Subjectivity Theory vs. Speaker Intuitions: Explaining the Results of a Subjectivity Regressor Trained on Native Speaker Judgements

Elena Savinova and **Jet Hoek**
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
elena.savinova@ru.nl, jet.hoek@ru.nl

Abstract

In this paper, we address the issue of explainability in a transformer-based subjectivity regressor trained on native English speakers' judgements. The main goal of this work is to test how the regressor's predictions, and therefore native speakers' intuitions, relate to theoretical accounts of subjectivity. We approach this goal using two methods: a top-down manual selection of theoretically defined subjectivity features and a bottom-up extraction of top subjective and objective features using the LIME explanation method. The explainability of the subjectivity regressor is evaluated on a British news dataset containing sentences taken from social media news posts and from articles on the websites of the same news outlets. Both methods provide converging evidence that theoretically defined subjectivity features, such as emoji, evaluative adjectives, exclamations, questions, intensifiers, and first person pronouns, are prominent predictors of subjectivity scores. Thus, our findings show that the predictions of the regressor, and therefore native speakers' perceptions of subjectivity, align with subjectivity theory. However, an additional comparison of the effects of different subjectivity features in author text and the text of cited sources reveals that the distinction between author and source subjectivity might not be as salient for naïve speakers as it is in the theory.

1 Introduction

Subjectivity analysis is the task of identifying opinions, attitudes, evaluations and beliefs in texts. State-of-the-art approaches to detecting subjectivity at the sentence level (e.g., Huo and Iwaihara, 2020; Kasnesis et al., 2021; Pachov et al., 2023; Schlicht et al., 2023; Zhao et al., 2015) are based on machine learning classifiers and often approach the problem of subjectivity detection as a binary task. This is largely due to the fact that subjectivity detection is often used as a preparatory step for fact-checking pipelines or sentiment analysis. However, there are a few problems with such an approach. Firstly, theoretical accounts of subjectivity in linguistics suggest that it is a gradual rather than a binary concept, meaning that some utterances can be more subjective than others (Langacker, 1990; Traugott, 1995). Secondly, because this common approach treats subjectivity as a preparatory task for fact-checking or sentiment analysis, the problem of explainabilty of state-of-the-art subjectivity detection models seems to have been outside the focus of attention of scholars working on subjectivity detection tools. In an attempt to tackle the continuous nature of subjectivity, Savinova and Moscoso Del Prado (2023) created a transformer-based subjectivity regressor trained on native English speakers' judgements. The aim of the present contribution is to address the issue of explainability of this subjectivity regressor applying a combination of two approaches: 1) a top-down approach using manual selection of theoretically defined subjectivity features and 2) a bottom-up approach using an automatic local interpretable model-agnostic explanation method (LIME). By collecting evidence from these two approaches, we can gain insights into the features that our transformer-based regressor utilizes for subjectivity analysis, as well as understand how they align with subjectivity theory. Although early rule-based subjectivity detection algorithms (Riloff and Wiebe, 2003; Riloff et al., 2003; Wiebe and Riloff, 2005) relied on some of the theory-based features to distinguish between subjective and objective texts, it is unclear how important these features are for the state-of-the-art machine learning-based models of subjectivity analysis.

Another contribution of the present study lies in the comparison of subjectivity theory with native speakers' perceptions of subjectivity. State-of-the-art subjectivity detection models for English are mostly trained on the gold standard subjectivity

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 305–315

dataset (Pang and Lee, 2004) that was automatically annotated using the source of a text as a proxy for its subjectivity: the dataset contains 5,000 sentences taken from movie review snippets, automatically labeled as subjective, and 5,000 sentences taken from movie plot summaries, automatically labeled as objective. Although this division undoubtedly correlates with the subjectivity distinction, this automatic annotation is not very accurate and does not reflect native speakers' intuitions about subjectivity (Savinova and Moscoso Del Prado, 2023). Similarly, datasets with manual annotations of subjectivity following specific theoretical guidelines (e.g., Antici et al., 2023), and therefore subjectivity detection models trained on such datasets (Pachov et al., 2023; Schlicht et al., 2023), may not coincide with the way subjectivity is perceived by naïve language users. In the present paper, we are looking at the explainability of a subjectivity regressor that was trained on the subjectivity judgements by naïve native English speakers who did not follow any explicit annotation guidelines. Therefore, we can assume that our regressor reflects an average native speaker's understanding of subjectivity. By directly testing the predictive value of theoretically defined subjectivity features for the regressor's subjectivity scores and comparing LIME's explanations with these features, we can understand how subjectivity theory corresponds to native speakers' perceptions of subjectivity.

