Analyzing Subjectivity Using a Transformer-Based Regressor Trained on Naïve Speakers' Judgements

Elena Savinova and **Fermín Moscoso del Prado Martín**
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
{elena.savinova, fermin.moscoso-del-prado}@ru.nl

Abstract

The problem of subjectivity detection is often approached as a preparatory binary task for sentiment analysis, despite the fact that theoretically subjectivity is often defined as a matter of degree. In this work, we approach subjectivity analysis as a regression task and test the efficiency of a transformer RoBERTa model in annotating subjectivity of online news, including news from social media, based on a small subset of human-labeled training data. The results of experiments comparing our model to an existing rule-based subjectivity regressor and a state-of-the-art binary classifier reveal that: 1) our model highly correlates with the human subjectivity ratings and outperforms the widely used rule-based *pattern* subjectivity regressor (De Smedt and Daelemans, 2012); 2) our model performs well as a binary classifier and generalizes to the benchmark subjectivity dataset (Pang and Lee, 2004); 3) in contrast, state-of-the-art classifiers trained on the benchmark dataset show catastrophic performance on our human-labeled data. The results bring to light the issues of the gold standard subjectivity dataset, and the models trained on it, which seem to distinguish between the origin/style of the texts rather than subjectivity as perceived by human English speakers.

1 Introduction

The task of subjectivity detection refers to identifying opinions, attitudes, beliefs and private states in a given text. Subjectivity detection as a task has received a lot of attention over the past decades, resulting in an abundance of methods and tools for subjectivity analysis. While in the earlier works, subjectivity was detected using rule-based approaches employing subjectivity lexicons (Riloff et al., 2003; Wiebe and Riloff, 2005), in the more recent studies, subjectivity detection is often approached with machine learning classifiers trained on existing gold standard datasets annotated for subjectivity (Huo and Iwaihara, 2020; Zhao et al., 2015).

Despite a relatively large body of literature on the topic, subjectivity detection has often been perceived as a preparatory step for sentiment analysis, that is, detection of positive or negative polarity of texts (Chaturvedi et al., 2018; Liu, 2010). Accurate sentiment analysis relies primarily on subjective fragments of the text. For this reason, subjectivity detection has been mostly viewed as a binary classification task. However, given the complex nature of the interplay of viewpoints in texts and numerous ways of expressing oneself with varying intensity, subjectivity can also be considered a gradual measure. To the best of our knowledge, the most widely used tool for subjectivity analysis that offers a gradual subjectivity estimate is *pattern* regressor (De Smedt and Daelemans, 2012), which is rule-based, and therefore, less accurate compared to state-of-the-art systems.

In the present paper, we approach the problem of identifying subjectivity as a regression task and use a semi-supervised approach to train a task-agnostic transformer model (RoBERTa) to produce sentence-level subjectivity scores based on a small subset of human annotations. The data that we use for training are a dataset of news articles and social media news posts produced by major UK news sources, with a small subset of it labeled by native English speakers. We describe the training procedure and compare the results of the model to the average native speaker's judgements, and to the widely used rule-based *pattern* regressor (De Smedt and Daelemans, 2012) to demonstrate that our model converges with native speaker intuitions and outperforms the rule-based regressor. The model is further evaluated as a binary classifier on our dataset and on the benchmark subjectivity dataset (Pang and Lee, 2004), showing good performance and generalizability to other discourse types. Finally, we show that our model generalizes better

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 305–314

to other discourse types than the current state-of-the-art systems trained on the gold standard subjectivity dataset (Pang and Lee, 2004). These results highlight the importance of relying on actual human annotations rather than automatic labeling for compilation of subjectivity datasets and open further discussion about the nature of subjectivity models trained on the gold standard dataset.

2 Related work

The problem of sentence-level subjectivity detection in the previous literature has been approached in two ways. The traditional approach is rule-based and uses subjectivity lexicons and linguistic pattern extraction to define subjective and objective text fragments (Riloff and Wiebe, 2003; Riloff et al., 2003; Wiebe and Riloff, 2005; De Smedt and Daelemans, 2012). A more recent state-of-the-art approach uses machine learning based text classification algorithms to detect subjectivity (e.g., Hube and Fetahu, 2019; Huo and Iwaihara, 2020; Lin et al., 2011; Sagnika et al., 2021; Zhao et al., 2015). Although training deep neural models can give impressive results, they require large annotated datasets and substantial computational resources, which are not always available. To overcome this issue, recent studies on subjectivity detection started employing pretrained language representation transformer models, such as BERT (Devlin et al., 2019), and fine-tuning them for subjectivity classification task, which showed very promising results (Huo and Iwaihara, 2020; Kasnesis et al., 2021; Pant et al., 2020). For instance, Kasnesis et al. 2021 repot an impressive accuracy of 98.3% in subjectivity detection in the benchmark subjectivity dataset (Pang and Lee, 2004) using a method based on ELECTRA-large transformer (Clark et al., 2020).

