The Learnability Hierarchy of News Values: What Makes Some Journalistic Concepts Harder to Classify?

Elisabeth Muth Andersen[1]

[1] Department of Culture and Language, University of Southern Denmark, Denmark

Abstract

This study analyzes the performance patterns of BERT-based classifiers trained to identify news values in Danish journalism, revealing a systematic learnability hierarchy among journalistic concepts. We trained multilabel classifiers on 59,108 LLM-annotated sentences across 10 news values and 62 subcategories, using perturbation-based analysis to examine linguistic decision-making patterns. Results demonstrate three distinct performance tiers. High-performing classifiers like 'unexpectedness' rely on consistent surface markers ("omvendt," "anderledes"), achieving reliable automated detection. Mid-tier classifiers such as 'personalization' and 'timeliness' show context-dependent but learnable patterns. Low-performing classifiers like 'eliteness' struggle due to complex pragmatic reasoning requirements and severe class imbalance. Word importance analysis reveals that successful classifiers attribute significance to single words, while struggling classifiers distribute importance across many words per sentence. These findings suggest that computational journalism requires methodological pluralism""combining transformer models with sentiment analysis and named entity recognition based on concept complexity rather than applying uniform approaches to all news values.

Keywords: news values, text classification, BERT, discourse analysis, computational journalism

1 Introduction

When attempting to automate the classification of news values""journalistic concepts like 'eliteness' and 'unexpectedness'""some prove remarkably easy to learn while others resist classification entirely. This performance hierarchy reveals fundamental differences in how these concepts are realized linguistically in news discourse.

News values, the criteria journalists use to assess newsworthiness [12], have traditionally been studied through manual analysis of small corpora or rule-based approaches that identify surface-level markers [6; 25]. However, recent advances in transformer models like BERT [10] offer new possibilities for understanding how these abstract journalistic concepts are actually realized in language at scale. When we train multilabel classifiers to identify ten different news values in Danish news discourse, a striking pattern emerges: some news values consistently achieve high performance while others systematically struggle, regardless of data volume or model architecture.

This performance variation is not merely a technical limitation""it reveals something fundamental about the linguistic nature of news values themselves. Concepts like 'unexpectedness' rely on consistent surface markers ("omvendt," (conversely) "anderledes," (different) "modsatte"

Elisabeth Muth Andersen. "The Learnability Hierarchy of News Values: What Makes Some Journalistic Concepts Harder to Classify?." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 336–350. https://doi.org/10.63744/svxDtDD45mvw.

(opposite)) that enable reliable automated detection. In contrast, 'eliteness' requires complex pragmatic reasoning about status attribution that extends far beyond naming specific individuals or institutions. Between these extremes lie news values that are context-dependent but learnable, such as 'personalization,' which depends on intensity rather than mere presence of personal markers.

This paper analyzes the performance patterns of BERT-based classifiers trained on 59,108 LLM-annotated Danish news sentences across 10 news values and 62 subcategories. Through perturbation-based word importance analysis, we examine what linguistic information different classifiers prioritize when making predictions. The findings reveal a three-tier learnability hierarchy: surface-realizable news values that rely on consistent linguistic markers, context-dependent values that require deeper semantic understanding, and inference-heavy values that demand pragmatic reasoning beyond sentence-level information.

These findings have significant implications for both computational journalism and discourse analysis. They suggest that not all journalistic concepts are equally amenable to automation, and that successful classification systems may require methodological pluralism""combining traditional NLP approaches with sentiment analysis, named entity recognition, and topic modeling depending on the linguistic complexity of the target concept. More broadly, this work demonstrates how computational methods can reveal patterns in discourse categories that traditional corpus linguistic approaches cannot detect at scale, offering new pathways for understanding how abstract journalistic concepts function in language.

2 Background

2.1 Discursive News Value Analysis

Within journalism, news values are commonly treated as a value system existing in the minds of journalists used for making judgements about the perceived newsworthiness of an event [6], originating in the work of Galtung and Ruge [12]. As an alternative, discursive news values analysis (DNVA) analyzes how events are presented as newsworthy in discourse.

Reserving news values for values in news actors and events, nine news values have been suggested: consonance, eliteness, impact, negativity, personalization, proximity, superlativeness, timeliness, and unexpectedness [6]. There is some discussion as to whether to include positivity as a news value as well [6; 30].

Using corpus analytic techniques and manual computer-aided annotation, pointers to each specific news value have been identified [6; 19; 25], in all cases stressing the need for combining qualitative and quantitative methods because of the role contextual factors play for conveying news values. With few exceptions, Chen & Liu [8], Huan [17; 18] and Guo, Mast & Vosters [16] on Chinese, Fruttaldo, & Venuti [11] on Italian, and Makki [22; 23] on Iranian, DNVA has predominantly been applied to English material.

Attempts to use corpus linguistic tools have proven to be difficult as, according to Javadinejad [19], common semantic taggers cannot be used to identify news values as their use and interpretation are culturally and context dependent. Common corpus linguistic techniques used include frequency analysis, collocation analysis and concordance analysis combined with statistical analyses and manual analyses [6; 19; 25].