2 Methodology

2.1 Dataset and model

The dataset and the model that we work with are described in detail in Savinova and Moscoso Del Prado (2023). The dataset contains sentences from news posts on Facebook and news articles on the websites of four major British news outlets (BBC, Sky News, Daily Mail, Metro) on the topics of "crime" and "Covid-19". There are 4,778 sentences (72,236 words) taken from Facebook news posts and 2,973 sentences (65,058 words) taken from news articles on the websites.

For a subset of 398 sentences from this dataset, subjectivity annotations of 19 native English speakers were collected in such a way that every speaker received 100 randomly assigned sentences for annotation and every sentence was annotated by 4 or 5 speakers. The annotators had to rate subjectivity of the sentences on a 7-point scale. There were no explicit annotation guidelines except for brief definitions of *subjective* as meaning "expressing personal opinions, emotions, feelings and tastes, hopes and wishes, self-made conclusions (e.g., *This is awful*)", and *objective* meaning "reporting facts, events, conclusions supported by data (e.g., *The President had a meeting with the Prime minister*)". We ensured that the annotators rated subjectivity by including comprehension checks in the form of clearly objective (*London is the capital of the UK*) and clearly subjective (*This is very beautiful*) sentences that had to receive a score of 1 and 7, respectively, in order for a participant's data to be included. The mean correlation between each rater and the other raters was .64. After transforming the mean subjectivity scores into a [0-1] scale, we split this labeled subset into training, validation and test sets (298/50/50) and trained a RoBERTa-base model (Liu et al., 2019) fine-tuned on our unlabeled sentences to produce subjectivity scores per sentence. Model performance on the test set showed that it correlated highly with the average speaker judgements (.79). The model was then applied to annotate the whole dataset for subjectivity. The annotated dataset is available in open access.

2.2 Approach to Explainability

In order to explain the predictions of our model on the dataset and to elucidate how they relate to subjectivity theory, we employ two methods: a top-down approach and a bottom-up approach. With the top-down approach, we manually selected the most common subjectivity features identified in linguistic theories on subjectivity (in a social media context) and annotated our dataset for the presence of these features. We then built a linear regression with the presence of each feature as a predictor, controlling for sentence length, to check whether theoretically defined subjectivity features indeed correlate with higher subjectivity scores and to estimate the relative importance of each feature in contributing to the subjectivity score. This approach can provide insights on the alignment between the model scores, and therefore average speaker judgements about subjectivity, and theoretically defined subjectivity features.

In contrast to the top-down approach that starts with the theory, the bottom-up approach starts with the data and allows us to to look at the features that are important for the model's scores for each sentence. To perform such a bottom-up inspection, we chose to look at the local explainability of our model on each sentence in our dataset us-

ing LIME (Local Interpretable Model-agnostic Explanations) method (Ribeiro et al., 2016). This method is model-agnostic and provides a good approximation for interpretation of any model's local behaviour. For textual data, LIME treats words as features and creates perturbations of the text entry by excluding different words. A local explainable model is then trained on the dataset consisting of these perturbations and their corresponding scores given by the original black box model. This results in every word/feature receiving a weight score indicating its contribution to the original model's prediction. For our bottom-up approach, we applied LIME to every sentence in our dataset and extracted the words/features and their mean weights and frequency in order to look at the top subjective and objective features. Comparing these top features to theoretically defined features can shed light on the local importance of different subjective features in the explainability of our regressor and in native speakers' local reasoning.

2.3 Subjectivity features

In order to interpret the results of the model using a top-down approach, we selected a number of theoretically defined subjectivity features from the literature. An overview of the features with examples and corroborating literature is provided in the Appendix (Table A1). We annotated each sentence in our dataset for the presence of these features. For every feature, the number of elements corresponding to this feature in every sentence was extracted. As a preparatory step, the sentences were preprocessed (i.e., tokenized, lemmatized, POS-tagged) using the *en_core_web_sm* pipeline for English from the Spacy library (Honnibal et al., 2020). The subjective elements were identified by their lemmas.