In the present work, we adopt the approach of using a pretrained task-agnostic language model and fine-tuning it on subjectivity detection task for our own news dataset. The dataset consists of Facebook news posts and online news articles produced by four major UK news sources, with the total size of 7,751 sentences. Subjectivity of utterances is known to be a gradual factor in the cognitive theoretical accounts of subjectivity (Langacker, 1990; Traugott, 1995); some utterances are perceived as more subjective than others. For this reason, in contrast to many previous studies on automatic subjectivity analysis, we approached the problem of subjectivity detection as a regression rather than a binary classification problem. To do so, we trained our model on a subset of our data that have been annotated for the degree of subjectivity by human raters using a 7-point scale.

In the previous literature, benchmark datasets used for training subjectivity detection models were often obtained via annotations based on certain guidelines. For example, one of the earliest resources, namely MPQA Opinion Corpus (Wiebe et al., 2005), was compiled following a precise event- and entity-level annotation scheme for what is considered subjective and objective. In languages other than English, e.g., in Italian corpus *subjectivITA* (Antici et al., 2021), sentence-level subjectivity annotations were also obtained by asking annotators to follow specific guidelines on what should or should not be considered subjective. Although the guidelines are often theory-based, it is difficult to estimate how they relate to the actual native speakers' intuitions. For example, telling annotators to label third person attitudes and beliefs as objective reflects a certain theoretical choice but may not reflect language users' perceptions (e.g., *"According to the guests, the show was extremely unprofessional"*). Even more difficult to relate to human judgements are automatically collected subjectivity datasets, such as the benchmark SUBJ dataset (Pang and Lee, 2004), which is a widely used dataset for model training and evaluation. This dataset contains 5,000 movie review snippets that are automatically labeled as subjective and 5,000 sentences from plot summaries that are automatically labeled as objective. However, a closer look at this dataset reveals many cases where objectivity of the sentences taken from the movie plot summaries is questionable (e.g., *"What better place for a writer to pick up a girl?"* is considered as being objective). In the present work, we train our model on subjectivity annotations by native speakers who were not asked to follow any guidelines except for brief definitions of subjective ("expressing opinions, attitudes and beliefs") and objective ("stating factual information"), which means that our model results represent how subjectivity would be perceived by naïve language users. Similar approach to obtaining annotations with only definitions of subjective and objective as guidelines was used in the compilation of a Czech subjectivity dataset (Přibáň and Steinberger, 2022).

3 Method

3.1 Dataset

The dataset contains articles and Facebook posts on the topics of "crime" and "Covid-19" by four major UK news sources: two "popular" newspapers focused on soft news content (*Daily Mail* and *Metro*) and two "quality" sources focused on hard news (*BBC News* and *Sky News*). Since the dataset was collected for the purpose of analyzing subjectivity in the news across different types of sources (quality and popular) and media channels (articles on the websites and Facebook posts), the topics of "crime" and "Covid-19" were chosen to ensure comparability between the popular and the quality sources, as these topics are covered by both types of sources. First, we acquired Facebook posts of the four sources using Facepager app (Jünger and Keyling, 2019). Around 2,000 posts per news source page were randomly selected, setting the app limits to 100 posts per page, 20 pages and a one-year time limit, meaning the posts were published between December 7, 2020 and December 7, 2021. The information collected included the text of the post, the news headline (if present), the link to the original news item (if present), date and time of publication. Topic selection for "crime" and "Covid-19" topics was performed using keywords (see Appendix A). During topic selection for crime news, several keywords for exclusion of items were used to make sure that the news items did not include stories about natural disasters or TV soap operas. At the preprocessing stage, the posts were split into sentences using the Python package NLTK (Bird et al., 2009).

After the Facebook posts were selected, 84 news articles were randomly chosen out of those posts that had a corresponding link to the original news item (21 articles per source, equal distribution of topics), in an attempt to match the articles and the Facebook posts subparts of the dataset in the number of words. The text of the news articles was scraped from the websites (using the Python package *beautifulsoup*; Richardson, 2007). The articles were also preprocessed and split into sentences using NLTK (Bird et al., 2009). The resulting dataset includes 4,778 sentences taken from Facebook posts (65,058 words) and 2,973 sentences taken from articles on the websites (72,236 words), including headlines in both cases.

