear in the given S. We did not use these information for our other two models as it was not required.

For TD-BERT and BERT-AEN, we use pre-trained BERT$_{\text{Base}}$ uncased (Devlin et al., 2019) and for BARTABSA, we use BART$_{\text{Base}}$ as the pre-trained model (Lewis et al., 2020). For TD-BERT and BERT-AEN, there are not any positional encoding.

Data Set:	*Lap14*	*Res14*	*Res15*
Number of sentences	3848	3844	2000
Number of opinion terms	3178	4492	1720
Average number of opinion words	0.82	1.16	0.86

Table 1: Details of the SemEval data sets used as part of sanity checks for the models we have described in Section 4.

5 SemEval Data Set and Sanity Check

We use three SemEval data sets: *Lap14*, *Res14* and *Res15* (Pontiki et al., 2015, 2016) to check our implementations. Initially, these data sets contained only aspects and sentiments, but Wang et al. (2016) annotated the data to contain opinion terms. Table 1 describes the distribution of items in the data sets. Wang et al. (2016) used crowdsourcing workers to annotate this data set. However, they did not provide the agreement level between the annotators. We hypothesise that the level of agreement between the annotators is high because models such as BARTABSA were able to obtain high F_1 scores. Therefore, we hypothesise that if the opinion identification task in ABSA is trivial, it would also mean that automated approaches can perform well in identifying JUDGEMENT and APPRECIATION.

These data have already been divided into training and test sets. We maintain those splits in our experiments. The purpose of using the SemEval data set is to validate our implementation to ensure that the scores we obtained are within the range of the scores reported in the literature (Zhang et al., 2022). By verifying if our implementation is correct, we can then evaluate the performance of these models in our data sets.

6 Experimental Setup

For our experiments, we used pyTorch-ABSA[4]. The framework is implemented in PyTorch[5] 1.71, spaCy[6] 1.9 and huggingface 3.4.0. We ran our experiments on Google Cloud Platform with 16 vCPUs (Intel Xeon E5 CPU @ 2.50Ghz), 16 GiB of RAM and an NVIDIA Tesla P100.

We use two baselines. First, we use the Naive Bayes (NB) classifier as our baseline. We trained the NB classifier on the three SemEval data sets

[4] https://github.com/songyouwei/ABSA-PyTorch

[5] https://www.tensorflow.org/

[6] https://pypi.org/project/spacy/

Model	Original implementation (Acc)	Ours (Acc)	Diff
TD-LSTM	0.764 (Tang et al., 2016)	0.746	-2.42%
TC-LSTM	0.760 (Tang et al., 2016)	0.721	-5.14%
Model	Original implementation (F_1)	Ours (F_1)	Diff
RAM	0.708 (Chen et al., 2017)	0.659	-6.86%
TD-BERT	0.769 (Gao et al., 2019)	0.780	1.35%
BERT-AEN	0.737 (Song et al., 2019)	0.712	-3.42%
BARTABSA	0.870 (Yan et al., 2021)	0.828	-4.88%

Table 2: Performance of our implementation compared with the authors' original performance on the *Res14* data set. TD-LSTM and TC-LSTM models comparisons are using accuracy score (Acc) and the others are using F_1 because those are the metrics reported by the original authors.

using the same split. We use SO-CAL (Taboada et al., 2011) as our second baseline because it is the only publicly available APPRAISAL classifier. SO-CAL produces a probability score for each category of the APPRAISAL taxonomy, but we are only interested in JUDGEMENT and APPRECIATION. So we only consider the word to be JUDGEMENT or APPRECIATION if either one of the labels is the highest of the probabilities and if the probability of JUDGEMENT or APPRECIATION is greater than a given threshold.[7]

For LSTM-based models, we set the dropout to 0.2 to avoid overfitting, and the number of hidden LSTM units was set to 300. We use Adam Optimizer with a learning rate of 10^{-5} for 30 epochs. The batch size was 64. We used our validation F_1 score as an early stopping criterion. Training stopped if we reached the maximum number of epochs or if the score did not increase for 20 epochs.

For our transformer-based models, the best parameters we found were with a batch size of 32, a maximum sequence length of 128, the maximum predictions per sequence of 20, and a learning rate of 10^{-5} using the Adam Optimizer.

We performed our experiments five times (using five different random seeds) and reported average performance except when we validated the SemEval scores, as we were interested in validating the correctness of our implementation.