Emoji were identified by adding *spacymoji* pipeline to the preprocessing step. **First and second person pronouns** consisted of a list of all possible pronoun forms. **Questions** and **exclamations** were identified by a question and an exclamation mark, respectively. The list of **modal adverbials and adjectives** (e.g., *possible*, *likely*, *indeed*) was taken from Biber and Finegan (1988) and Biber (2004). We selected only those elements that have a modal meaning (factive, non-factive, evidential, certainty, doubt and likelihood adverbials). **Modal verbs** (e.g., *can*, *could*, *should*) were taken from Biber (2004). **Evaluative adjectives and adverbs** (e.g., *adorable*, *terrified*, *incredibly*) were taken from several sources: 1) attitudinal stance adverbials and adjectives from Biber and Finegan (1988) and Biber (2004), 2) adjectives from the Spacy sentiment lexicon, which uses the TextBlob library (De Smedt and Daelemans, 2012), with a subjectivity score above .7, 3) adjectives from MPQA subjectivity lexicon (Wiebe et al. 2005) tagged as "strong subjectivity", which means that they should be subjective in most contexts (Wilson et al., 2005). Since subjectivity lexicons are compiled using corpus data, usually from a specific genre or text type, they may miss out on subjective adjectives when applied in a different context. Therefore, after compiling this list of items, we extracted all adjectives from our dataset that were not part of the list and manually added 72 adjectives we considered subjective, such as, for instance, *worrying*, *hellish*, and *vile*. **Focus particles** (e.g., *only*, *just*, *too*) were taken from König (1991). **Intensifiers** (e.g., *very*, *really*, *totally*) were taken from Zhiber and Korotina (2019). **Epistemic phrases** of the form 'I + cognitive verb' (e.g., *I think*, *I believe*) were taken from Wierzbicka (2006). They were identified by searching for "I" followed by one of the cognitive verbs in present tense with optional negation in between.

3 Explaining the model using manual feature selection

3.1 Procedure

To estimate whether our model's predictions, and therefore native speakers' intuitions, correspond to the theoretically defined features outlined above, we built a linear regression model in R (R Core Team, 2022) predicting the subjectivity score with subjectivity features as categorical factors (presence/absence of the feature). All factors were coded using treatment contrasts so that the effect of every feature is estimated when the other features are absent. To control for sentence length, we also included log-transformed word count as a predictor in the regression. Logarithmic transformation allowed us to account for the non-linear relationship between the word count and the subjectivity score. The results of the regression model can be found in Table 1 with estimates ranked from largest to smallest.

3.2 Results

The results show that all predictors were significant, suggesting that the theoretically defined subjectiv-

Predictor	Estimate	Estimated means[1]	Std. error	*t*	*p*
Emoji	0.29	0.40-0.69	0.01	23.28	<.001
First and second person pronouns	0.18	0.38-0.57	0.01	31.36	<.001
Exclamations	0.16	0.40-0.57	0.02	8.20	<.001
Questions	0.14	0.40-0.54	0.01	12.29	<.001
Intensifiers	0.14	0.40-0.54	0.01	12.03	<.001
Evaluative adjectives and adverbs	0.12	0.38-0.50	0.005	25.83	<.001
Epistemic phrases	0.11	0.40-0.51	0.02	5.28	<.001
Modal verbs	0.10	0.39-0.49	0.01	18.98	<.001
Modal adverbials and adjectives	0.05	0.40-0.45	0.01	6.43	<.001
Focus particles	0.05	0.40-0.45	0.01	8.37	<.001
Word count	-0.09	NA	0.003	-30.32	<.001

[1] Estimated means of the model when the predictor is absent versus present, obtained using the *effects* package (Fox and Weisberg, 2018).

Table 1: Model output of features as subjectivity predictors, ordered by the estimates.

ity features correspond to the speakers' intuitions that our model was trained on. Comparison of the estimates suggests that the presence of emoji leads to the most substantial change in subjectivity score. Together with first and second person pronouns, exclamations, questions, intensifiers and epistemic phrases, these features bring the score over .5, assuming this threshold roughly indicates the transition from objective to subjective. The other features, in particular modal adverbials and focus particles, contribute to a minimal shift in subjectivity scores. Log-transformed word count turned out to be a significant predictor of subjectivity scores as well: Figure 1 shows that higher word count leads to lower subjectivity scores, which is most noticeable in the 1-15 word count range. This is understandable given social media data, since many one- and two-word posts on social media contain an evaluative adjective and/or an emoji (e.g., *Awful* 😔). It should be noted that the linear regression model explains 40% of the variance (adjusted R^2=.40), suggesting that the selected subjectivity features cannot fully explain the subjectivity regressor. This is not surprising: a sentence can be subjective even without explicit subjective elements (e.g., *The lights are on, so he is home*), which is why using rule-based subjectivity feature extraction will always result in an underestimation of the scores compared to the machine learning-based subjectivity detection. However, it could also be the case that the theory on subjectivity misses out on features that are deemed important to naïve speakers. Any such features could be identified by a bottom-up, theory-agnostic explainability approach.

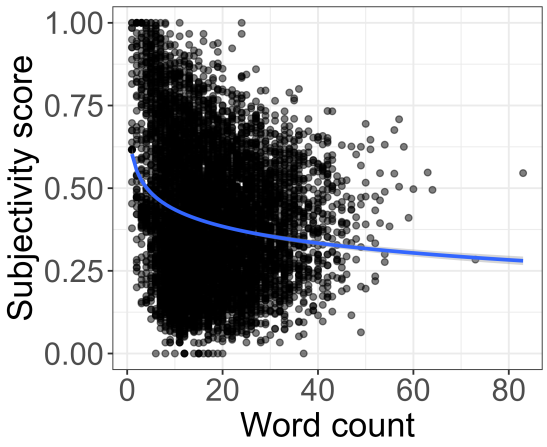

Figure 1: Effect of word count on subjectivity score.