2.2 Text classification of news discourse

Notably, while corpus linguistics include computer-assisted methods, machine learning approaches such as classification have not been applied in DNVA analyses. Text classification has, however, been applied to related fields. Some studies use more traditional machine learning techniques such as Support Vector Machine, logistic regression, KNN, and Naïve Bayes etc. to predict articles' popularity on Twitter [5], identify topics [2], or do framing analysis [7].

More recently, pretrained foundation models such as the BERT model have been applied to news discourse with several studies finding that BERT models perform exceptionally well for classifying news [14] and detecting fake news [3; 20; 21].

Rybinski [27] suggests using emotion detection by applying a BERT-based NLP model to predict longevity of news, finding that positive news tend to stay on the main page of a news site for longer than negative news, an analysis based scraped news from Russia, Poland, Ukraine, and Kazakhstan. Adelakun and Baale [1] finds that implementation of the BERT model in sentiment analysis of financial news outperforms existing sentiment algorithms performing with an accuracy score above 95 % on three-class news data. Adding to this knowledge, Ndama and Bensassi [24] compares sentiment analysis accuracy on financial news when comparing different machine learning models, finding that model performance is improved when implementing BERT into neural network architectures such as GRU (gated recurrent units), giving an accuracy score of 97 % for news data with three sentiment classes.

3 Method and data

With a goal of developing multilabel classifiers for each of the 10 news values from the DNVA literature, conceptualizations of the news values were developed iteratively using manual annotation of 100 Danish news articles collected in September 2024, a literature review of previously identified linguistic markers [6; 19; 25; 30] as well as data explorations of preliminary findings which led to the final identification of 5-10 subclasses for each news value (62 total). Each news value subcategory was defined using a semantic conceptualizing title, and initially rule-based patterns were identified for each subclass. BERT multilabel classifiers were implemented (using Maltehb/danish-bert-botxo) for each news value by training them on a rule-based annotated data set consisting of 159,648 sentences from Danish news [13], leading to models with very high precision (above 98%), but with a limited usability as the models learned basic linguistic patterns rather than achieve a context-sensitive understanding of how news values are constructed in news discourse.

To address this, the sampling methods, annotation technique and classification method were changed. Instead of using rule-based annotation, LLM annotation was used. Annotating text data is costly because it is ideally done manually, referred to as the ”˜gold standard”™ [28]. LLM annotation is advancing [29], making the production of annotated data for e.g., NLP classification manageable, but compromising on having human experts manually annotate every text element. Claude.ai was prompted to be a linguistic expert and asked to rate batches of sentences from Danish news discourse from a scale from 1 (absent) to 5 (dominant) in terms of how well a specific news value subcategory matched a sentence. Claude.ai was used for annotation with quality controls and reasonable inter-rater agreement with a human annotator[1].

Various sampling methods were explored iteratively, leading to an increase in data set size and quality. First, the initial rule-based classifiers were used to sample balanced data (low/medium/high confidence), but the combination of the classifier types and the sampling strategy led to many low ratings. To increase the number of high ratings, an active learning approach was used that targets high-confidence examples and sentences with subcategory correlations. Initially, binary multilabel classifiers were refined based on the data annotated on an ordinal scale, but as the size and quality of the data progressed, soft label classification was implemented[2]. The classifiers were trained

[1] Following Alizadeh et al. [4], a simple instruction approach was used. The LLM was provided with subcategory definitions and example sentences. As a quality check, the LLM identified the sentence it was most uncertain about in each batch. Human-LLM inter-rater agreement on 700 sentences: r=0.69, mean difference=0.14.

[2] The model's confidence in predicting news values is defined as the sigmoid-activated output of the intensity regression head: $\text{confidence}(s, c) = \sigma(W_i \cdot \text{BERT}(s)[\text{CLS}] + b_i)$, where s is the input sentence and c is the news value subcategory. During training, human ratings (1–5 scale) were mapped to intensity targets: $\{1 \rightarrow 0.0, 2 \rightarrow 0.3, 3 \rightarrow 0.6, 4 \rightarrow$

using standard hyperparameters[3].

An overview of the final number of subclasses and the number of sentences rated per subclass can be seen in Table 1. The same sentence has been rated for its fit to all subclasses for a news value, with a total of 59.108 being rated.

News value	Number of subclasses	Data annotated	Total sentences rated
Consonance	5	765	3825
Eliteness	6	1154	6924
Impact	6	686	4116
Negativity	7	1226	8582
Personalization	4	1670	6680
Positivity	5	1006	5030
Proximity	8	871	6968
Superlativeness	6	667	4002
Timeliness	10	1000	10000
Unexpectedness	5	596	2980

Table 1: Final samples of annotated data per subclass.

The analysis focuses on dissecting performance patterns rather than absolute scores. The news values represent quite different semantic concepts that are expressed linguistically using different means. Furthermore, the news values and their subclasses are not equally distributed in news discourse, making it difficult to sample data in a balanced way. Therefore, the analysis outlines and interprets the meaning and significance of subclasses with low support, choice of amount and conceptualizations of subclasses, consistency, and class imbalance.