3.2 Annotation experiment

A random subset of 400 sentences (controlled for equal distribution of topics, news sources and media channels) was selected from the dataset for the annotation experiment. We used Prolific to collect annotations from 20 native speakers of English. We semi-randomly split the subset of 400 sentences into 20 experimental lists matching the number of annotators, in such a way that every annotator received 100 sentences for labeling and every sentence was labeled by 5 different speakers. The participants were instructed to evaluate subjectivity of the sentences on a 7-point scale, with extremes marked as "objective" and "subjective". They were informed that the sentences were taken from news articles on the newspaper websites and news posts on social media, and that some sentences are headlines. The participants were given simple conceptual definitions of the terms, namely, they were told that "subjective" meant "expressing personal opinions, emotions, feelings and tastes, hopes and wishes, self-made conclusions (e.g., "This is awful")", while "objective" meant "reporting facts, events, conclusions supported by data (e.g., "The President had a meeting with the Prime minister")". There were four attention checks asking participants to select a specific answer option and four comprehension checks representing clearly subjective ("This is very beautiful") and objective ("London is the capital of the UK") sentences that were expected to be rated with 7 and 1, respectively. Only those participants who passed all the attention checks were paid for participation (4 GBP) and only those who also passed the comprehension checks were included in the dataset. One participant failed to pass the attention checks; additionally, two participants failed to pass the comprehension checks. After rejecting a participant, their list was reposted to Prolific until all 20 lists were successfully annotated. The mean age of participants in the final dataset was 36 (SD=15, range 19-67). The experiment was approved by the Ethics Assessment Committee Humanities of Radboud University (reference number 2022-9393).

Since our participants each rated a different subset of 100 sentences from all other participants, in order to estimate the inter-rater agreement, we computed the correlation of each participant's ratings with the mean of the remaining participant's ratings. We chose a correlation score of r=.4 as an inclusion threshold, leaving out one participant whose score

was lower than .4. The mean correlation score of the remaining 19 raters was r=.64. We also excluded two sentences from the annotated subset as those were discovered to be duplicates (although in the full dataset these sentences come from different news items, they share the same text: "BREAKING"). For the remaining 398 sentences rated by 19 subjects, we computed mean scores and standard deviations. Figure 1 shows the distribution of standard deviations over the scores: as expected for this type of data, the more extreme scores have smaller deviations since people tend to agree on what is clearly subjective and objective, while the scores towards the middle have larger deviations reflecting weaker agreement among raters.

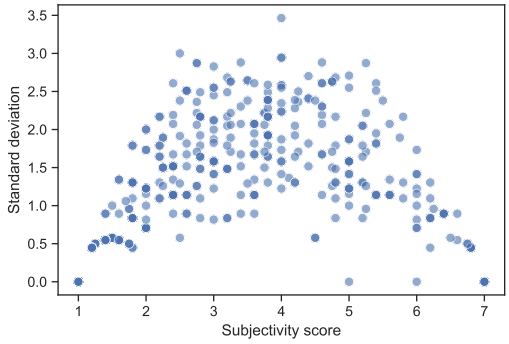

Figure 1: The distribution of standard deviations over mean subjectivity scores for annotated sentences.

3.3 Model training

In order to improve the performance of our text classifier/regressor, we began by fine-tuning the robustly optimized BERT transformer RoBERTa-base (Liu et al., 2019), which is based on dynamic masked priming, by adapting it to the unlabeled part of our dataset using the *simpletransformers* library (Rajapakse, 2019). Doing this for just a single epoch provided a small improvement in the final regression/classification results. Subsequently, we trained a text regression model on the labeled subset of our data using our version of RoBERTa-base fine-tuned to our specific dataset. We split the labeled data into a training set (298 sentences), a validation set (50 sentences) and a test set (50 sentences). For each of the 298 sentences, the model was trained to produce an average rating for that sentence provided by the human annotators. The human rating scores were normalized from the 7-point scale into a [0-1] scale. The convergence of the evaluation loss indicates that 20 epochs are sufficient training for this model (Figure 2). After testing the model's performance, it was applied to the full dataset to obtain subjectivity scores per sentence. The dataset with sentence-level subjectivity scores predicted by our model is available online.

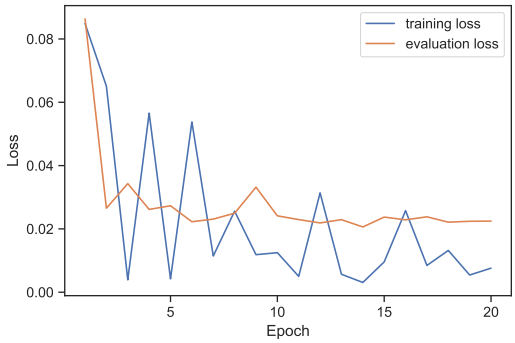

Figure 2: Training and evaluation loss.