7 Results

First, we reran our models on the SemEval data set to ensure that our implementation was correct. We then evaluated our models in our data sets. Finally, we present our findings of the similarities between opinion-ATTITUDE$_{\text{Su}}$ words in APPRAISAL and opinion words in ABSA tasks.

7.1 Validating SemEval Scores

Here, we performed a sanity check on the correctness of our implementation. We applied the six models to the aspect extraction task on the SemEval data. We chose aspect extraction because these were the scores that all of the papers reported, making it a fair basis on which to perform our comparisons. Validating for all data sets requires tremendous computing resources; therefore, we scoped our sanity check on the *Res14* data set. The results are reported in Table 2.

We compared the performance of the LSTM-based models using accuracy, since that was the metric that the original authors used. However, we compared the other models using F_1 because that was the metrics used by the original authors of these models. In all cases, except for TD-BERT, our performance is slightly lower than the performance published by the original authors.

This is not unexpected, as the implementations we use come from the pyTorch library, so they might be slightly different from those of the original authors who would have implemented the systems themselves. This difference in implementation may introduce subtle differences in performance that could easily account for a few percent of the final score. Furthermore, we do not have the same hardware setup as the original authors, which is also known to affect the final performance of machine learning (Crane, 2018). Although we are using transfer learning in transformer networks, the order of operations, the GPU, and the accuracy of the numerical representation all play a role in the final performance. We expect that this might explain a few percent differ-

[7] We set the threshold to 0.5, noting that the probabilities do not need to add to 1.0

ence in the final results.

Nevertheless, from the results in Table 2, it is reasonable to believe that our implementations are sound because the performance is close to that reported in the original papers. Our paired *t*-test did not show statistically significant differences at the $p < 0.05$ level, so we are confident that our implementations are valid.

7.2 Effectiveness on extracting JUDGEMENT and APPRECIATION words

We evaluated the effectiveness of our LSTM-based models and transformer-based models in identifying $\text{ATTITUDE}_{\text{Su}}$ words from *Bio*, *Psyc* and *Psyc$_a$*. We present our scores in Table 3.

Across the three data sets that we evaluated, we have observed that the data set on which our models were trained played an essential role in terms of the F_1 scores we obtained. For example, across the six models, we can observe that using a trained *Lap14* results in poor performance in the *Psyc* and *Psyc$_a$* data sets. The poor performance could be explained by the fact that the vocabularies used in *Psyc$_a$* differ from *Lap14*. On the other hand, we can see that our models perform reasonably well in *Bio* as shown by the F_1 scores on the *Res14*-trained models and *Res15*-trained models. Our visual inspection of the *Res14* and *Res15* data sets found that they contain a mixture of APPRECIATION and JUDGEMENT words which is similar to *Bio*. Therefore, our six models could take advantage of these similarities and perform well in the *Bio* data set.

The baseline, SO-CAL, does not perform well compared to machine learning models. This could be due to the use of a lexicon. By their very nature, lexicons are domain-specific, and if the source domain does not match the domain of the data set, then performance can be expected to be impacted. Closer inspection shows that about 39% of the opinion-$\text{ATTITUDE}_{\text{Su}}$ phrases used in the *Bio* data set is in the SO-CAL lexicon, and about 21% of the opinion-$\text{ATTITUDE}_{\text{Su}}$ phrases in the *Psyc* data set are in the lexicon.

We find that lexicon based are more susceptible to ambiguity. For example, in the sentence from *Bio*, *'It was lovely of them to help me'*, and for the word *'lovely'*, SO-CAL gave an AFFECT score of 0.60 and a JUDGEMENT score of 0.48; and so incorrectly classified the word. In this case, the context of the sentence is essential for a correct classification. All LSTM and transformer models correctly identified this context and correctly classified the word. We have also observed that the NB Classifier's performance is comparable to SO-CAL. We hypothesise that if we further expand the vocabularies in SO-CAL from our training data set, the performance of SO-CAL could be further improved.