4 Explaining the model using LIME

4.1 Procedure

As stated above, LIME can offer model-agnostic local approximations of model explanations based on textual features, i.e. words (Ribeiro et al., 2016). Although LIME has been used to explain sentiment analysis models (e.g., Chowdhury et al., 2021; Jain et al., 2023), its applicability to subjectivity analysis models appears to have been largely overlooked. It is important to note that LIME cannot provide an explanation of the internal working of the black box model, which goes beyond human-understandable features like words. However, since the black box, especially in the case of transformer models, cannot be understood as such, LIME provides a useful tool for interpreting what the black box model does locally on a level comprehensible to humans.

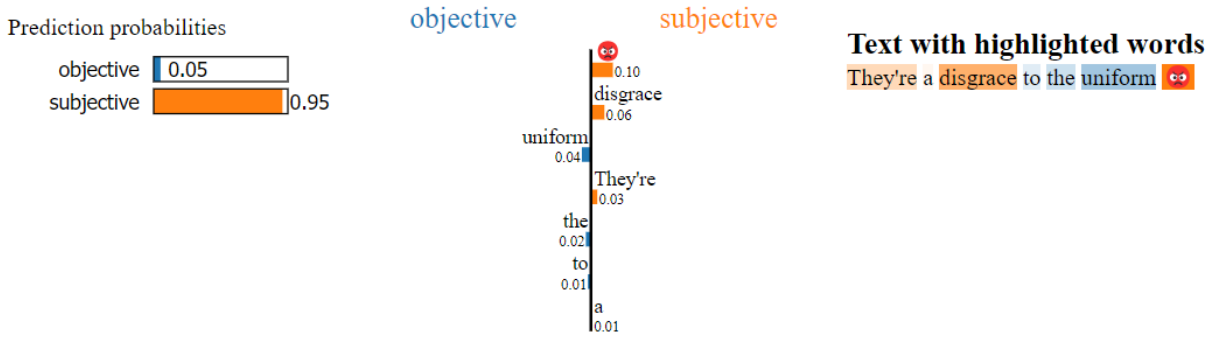

Figure 2: Example of LIME output for a sentence.

To explain our subjectivity regressor locally, we applied LIME to every sentence in our dataset and collected the weights associated with every feature in every sentence. We then computed the frequency of every feature in our dataset (case-insensitive), its mean weight in the dataset and the minimum and maximum weights. We modified the original LIME code, which uses a simple regular expression to tokenize the input and identify only words as potential textual features. Instead, we employed the TweetTokenizer from NLTK package (Bird et al., 2009), enabling the recognition of punctuation marks and emoji as potential features as well. Figure 2 provides an example of LIME output for one of the sentences from our dataset with weights per feature. As the output indicates, features that contribute to objectivity are associated with negative scores, while features that contribute to subjectivity are associated with positive scores.

4.2 Results

The complete list of 12,535 unique features resulting from application of LIME to our dataset is available in open access. To interpret the LIME output for our dataset, we decided to look at the top 200 subjective and objective features with the highest and lowest weights, respectively. For illustration, Table 2 shows the top 10 most subjective and objective features identified by LIME. Among the top objective features, the most prominent categories were numbers (*6400*), dates and times (*13:00*), proper nouns (*Churchill*) and concrete nouns (*candles*, *airbag*), links (*https://trib.al/7nvqdio*), verbs (*matches*, *redeployed*) and non-evaluative adjectives (*month-long*, *water-related*). In contrast, the top subjective features were dominated by emoji, evaluative adjectives (*unforgivable*, *reckless*, *stunning*) and evaluative nouns (*hypocrite*, *downfall*, *gamechanger*). It is notable that all types of emoji, not only the ones representing faces/emotions, were found to be very subjective according to LIME results.

Inspection of both subjective and objective features revealed a strong frequency bias: almost all of the top 200 features were encountered in our dataset only once. It is not surprising that these infrequent words received the most extreme scores, since their weights were based only on one example sentence. To compare the findings with the results of the top-down approach that includes theory-based subjectivity features, we eliminated the frequency bias by excluding features with a frequency of less than five. The remaining data consisted of 3,285 features, which is approximately a quarter of the original list. Table 3 shows top 10 subjective and objective features in this subset of the data. After selecting only those LIME features that are encountered at least 5 times in the dataset, we annotated them for the presence of the theoretical subjectivity features described in Section 2.3. Epistemic phrases were not used for annotation because they require multiword expressions. Subsequently, we checked which theoretical features appeared in the top 200 subjective and objective LIME features.