4 Results

4.1 Evaluating the model as a text regressor

Our model's predictions on the test set show that there is a very high correlation with the average human ratings (r=.79), accounting for over 62% of the variance. Figure 3 shows a plot of the correlation between the model's prediction and the true human ratings. Beyond being quite a good correlation, this is above the correlation achieved by any of the raters with the average of the remaining raters (the maximum achieved by the raters was r=.76, the average correlation was r=.64, and the median correlation was r=.67). In other words, our model is a better match to the average human rater than any of the human raters was.

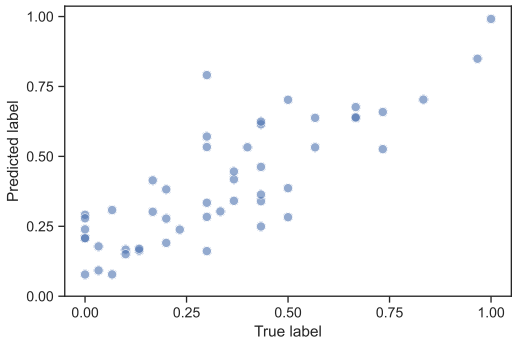

Figure 3: Correlation between our model's predictions and the human ratings on the test set.

For comparison, we provide subjectivity annotation with TextBlob (using *SpacyTextBlob*; Loria,

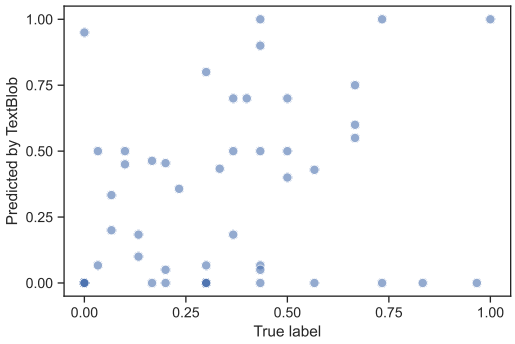

Figure 4: Correlation between *pattern*'s (TextBlob) predictions and the human ratings on the test set.

2018) as a baseline performance. TextBlob uses the rule-based sentiment valence and subjectivity tagger from the *pattern* library described in De Smedt and Daelemans (2012). This is one of the most popular sentiment analyzers for English. In addition to sentiment, it provides subjectivity ratings on a scale from 0.0 (totally objective) to 1.0 (totally subjective). We evaluate *pattern*'s performance on our test set by comparing *pattern*'s subjectivity ratings with the average ratings provided by our annotators. The correlation plot is presented in Figure 4. As we can see from the plot, *pattern*'s predictions correlate very poorly with human raters (r=.28), accounting for barely 8% of the variance in the rating means. Thus, our model substantially outperforms a widely used tool for subjectivity annotation based on regression. To give an example for comparison, a sentence *"Rose West was convicted of 10 murders in November 1995 and is serving life"* from our test set received an average score of 0 by the human annotators (which corresponds to 1 on the 7-point scale), suggesting that it was perceived as objective. While our model predicted a similar subjectivity score of .08 for this sentence, the *pattern* regressor estimated it as very subjective with a score of .95. From the above, we can conclude that using our model for tagging a corpus should result in an annotation that would be at least as good as an annotation that would be obtained if a single person rated all sentences for subjectivity.

4.2 Additional test set

In order to test our model's performance further, we collected an additional human-labeled test set by randomly selecting 100 sentences from the unlabeled part of our dataset. We obtained subjectivity ratings for theses sentences from 5 native English speakers (M=29, SD=7, range 19-35) using the same procedure as in the first annotation experiment (Section 3.2). Comparing each participant's ratings and their correlation with the mean of the other participants' ratings led to exclusion of one outlying participant, whose correlation with the others was below .4 threshold. The mean correlation score of the remaining participants was r=.60. Since all participants rated the same set of items in this experiment, we also computed an intraclass correlation coefficient (ICC) to estimate inter-rater agreement between participants. The ICC estimate at 95% confidence level using a two-way random-effects model (using the *pingouin* Python package; Vallat, 2018) was .41 for a single rater, suggesting fair bordering with poor reliability of raters as individuals. The ICC estimate for the average of raters was .74 indicating moderate bordering with good inter-rater agreement. The correlation between the average human ratings and our model's predictions are presented in Figure 5. The correlation score was r=.61, explaining 37% of the variance. Although this score is lower than the score obtained for the original test set, it is just above the mean correlation between the raters of this additional set (r=.60). Therefore, as in the previous tests, our model is indeed a very good model of the average human rater. Notice that one should not expect the model to show very high correlation scores with the raters' evaluations when the raters themselves do not agree on the evaluation of these sentences, as is indicated by the low inter-rater agreement scores. Further research is needed to investigate whether there are natural clusters among the raters which would imply that there are different ways of understanding what subjectivity is among English speakers. We believe that the performance of our model on the additional test set is not surprising given the low level of agreement among human raters themselves, and together with relatively good performance on the benchmark dataset (see Section 4.4 below), indicates that our model is a good subjectivity predictor.