RAM was the best of the LSTM-based models. Although we did not find statistically significant differences between the LSTM-based models when we performed a one-way ANOVA ($p < 0.05$), we believe that incorporating the potential opinion-$\text{ATTITUDE}_{\text{Su}}$ word in its memory slices allowed the RAM model to understand the nuances of sentences, even if the potential words are far away. In TC-LSTM, the incorporation of target information in each step during training further reduces the scores compared to not using it in TD-LSTM. TD-LSTM, on the other hand, was a little chaotic. The chaotic behaviour could be due to how the opinion-$\text{ATTITUDE}_{\text{Su}}$ words are located further away in the sentence. We cannot be sure, as the data set on which we evaluated our models was small.

Regarding the transformer-based approach, the best-performing model is BARTABSA: *Bio* (F_1 = 0.623), *Psyc* (F_1 = 0.414) and *Psyc$_a$* data set (F_1 = 0.436), suggesting that the sequence-to-sequence paradigm and the use of BART are an accurate way of extracting opinion-$\text{ATTITUDE}_{\text{Su}}$ phrases. BARTABSA substantially outperformed our baseline, SO-CAL, scoring more than double on all metrics we used. As for the other transformer models (TD-BERT and BERT-AEN), we find that the performance of these models is similar; in particular, we were impressed by TD-BERT's performance, as the performance is comparable to a more complex transformer-based approach (BERT-AEN). We then performed a paired *t*-test, which did not show statistically significant differences at the $p < 0.05$ level between these two models.

Our results suggest that positional information helps BARTABSA achieve strong performance. Further improvements in BARTABSA might be possible by incorporating Part-of-Speech (PoS) information. Su and Hunston (2019) demonstrated that JUDGEMENT and APPRECIATION could be identified by their adjective patterns, including the prepositions or clauses that follow after the word

Model	*Bio* (*Lap14*)	*Psyc* (*Lap14*)	$Psyc_a$ (*Lap14*)	*Bio* (*Res14*)	*Psyc* (*Res14*)	$Psyc_a$ (*Res14*)	*Bio* (*Res15*)	*Psyc* (*Res15*)	$Psyc_a$ (*Res15*)
NB	0.101 ± 0.000	0.084 ± 0.000	0.095 ± 0.000	0.188 ± 0.000	0.104 ± 0.000	0.110 ± 0.000	0.164 ± 0.000	0.124 ± 0.000	0.116 ± 0.000
SO-CAL	0.143 ± 0.000	0.122 ± 0.000	0.145 ± 0.000	0.224 ± 0.000	0.148 ± 0.000	0.160 ± 0.000	0.228 ± 0.000	0.144 ± 0.000	0.155 ± 0.000
TD-LSTM	0.428 ± 0.144	0.244 ± 0.112	0.210 ± 0.123	0.528 ± 0.132	0.410 ± 0.135	0.402 ± 0.118	0.468 ± 0.152	0.344 ± 0.153	0.360 ± 0.145
TC-LSTM	0.401 ± 0.202	0.232 ± 0.310	0.298 ± 0.225	0.501 ± 0.199	0.406 ± 0.194	0.398 ± 0.205	0.456 ± 0.188	0.332 ± 0.198	0.358 ± 0.205
RAM	0.450 ± 0.168	0.291 ± 0.197	0.197 ± 0.188	0.548 ± 0.158	0.461 ± 0.174	0.397 ± 0.181	0.492 ± 0.155	0.365 ± 0.158	0.367 ± 0.176
TD-BERT	0.487 ± 0.144	0.315 ± 0.157	0.341 ± 0.169	0.617 ± 0.135	0.412 ± 0.124	0.422 ± 0.143	0.547 ± 0.142	0.399 ± 0.149	0.382 ± 0.140
BERT-AEN	0.504 ± 0.153	0.323 ± 0.170	0.359 ± 0.221	0.618 ± 0.161	0.408 ± 0.175	0.416 ± 0.152	0.564 ± 0.173	0.381 ± 0.162	0.378 ± 0.146
BARTABSA	**0.598** ± **0.185**	**0.364** ± **0.189**	**0.386** ± **0.198**	**0.623** ± **0.196**	**0.414** ± **0.182**	**0.436** ± **0.185**	**0.588** ± **0.185**	**0.403** ± **0.199**	**0.394** ± **0.181**

Table 3: F_1 scores (with standard deviation) of the models evaluated on *Bio* and *Psyc* when trained on *Lap14*, *Res14* and *Res15* data set. BARTABSA is the best-performing model across all three data sets (highlighted in **bold**).