In line with our expectations, there were no theoretically defined subjectivity features among the top 200 objective LIME features. On the contrary, all theoretical subjectivity features were present among the top 200 subjective LIME features. The latter included the question and exclamation marks, 58 evaluative adjectives, 11 emoji, 10 modal verbs, 6 modal adverbials, 6 intensifiers, one personal pronoun (*I*) and one focus particle (*too*). The theoretically defined features thus accounted for around 48% of the top 200 subjective LIME features. The other subjective features identified by LIME mostly included emotionally laden nouns (*horror*, *shock*, *hope*) and verbs (*missed*, *enjoy*, *worry*). While

Feature	Mean weight	Frequency
-202012/01	-0.32	1
four-year-old	-0.28	1
accounted	-0.28	1
re-arrested	-0.27	1
Churchill	-0.27	1
murder-suicide	-0.27	1
https://trib.al/7nvqdio	-0.26	1
mugshots	-0.25	1
plea	-0.24	1
120ft	-0.24	1

Feature	Mean weight	Frequency
🔽	0.30	2
🐠	0.30	1
🎿	0.30	1
😤	0.29	1
😦	0.28	1
heartlessly	0.27	1
😭	0.26	1
🚨	0.24	2
😰	0.24	1
🙏	0.23	1

Table 2: Top 10 objective (left) and subjective (right) LIME features in the dataset.

Feature	Mean weight	Frequency
detect	-0.16	5
jailed	-0.16	104
arrests	-0.15	12
fined	-0.14	8
homicide	-0.14	5
two-year-old	-0.14	5
arrested	-0.13	121
25,000	-0.13	5
eight-year-old	-0.13	5
anti-vaxxer	-0.12	11

Feature	Mean weight	Frequency
💔	0.19	20
🙌	0.19	10
💉	0.18	5
🙏	0.16	7
😱	0.13	5
terrifying	0.11	6
awful	0.11	8
shocking	0.11	9
👏	0.09	12
wonderful	0.09	8

Table 3: Top 10 objective (left) and subjective (right) LIME features with frequency above 5 in the dataset.

these features were not included in our top-down analysis, they are in line with the general definition of subjectivity provided by the theory: a speaker conveying their judgement, opinion, or emotion. Interestingly, the top 200 subjective LIME features also included multiple negative words, such as *wasn't*, *didn't*, *none*, *no*. Some theoretical literature (e.g., Dancygier, 2012) suggests that negation can be considered a subjective viewpoint device because it evokes an alternative set-up and expresses the speaker's negative stance towards this set-up (for instance, *This is not funny* could be interpreted as expressing the speaker's negative attitude to the alternative *This is funny*). More work is needed to investigate the role of negative words in signalling subjectivity.

It is noteworthy that while for the LIME results, evaluative adjectives were clearly dominating the top subjective features list, in the manual approach, they seem to be of a lesser importance and do not bring the score over .50. We believe that this could have several reasons. Firstly, our list of evaluative adjectives in the top-down approach consisted of a large number of adjectives (1611), not all of which are *always* subjective. In contrast, the evaluative adjectives appearing in the top subjective LIME features seem to be those that are very subjective independent of the context. In other words, it appears that our top-down approach may lack accuracy with respect to evaluative adjectives. Secondly, the relatively low importance of evaluative adjectives in the top-down approach could be related to the fact that they often co-occur with stronger subjectivity indicators, such as, for instance, emoji or exclamation marks.

Interestingly, the top 200 subjective LIME features also contained three versions of quotation marks (", ', '), suggesting that the quotations of third party sources in news texts were treated as subjective. This goes against the strict approach to subjectivity where subjectivity of a third person cited source should not count, since it is being merely reported by the author (Sanders, 1994). Moreover, quoting sources, however subjective their comments are, has a place within the tradition of objective news reporting. In other words, while *This is terrible* is subjective, *He said: "This is terrible"* should be (more) objective. In an attempt to

take a closer look at whether our model, and consequently native speakers, distinguish between author and source subjectivity, we conducted additional analyses that are reported in the next section.

5 Author vs. source subjectivity

In texts that cite other sources, such as news, two types of subjectivity can be distinguished: author subjectivity and the subjectivity of the reported sources (Banfield, 1982; Pit, 2003; Sanders, 1994). This distinction is unclear in the case of indirect reported speech, when it is unknown who exactly is responsible for the wording of the cited fragment (e.g., *Shepherd's mother said that medical staff treated her daughter well and did everything in their power to save her*). In contrast, in the case of direct speech it is always clear that the quoted part corresponds to the voice of the source; the sentence could therefore be considered objective since it is merely (objectively) reporting the subjectivity of the source. When collecting annotations for the subjectivity regressor, we did not explicitly instruct the raters about this distinction, since we do not know whether naïve speakers share the intuition that fragments with citations are objective.