4.3 Evaluating the model as a binary classifier

Although we have trained the model as a text regressor, it can also be used as a classifier, by discretizing the continuous scores on one or more thresholds for both the true and the predicted labels. Given that the human annotators were instructed to rate subjectivity on a 7-point scale, it is clear

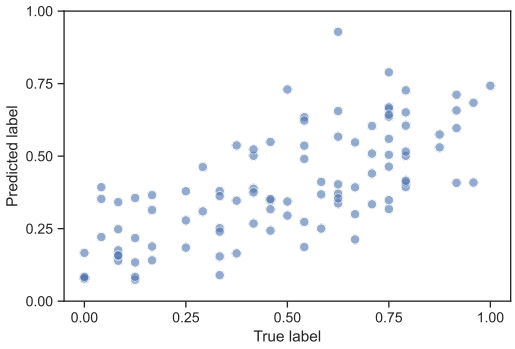

Figure 5: Correlation between our model's predictions and the human ratings on the additional test set.

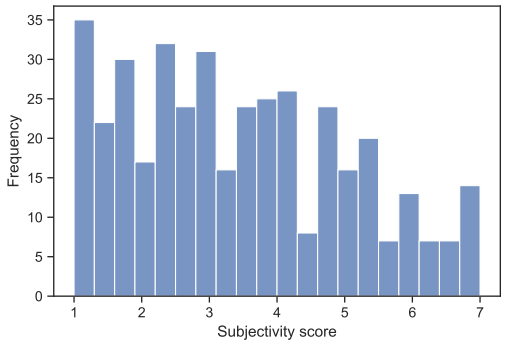

Figure 6: The distribution of subjectivity scores per sentence in the labeled set.

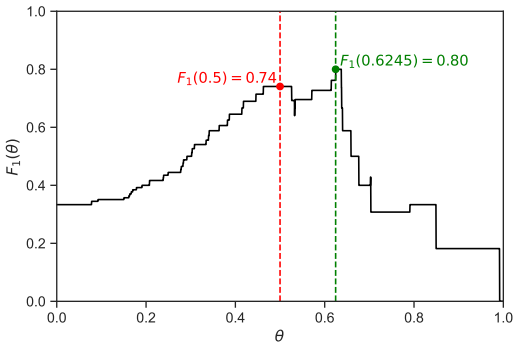

Figure 7: Threshold value for the binary classifier as a function of F1 score.

that anything with an average rating above .5 (i.e., above 4 on the 7-point scale), was considered as more subjective than objective by the annotators, and therefore, human-labeled data can be split on the .5 threshold. On the other hand, our model was trained on minimizing the mean squared error between predictions and true labels, and not on classification. Furthermore, the data in the training set were unbalanced towards objective labels (Figure 6). For this reason, one might want to consider a threshold different from .5 for taking a prediction of the model as subjective. We found the optimal threshold value for the model output by considering the model's predictions and optimizing the value of the F1 score (for the minority category "subjective") as a function of the threshold value. The result of this optimization can be seen in Figure 7. It shows that taking a classification threshold of θ=.6245, leads to the optimal classification behavior with an impressive accuracy of 92% and F1=.80. This is a slight improvement over the classification that would be obtained by a plain .5 threshold, which still leads to a very good classification performance with an accuracy of 86% and F1=.74. The overall performance of our model as a classifier can be appreciated in the diagonal confusion matrix (Figure 8).