Model	Bio_{opi} (*Lap14*)	$Psyc_{opi}$ (*Lap14*)	Bio_{opi} (*Res14*)	$Psyc_{opi}$ (*Res14*)	Bio_{opi} (*Res15*)	$Psyc_{opi}$ (*Res15*)
RAM	0.446 ± 0.215	0.297 ± 0.198	0.562 ± 0.232	0.487 ± 0.224	0.506 ± 0.000	0.388 ± 0.000
BARTABSA	**0.582** ± **0.153**	**0.384** ± **0.146**	**0.663** ± **0.145**	**0.448**± **0.138**	**0.592** ± **0.000**	**0.457** ± **0.000**

Table 4: F_1 scores of the best-performing models (with standard deviation) evaluated on Bio_{opi} and $Psyc_{opi}$ when trained on *Lap14*, *Res14* and *Res15* data sets. The best-performing model is highlighted in **bold**.

(for example, if an adjective is followed by a *that* clause, it is likely to be JUDGEMENT). We leave the investigation of PoS in BARTABSA for future work.

7.3 Are ATTITUDE$_{\text{Su}}$ and Opinion similar?

Our above findings do not yet provide a clear indicator of whether opinion-ATTITUDE$_{\text{Su}}$ in APPRAISAL tasks and opinions in ABSA tasks are the same. To accurately determine whether they are similar, we then asked three annotators (two undergraduates and a postgraduate) to re-annotate the *Bio* and *Psyc* data set by following ABSA Opinion extraction guidelines. We will refer to these newly annotated data sets of *Bio* and *Psyc* data sets as Bio_{opi} and $Psyc_{opi}$. As a guideline, we provide samples from SemEval tasks with randomly selected examples from the training portion of the SemEval data set. These were the same samples that Wang et al. (2016) used to annotate the SemEval data set.

We present our findings in Table 4. It is noticeable here that the scores we obtained are similar to the scores we reported in Table 3. Observing only the scores would make it difficult to quantify opinion-ATTITUDE$_{\text{Su}}$, so we needed to perform a statistical analysis to determine whether opinion-ATTITUDE$_{\text{Su}}$ and opinion in ABSA are the same. We first performed a pair chi-square test by comparing the performance of BARTABSA, that is trained on the *Res14* data set, at identifying opinion words in *Bio* and Bio_{opi}. We then proceeded to rerun the same test on the different models (i.e., BARTABSA trained on the *Res15* data set) but evaluated on the same pair of data sets (i.e., Bio and Bio_{opi}). We then repeated the same test on different combinations of models and with *Psyc* and $Psyc_{opi}$.

The analysis did not show statistically significant differences between opinion-ATTITUDE$_{\text{Su}}$ and opinion-ABSA in all combinations at the $p < 0.05$ level. Although our finding of no statistical significance supports the argument of Su and Hunston (2019) that opinion bearing words are a combination of JUDGEMENT and APPRECIATION, we cannot be sure that this would always be the case. This is because our data set is too small to draw a solid conclusion, so we cannot be certain that our findings are applicable on other APPRAISAL data sets.

Annotating a large APPRAISAL data set from scratch can be challenging due to the costs of linguists needed for the process (Snow et al., 2008; Lease, 2011). We suggest that this problem can be addressed by using the SemEval data set as a base and annotate the opinions following the APPRAISAL taxonomy.

8 Conclusion & Limitations

In this work, we investigated whether JUDGEMENT and APPRECIATION branches of the APPRAISAL taxonomy and opinion in Aspect-Based Sentiment Analysis (ABSA) tasks are similar. We use existing ABSA data sets and models to evaluate on two publicly available APPRAISAL data sets. Our empirical results show that there are similarities between the two tasks. Our proposed methodology needs to be carefully tested when reapplied: we were only able to perform experiments on small data sets. Secondly, we focus on the JUDGEMENT and APPRECIATION branches of APPRAISAL, although it would be interesting to see if we could use triplet-extraction task from ABSA. We hope that our work here could motivate Systemic Functional linguists community and NLP community to work together.

References

Surabhi Adhikari et al. 2020. NLP Based Machine Learning Approaches for Text Summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 535–538. IEEE.

Shlomo Argamon, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2007. Automatically Determining Attitude Type and Force for Sentiment Analysis. In *Human Language Technology. Challenges of the Information Society, Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 5-7, 2007, Revised Selected Papers*, volume 5603 of *Lecture Notes in Computer Science*, pages 218–231. Springer.