In order to test whether this distinction is indeed important to naïve language users, we identified author and source fragments in every sentence in our dataset using quotation marks. We then counted the presence of subjectivity features in author and source text separately using the top-down approach with theoretically defined features. The dataset was then extended in such a way that the sentences containing both author and source subjectivity were split into two separate entries, and Origin of subjectivity (author vs. source) was added as a separate variable. We built another linear regression model specifying interactions of Origin with all subjectivity features except for emoji, which was entered as a main effect, since there was only one case of emoji used in the source text.

The results revealed four significant interactions with Origin: personal pronouns (t=-4.48, p<.001), questions (t=-2.40, p=.02), intensifiers (t=-3.11, p=.002) and epistemic phrases (t=-2.82, p=.005). The interaction plots (Figure 3) show that the effect of encountering these features in the author text leads to a bigger change in subjectivity score as compared to the source text. On the one hand, such an outcome would be predicted if speakers distinguish between author and source subjectivity: source subjectivity is reported and therefore should influence speakers' judgements less than author subjectivity. On the other hand, there were no significant interactions for the other subjectivity features. In addition, there was a main effect of Origin (t=22.51, p<.001), indicating that the source text was overall more subjective than the author text, which supports our findings from LIME where quotation marks appeared in the top subjectivity features. This suggests that subjective quotations of third person sources are still considered subjective by the model and native speakers. However, by not commenting on the author versus source distinction when collecting annotations, we may have implicitly prompted the participants to rate all kinds of subjectivity, regardless of whether it stemmed from the author or the source. Whether author versus source subjectivity is indeed a relevant distinction for naïve language users and whether the relevance differs between subjectivity features seems like a fruitful direction for future research.

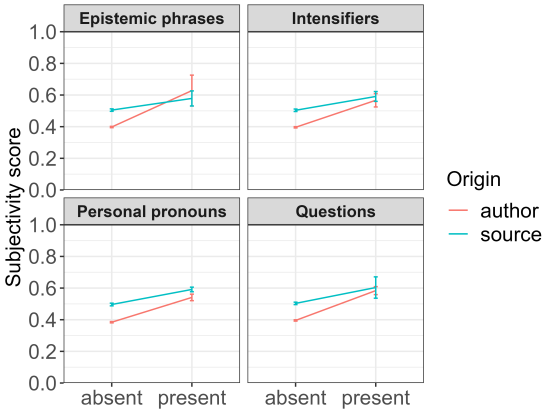

Figure 3: Interaction plots for model estimated marginal means, with confidence intervals.

It should be noted that in this more complex model, modal adverbials (t=.60, p=.55) and focus particles (t=1.39, p=.16) showed no significant effect on subjectivity scores. Upon closer inspection of our data, we found that certain evidential modal adverbials, such as *allegedly*, *reportedly*, *apparently*, were on average associated with rather low subjectivity scores (.25, .31, .33, respectively). These adverbials seem to be used in news discourse, and especially in crime news, to indicate the common agreement/existence of evidence about what is being introduced and, as such, are rather employed to underline objectivity and impartiality of the author. Among the focus particles, *merely* and

at least were associated with rather low subjectivity scores (.24 and .26). A closer inspection revealed that these were used in the context of news in their rather factual non-focus meanings (e.g., *people who merely tested negative*; *at least 20 killed*). These qualitative observations underline that whether a specific feature is a subjectivity indicator can be dependent on the context.

6 Density of subjectivity features

A single utterance may contain multiple subjective elements (e.g., *Delivering smiles during a tough time!* contains both an evaluative adjective and an exclamation mark). The output of LIME and the linguistic theory on subjectivity suggest that, at least in some cases, more subjective features in the sentence should lead to increased subjectivity (for example, *This is really really bad* seems more subjective than *This is bad*). At the same time, it is also intuitively clear that adding an exclamation mark to the sentence that already has an emoji at the end will probably not make it much more subjective than it already is. This division of labour between subjectivity features in different contexts is clearly visible in the different weights that the LIME features get depending on the sentence that is being analyzed. To test the relationship between the number of subjective elements in a sentence and its subjectivity score, we built a generalized additive model (GAM) with smooth terms for the count of all subjective elements in the sentence (including multiple instances of one feature) and for word count as a control. The results of the model, which explained 39.5% of deviance, showed significance of both the smooth term for the word count (p<.001) and the smooth term for the number of subjective elements in a sentence (p<.001). The effect of the latter is visualized in Figure 4. The visualization shows a logarithmic curve, which illustrates that increasing the number of subjective features from 0 to 4 has a strong positive effect on the subjectivity score, while for any subsequent increase the effect levels off.