4.4 Evaluation on the benchmark subjectivity dataset

As mentioned above, the most used dataset for evaluating subjectivity labels is the SUBJ dataset introduced in Pang and Lee (2004). This dataset contains 10,000 short texts. Of these, 5,000 – automatically labeled as subjective – are movie review snippets (e.g., *"bold, imaginative, and impossible to resist"*) from www.rottentomatoes.com. The remaining 5,000 – automatically labeled as objective – are sentences from plot summaries taken from the Internet Movie Database (IMDb). This dataset is generally taken as a gold standard for subjectivity. However, although there is a clear correlation between subjectivity and the source of the text (review snippets being subjective vs. plot sentences being objective), one can find many examples in the dataset where this assumption fails. For instance, the IMDb sentence: *"What better place for a writer to pick up a girl?"* is labeled as objective in the SUBJ dataset, but the objectivity of this sentence is rather questionable. In all fairness, such a sentence might indeed have been objective in the context of the plot summary in which it appeared but, without such context, as it appears in the dataset, it is less clearly objective. Such examples of not-so-objective IMDb sentences abound in the SUBJ dataset. The opposite, however, is less common, with the snippets taken from www.rottentomatoes.com appearing consistently subjective, at least on visual inspection.

Comparing a variety of traditional (i.e., non-

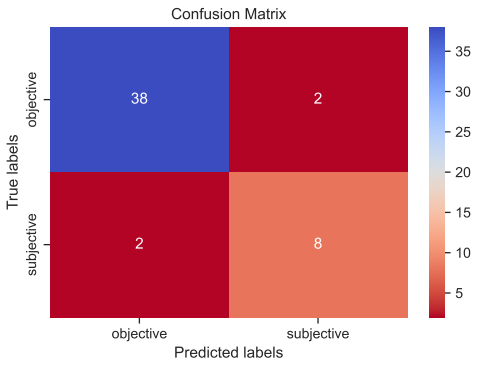

Figure 8: Confusion matrix of our model on the test set.

deep-learning) methods, Wang and Manning (2012) report maximum accuracy of 93.6% in tagging this corpus. In a recent review, Kasnesis et al. (2021) raise this maximum accuracy to 98.3% using a method based on the ELECTRA-large transformer. Our text regression model was trained on a completely different type of texts. Such texts were also used for setting the classification threshold. Despite the mislabeling present in the SUBJ dataset, it is still interesting to evaluate how our model performs on the test set of the SUBJ dataset. The density plot of the predicted subjectivity scores is presented in Figure 9.

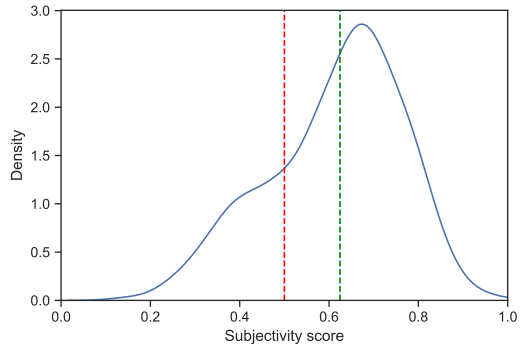

Figure 9: Density plot of our model's predictions on the test set of the SUBJ dataset.

As mentioned above, it appears that many of the "objective" sentences in the SUBJ dataset are in fact more than a bit subjective. The opposite (i.e., rather objective sentences labeled as "subjective") is less common in the dataset. If this intuition is true, and our model captures the actual subjectivity of the texts in the corpus, we should expect to see that the predictions of our model are visibly skewed towards the subjective, instead of being balanced as it is assumed in the dataset design. The kernel density estimate plot of our model's predictions (Figure 9) confirms this intuition: there are substantially more subjectively labeled sentences than one would have expected in a balanced labeled corpus. The green dashed line on the density plot denotes the optimized classification threshold, and the red dashed line plots the suboptimal .5 classification threshold. The slight shoulder on the left side of the density plot is a trace of bimodality. This bimodality arises because, on average, the sentences from movie plots are indeed more objective than the movie review snippets, but this is far from a clear-cut distinction in terms of objectivity.

If we use the classification threshold that we established on our own testing set, without further optimization, we obtain an accuracy of 78.2%, and an impressive F1=.79 on the SUBJ test set, just slightly below what we obtained for our own testing set. This is remarkable, considering that the SUBJ dataset is substantially different from the dataset that we trained our model on. Even if we had chosen to keep the suboptimal classification threshold at .5, we would still obtain an accuracy of 69.8% and a very good F1=.76. Examining the confusion matrix for this dataset with the optimized threshold value (Figure 10), we find that the accuracy especially suffers from cases that were labeled as "objective" in the corpus, but our model in fact considers them subjective. However, if we bear in mind the mislabeling present in SUBJ dataset that we discussed above, these might in fact not be errors, but sometimes cases where our model is actually outperforming the supposed gold standard. For instance, the aforementioned sentence *"What better place for a writer to pick up a girl?"*, which is labelled as objective in the SUBJ dataset, but appears subjective to us, is given a subjectivity score of .66 by our model. Thus, this sentence is evaluated as unclear but slightly subjective by our model. To us this appears to be a better assessment of this sentence's subjectivity than the gold standard label of "objective". To give another example, a sentence *"Moving cross country isn't even a problem for her"* is tagged as objective in the SUBJ dataset. Without the context, this sentence seems to represent an opinion/judgement, which is in essence subjective. Our model's prediction for subjectivity of this sentence is .78, which, in our opinion, is a more accurate estimate.