Kenneth Bloom and Shlomo Argamon. 2010. Unsupervised Extraction of Appraisal Expressions. In *Advances in Artificial Intelligence, 23rd Canadian Conference on Artificial Intelligence, Canadian, AI 2010, Ottawa, Canada, May 31 - June 2, 2010. Proceedings*, volume 6085 of *Lecture Notes in Computer Science*, pages 290–294. Springer.

Muxuan Chen. 2022. An Appraisal Analysis of Sina Weibo Texts about Reforms of Undergraduate Education. *Scientific and Social Research*, 4(1):1–16.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 452–461. Association for Computational Linguistics.

Matt Crane. 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. *Trans. Assoc. Comput. Linguistics*, 6:241–252.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Hai Ha Do, P. W. C. Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Syst. Appl.*, 118:272–299.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-Dependent Sentiment Classification With BERT. *IEEE Access*, 7:154290–154299.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In *EMNLP 2006, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia*, pages 355–363. ACL.

Hee-Gon Kim, Do-Young Lee, Se-Yeon Jeong, Heeyoul Choi, Jae-Hyung Yoo, and James Won-Ki Hong. 2019. Machine Learning-Based Method for Prediction of Virtual Network Function Resource Demands. In *2019 IEEE Conference on Network Softwarization (NetSoft)*, pages 405–413. IEEE.

Matthew Lease. 2011. On Quality Control and Machine Learning in Crowdsourcing. In *Human Computation, Papers from the 2011 AAAI Workshop, San Francisco, California, USA, August 8, 2011*, volume WS-11-11 of *AAAI Technical Report*. AAAI.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jun Liu, Tong Zhang, Guangjie Han, and Yu Gou. 2018. TD-LSTM: Temporal Dependence-Based LSTM Networks for Marine Temperature Prediction. *Sensors*, 18(11):3797.

Shaiq Munir Malik, Fnu Mohbat, Muhammad Umair Haider, Muhammad Musab Rasheed, and Murtaza Taj. 2021. Teacher-Class Network: A Neural Network Compression Mechanism. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 58. BMVA Press.

J. R. Martin and P. R.R. White. 2005. *The Language of Evaluation*. Palgrave/Macmillan.

Diego Mollá. 2020. Overview of the 2020 ALTA Shared Task: Assess Human Behaviour. *ALTA 2020*, page 127.

Diego Mollá. 2021. Overview of the 2021 ALTA Shared Task: Automatic Grading of Evidence, 10 years later. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 201–204.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of Affect, Judgment, and Appreciation in Text. In *COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China*, pages 806–814. Tsinghua University Press.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2022. Reproducibility and automation of the appraisal taxonomy. *Proceedings of the 29th International Conference on Computational Linguistics*.

Jasabanta Patro, Srijan Bansal, and Animesh Mukherjee. 2019. A Deep-Learning Framework to Detect Sarcasm Targets. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6335–6341. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel, Salud María Jiménez Zafra, and Gülsen Eryigit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 19–30. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495. The Association for Computer Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Andrew S. Ross and David Caldwell. 2020. 'Going negative': An APPRAISAL analysis of the rhetoric of Donald Trump on Twitter. *Lang. Commun.*, 70:13–27.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional Encoder Network for Targeted Sentiment Classification. *CoRR*, abs/1902.09314.

Susan Standing and Craig Standing. 2018. The Ethical Use of Crowdsourcing. *Business Ethics: A European Review*, 27(1):72–80.

Sue Starfield, Brian Paltridge, Robert McMurtrie, Allyson Holbrook, Sid Bourke, Hedy Fairbairn, Margaret Kiley, and Terry Lovat. 2015. Understanding the Language of Evaluation in Examiners' Reports on Doctoral Theses. *Linguist. Educ.*, 31:130–144.

Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 341:1–341:14. ACM.

Hang Su and Susan Hunston. 2019. Language patterns and attitude revisited: Adjective patterns, Attitude and Appraisal. *Functions of Language*, 26(3):343–371.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Comput. Linguistics*, 37(2):267–307.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for Target-Dependent Sentiment Classification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3298–3307. ACL.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615. The Association for Computational Linguistics.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using Appraisal Groups for Sentiment Analysis. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 625–631. ACM.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A Unified Generative Framework for 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2416–2429. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *CoRR*, abs/2203.01054.