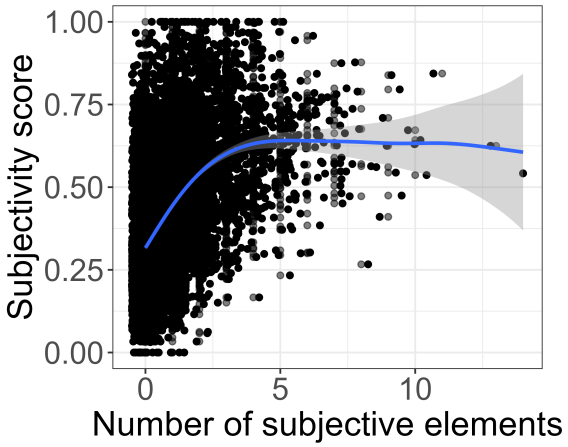

Figure 4: Results of GAM for the density of subjective elements.

7 Conclusion

In this paper, we approached the problem of explainability of a transformer-based subjectivity regressor trained on native English speakers' judgements using two methods: a top-down manual selection of theoretically defined subjectivity features and a bottom-up extraction of top subjective and objective features using LIME explanation method. The explainability was tested on a news dataset containing sentences from social media news posts and articles on the websites of the corresponding news outlets. The results of the two methods provided a similar picture: most of the theory-based subjectivity features turned out to be important for our regressor's predictions. According to both methods, emoji, exclamations, questions, intensifiers and first person pronouns turned out to be prominent predictors of subjectivity scores. The results of the bottom-up approach also revealed the significance of evaluative adjectives, especially the ones that are highly subjective across contexts, as a top subjective feature. We also found that the more subjective elements are present in a sentence, the more subjective it becomes. Overall, our findings suggest that the features used by the subjectivity regressor in its judgements align with the subjectivity theory.

Since the regressor was trained on native English speakers' intuitions and, therefore, represents an average speaker's perception of subjectivity, our findings mentioned above seem to indicate that the naïve speakers' perceptions correspond to the theoretical accounts of subjectivity in linguistics. At the same time, our regressor does not seem to distinguish between author and source subjectivity, contrary to what theory predicts. Future work could investigate what role this distinction plays in naïve speakers' perceptions of subjectivity.

Limitations

The list of subjectivity features that we used in this work was not exhaustive. For instance, we did not include affective nouns and verbs (e.g, *enjoy*, *horror*, *love*) as subjective features in our top-down approach, but they did show up in the LIME output, which suggests that they belong to some of the most influential features for the subjectivity regressor and for native speakers' judgements. We also employed a rather coarse measure in our analysis of the density of subjectivity features. The effect of number of subjective elements might vary depending on the particular type of subjectivity feature or the specific combination of such features. In addition, subjectivity cannot be reduced to explicit subjective elements. In that sense, both the manual selection of features and the LIME method are limited in their explanation capacity since they can only take into account explicit subjective markers.

Our limitation in the approach to author vs. source subjectivity lies in the fact that we provided minimal instructions for annotators for the sake of obtaining their natural intuitions about subjectivity. This could have prompted participants to rate any kind of subjectivity. In addition, splitting the texts by sentence resulted in some quotations being fragmented and unrecognizable as cited text without context. Finally, our approach to annotation of Origin was not conceptually ideal, as it resulted in splitting some sentences into author and source parts and assigning the same score to them. Future work is needed to address the issue of author vs. source subjectivity in speakers' intuitions more comprehensively, potentially within context.

Acknowledgements

We would like to thank Jetske Adams and Michael Voronov for their help with the LIME code. We would also like to thank Wilbert Spooren for his assistance in selecting theory-based subjectivity features and his help with editing the section on model explanation using manual feature selection.

References

Francesco Antici, Andrea Galassi, Federico Ruggeri, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2023. A corpus for sentence-level subjectivity detection on English news articles. *arXiv:2305.18034*.

Angeliki Athanasiadou. 2006. *Subjectification: Various paths to subjectivity*. Mouton de Gruyter.

Ann Banfield. 1982. *Unspeakable sentences: Narration and representation in the language of fiction*. Routledge.

Douglas Biber. 2004. Historical patterns for the grammatical marking of stance: A cross-register comparison. *Journal of historical pragmatics*, 5(1):107–136.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in English. *Discourse processes*, 11(1):1–34.

Douglas Biber and Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language*, 65(3):487–517.

Douglas Biber and Edward Finegan. 2001. Diachronic relations among speech-based and written registers in English. In S. Conrad and D. Biber, editors, *Variation in English*, pages 66–83. Routledge.

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.

Kounteyo Roy Chowdhury, Arpan Sil, and Sharvari Rahul Shukla. 2021. Explaining a black-box sentiment analysis model with local interpretable model diagnostics explanation (LIME). In *Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021*, pages 90–101. Springer.