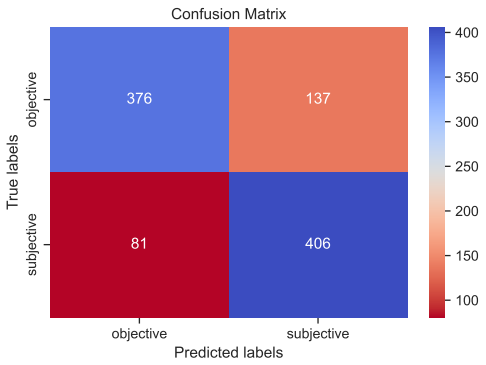

Figure 10: Confusion matrix of our model on the test set of the SUBJ dataset.

4.5 Performance of state-of-the-art classifiers on our data

At an accuracy of 78.2% and an F1 of .79, our system performs relatively well on the SUBJ dataset benchmark. Nevertheless, this performance is well below the top performance (with an accuracy of 98.3%) reported by Kasnesis et al. (2021) for the same dataset. We suspect, however, that the outstanding performance of subjectivity classifiers trained on this dataset is in fact misleading. As we have seen, many of the examples (certainly more than 2%) in this gold standard are actually mislabeled with respect to subjectivity itself. This suggests that such top performing systems, rather than learning to distinguish subjective from objective passages, are in fact learning to distinguish the language used in movie review snippets from the language used in movie plots. The fact that this distinction indeed correlates with subjectivity explains why our system, trained on data explicitly labeled for subjectivity, is still able to perform well on this dataset.

To investigate this further, we trained a two-way classifier (based on a distilBERT-base-uncased transformer) on the 8,100 training passages of the SUBJ dataset, using an additional 900 passages as a validation set. Our system performed slightly below the best reported performances, with an accuracy of 93.5% and F1=.93 on the 1,000 test passages from the SUBJ test set. We did not spend much time optimizing this system because we believe that improving the system's performance would not lead to results much different from those we report below. Once this model was trained, we tested the model on the 50 test sentences from our human-labeled dataset.

The two-way classifier seemed to perform relatively well on the 50 test sentences from our dataset, with an accuracy of 75.5%. However, examining the performance in more detail revealed that on our dataset (which reflects human subjectivity ratings), the model obtained a dismal F1=.25 in classifying subjective sentences. The very low F1 is explained by the confusion matrix below (Figure 11): the model shows more false positives and misses than it shows hits in labeling a sentence as subjective. This confirms our suspicion that the outstanding performance of this model on the SUBJ dataset reflects not the fact that the model is a good classifier of subjectivity, but rather the fact that this model instead learned how to distinguish the language in movie review snippets from that used in movie plot descriptions. Given this finding, it is to be expected that even the top state-of-the-art models reaching accuracies above 98% on the SUBJ dataset, would not succeed in distinguishing what is really subjective from what is really objective.

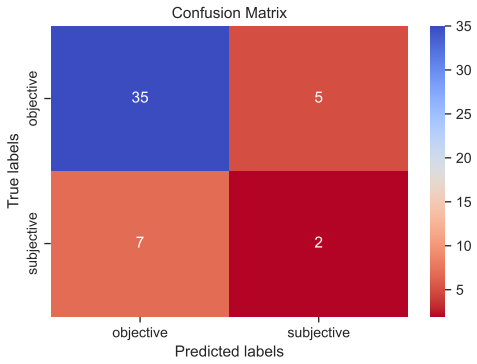

Figure 11: Confusion matrix of the state-of-the-art classifier trained on SUBJ dataset on our human-labeled test set.