Barbara Dancygier. 2012. Negation, stance verbs, and intersubjectivity. In B. Dancygier and E. Sweetser, editors, *Viewpoint in language: A multimodal perspective*, pages 69–93. Cambridge University Press.

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

John Fox and Sanford Weisberg. 2018. *An R companion to applied regression*. Sage.

Mario Haim, Michael Karlsson, Raul Ferrer-Conill, Aske Kammer, Dag Elgesem, and Helle Sjøvaag. 2021. You should read this study! It investigates scandinavian social media logics. *Digital Journalism*, 9(4):406–426.

Matthew Honnibal, Sofie Montani, Ines Van Landeghem, and Boyd Adriane. 2020. Spacy: Industrial-strength natural language processing in python.

Marianne Hundt and Christian Mair. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, 4(2):221–242.

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.

Rachna Jain, Ashish Kumar, Anand Nayyar, Kritika Dewan, Rishika Garg, Shatakshi Raman, and Sahil Ganguly. 2023. Explaining sentiment analysis results on social media texts through visualization. *Multimedia Tools and Applications*, 82(15):22613–22629.

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.

Christopher Kennedy. 2013. Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, 56(2-3):258–277.

Ekkehard König. 1991. *The meaning of focus particles: A comparative perspective*. Routledge.

Ronald W Langacker. 1990. *Subjectification*. Walter de Gruyter.

Geoffrey N. Leech. 2009. *Change in contemporary English: A grammatical study*. Cambridge University 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*.

Petra K. Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Georgi Pachov, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2023. Gpachov at checkthat! 2023: A diverse multi-approach ensemble for subjectivity detection in news articles. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, pages 404–412.

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.

Michael Pearce. 2005. Informalization in UK party election broadcasts 1966-97. *Language and Literature*, 14(1):65–90.

Mirna Pit. 2003. *How to express yourself with a causal connective: Subjectivity and causal connectives in Dutch, German and French*. Rodopi.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

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.

José Sanders. 1994. *Perspective in narrative discourse*. Doctoral dissertation. Tilburg University, Tilburg.

Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314, Toronto, Canada. Association for Computational Linguistics.

Ipek Baris Schlicht, Lynn Khellaf, and Defne Altiok. 2023. Dwreco at checkthat! 2023: Enhancing subjectivity detection through style-based data sampling. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, pages 306–317.

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.

Kirsten Vis, José Sanders, and Wilbert Spooren. 2012. Diachronic changes in subjectivity and stance – a corpus linguistic study of Dutch news texts. *Discourse, Context & Media*, 1(2-3):95–102.

Kasper Welbers and Michaël Opgenhaffen. 2019. Presenting news on social media: Media logic in the communication style of newspapers on facebook. *Digital journalism*, 7(1):45–62.

Ingrid Westin and Christer Geisler. 2002. A multidimensional study of diachronic variation in British newspaper editorials. *ICAME Journal*, 26:133–152.

Janyce Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

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.

Anna Wierzbicka. 2006. *English: Meaning and culture*. Oxford University Press.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

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.

Evgeniya Zhiber and Larisa Korotina. 2019. Intensifying adverbs in the English language. *Training, Language and Culture*, 3(3):70–88.

A Appendix

Subjectivity feature	**Supporting literature**[1]	**Examples**
Emoji	N, W&O	*Horrific news* 😱
First and second person pronouns	B&F, H&M, P, V, W&G	*I, me, my, you, our, yourself*
Questions	B&F, V, W, W&G	*Will I be protected if I have a booster?*
Exclamations	H, L, V, W	*Watch live!*
Modal adverbials and adjectives	B, B&F, V, W	*sure, possibly, in fact, apparently*
Modal verbs	B, B&F, V, W, W&G	*can, could, should, seem to*
Evaluative adjectives and adverbs	A, Ke, B&F, W2	*honestly, amazing, horrible, immense*
Focus particles	K, V	*only, just, already, exactly*
Intensifiers	B&F, V, W, W&G, Z&K	*very, really, extremely, so*
Epistemic phrases	B, Wie	*I think, I believe, I would say, I guess*

[1] A = Athanasiadou (2006), B = Biber (2004), B&F = Biber and Finegan (1988, 1989, 2001), H = Haim et al. (2021), Hundt and Mair (1999), K = König (1991), Ke = Kennedy (2013), L = Leech (2009), N = Novak et al. (2015), P = Pearce (2005), V = Vis et al. (2012), W = Wiebe (1994), W2 = Wiebe et al. (2005); Wie = Wierzbicka (2006), W&G = Westin and Geisler (2002), W&O = Welbers and Opgenhaffen (2019), Z&K = Zhiber and Korotina (2019)

Table A1: List of subjectivity features used in the top-down approach, with the corresponding theoretical literature and examples.