5 Conclusion

In this paper, we approached the problem of subjectivity analysis as a regression task and tested the efficiency of transformer language representation models, such as RoBERTa, in annotating subjectivity using a paradigm of semi-supervised approach based on a small subset of human-labeled data. Our model showed a very high correlation with the average human rater and significantly outperformed a widely used rule-based *pattern* subjectivity regressor. The model also performed well as a binary classifier, both on our news dataset and on the benchmark subjectivity dataset exemplifying different discourse types. In contrast, we found

that the state-of-the-art classifiers with best performance on the benchmark dataset and trained on that dataset show catastrophic performance on our human-labeled dataset, which is not very different from the baseline. This means that our model generalizes across domains much better than the current best systems. Moreover, this brings to light the issues of the subjectivity dataset that is considered the gold standard for subjectivity detection task. Rather than labelling for actual subjectivity, the gold standard dataset uses the origin of the texts (movie review snippets vs movie plot descriptions) as a proxy for subjectivity. Although the origin of the text undoubtedly correlates with subjectivity, these distinctions are not the same. As a result, state-of-the-art subjectivity classifiers trained on this dataset might be learning how to distinguish the language of movie review snippets from that of movie plot descriptions, rather than classifying subjectivity, as perceived by native speakers. Future work could further analyze how the performance of state-of-the-art classifiers trained on the benchmark subjectivity dataset compares to human-labeled subjectivity ratings in order to shed light on what exactly these systems are learning. Our work highlights the importance of using human annotations in such complex tasks as subjectivity detection. Future work can also be done in further comparing the performance of systems that are trained on the datasets labeled following explicit theoretical instructions to those trained on naïve human judgements about subjectivity. In addition, future studies on automatic subjectivity detection systems could investigate the origins of the differences in subjectivity perception across native speakers.

References

Francesco Antici, Luca Bolognini, Matteo Antonio Inajetovic, Bogdan Ivasiuk, Andrea Galassi, and Federico Ruggeri. 2021. *SubjectivITA*: An italian corpus for subjectivity detection in newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 40–52. Springer.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media Inc., Sebastopol, CA.

Iti Chaturvedi, Erik Cambria, Roy E. Welsch, and Francisco Herrera. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(66):2063–2067.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 195–203, New York, NY, USA. Association for Computing Machinery.

Hairong Huo and Mizuho Iwaihara. 2020. Utilizing bert pretrained models with various fine-tune methods for subjectivity detection. In *4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II*, pages 270–284. Springer.

Jakob Jünger and Till Keyling. 2019. Facepager. An application for automated data retrieval on the web.

Panagiotis Kasnesis, Lazaros Toumanidis, and Charalampos Z Patrikakis. 2021. Combating fake news with transformers: A comparative analysis of stance detection and subjectivity analysis. *Information*, 12(10):409.

Ronald W Langacker. 1990. *Subjectification*. Walter de Gruyter, Berlin/New York.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th international joint conference on natural language processing*, pages 1153–1161.

Bing Liu. 2010. Sentiment analysis and subjectivity. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of natural language processing*, 2 edition, pages 627–666. CRC Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

Steven Loria. 2018. Textblob documentation.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Kartikey Pant, Tanvi Dadu, and Radhika Mamidi. 2020. Towards detection of subjective bias using contextualized word embeddings. In *Companion Proceedings of the Web Conference 2020*, pages 75–76, New York, NY, USA. Association for Computing Machinery.

Pavel Přibáň and Josef Steinberger. 2022. Czech dataset for cross-lingual subjectivity classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1381–1391, Marseille, France. European Language Resources Association.

Thilina Rajapakse. 2019. Simple transformers.

Leonard Richardson. 2007. Beautiful soup documentation.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 25–32.

Santwana Sagnika, Bhabani Shankar Prasad Mishra, and Saroj K Meher. 2021. An attention-based cnn-lstm model for subjectivity detection in opinion-mining. *Neural Computing and Applications*, 33:17425–17438.

Elizabeth Closs Traugott. 1995. Subjectification in grammaticalization. In D. Stein and S. Wright, editors, *Subjectivity and subjectivisation: Linguistic perspectives*, volume 1, pages 31–54. Cambridge University Press.

Raphael Vallat. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05)*, pages 486–497. Springer.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Han Zhao, Zhengdong Lu, and Pascal Poupart. 2015. Self-adaptive hierarchical sentence model. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 4069–4076. AAAI Press.

A Appendix

Keywords for “Covid-19” news: "pandemic", "epidemic", "covid", "vaccin", "vaxx", "lockdown", "coronavirus", "omicron", "quarantine".

Keywords for “crime” news: "[^a-z]kill", "jail", "arrest", "crime", "murder", "kidnap", "[^a-z]rape", "[^a-z]rapi[^d]", "criminal", "terrorist", "shooting", "homicide", "robbery", "sentenced", "felony", "fraud".

Keywords for exclusion of news items about soap opera and natural disaster: “soaps”, “spoiler”, “storm”, “avalanche”, “volcano”, “lightning”, “tornado”, “flood”.