4 Analysis

4.1 Trends: A performance hierarchy

Figures 1-5 visualize precision-recall plots for the performance of the subclasses of each of the 10 news values.

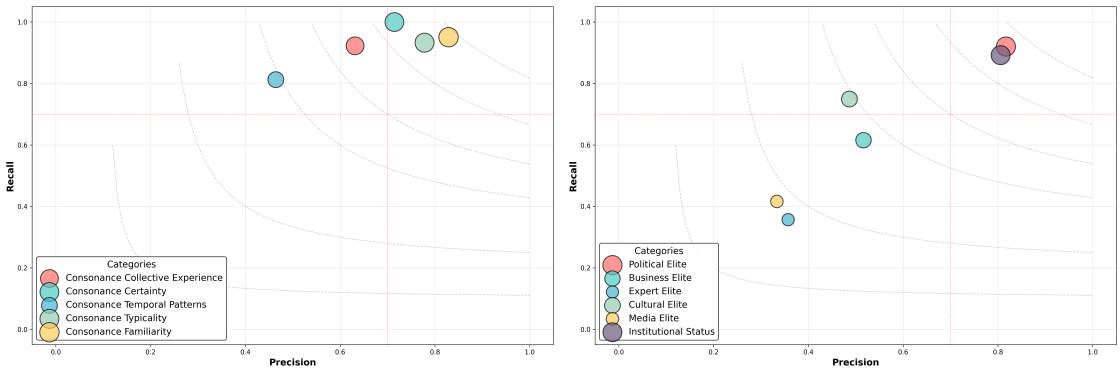

Figure 1: Precision-recall plots for consonance and eliteness subclasses.

$0.9, 5 \rightarrow 0.9\}$.

[3] Training specifications: 25 epochs, batch size 16, learning rate 1e-5. During training, recall was deliberately prioritized over precision due to limited annotated data.

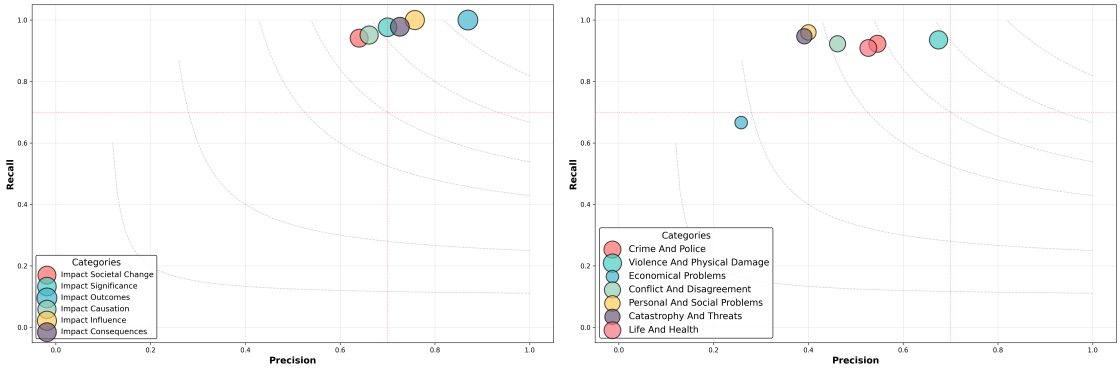

Figure 2: Precision-recall plots for impact and negativity subclasses.

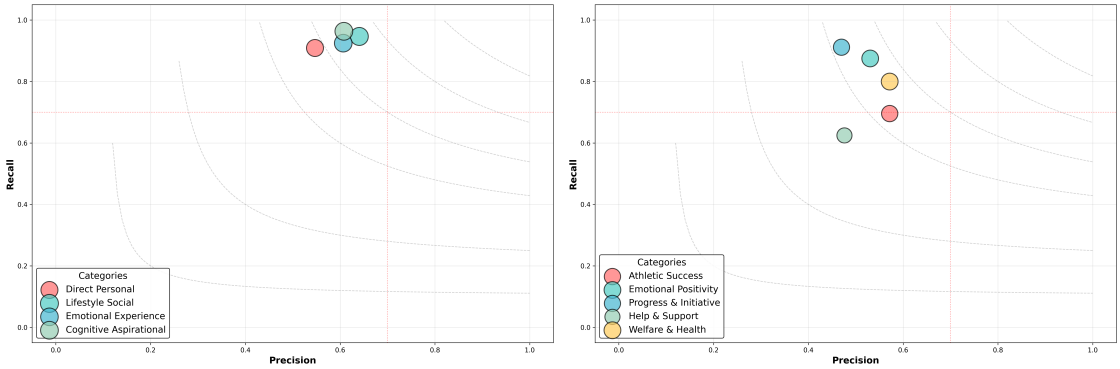

Figure 3: Precision-recall plots for personalization and positivity subclasses.

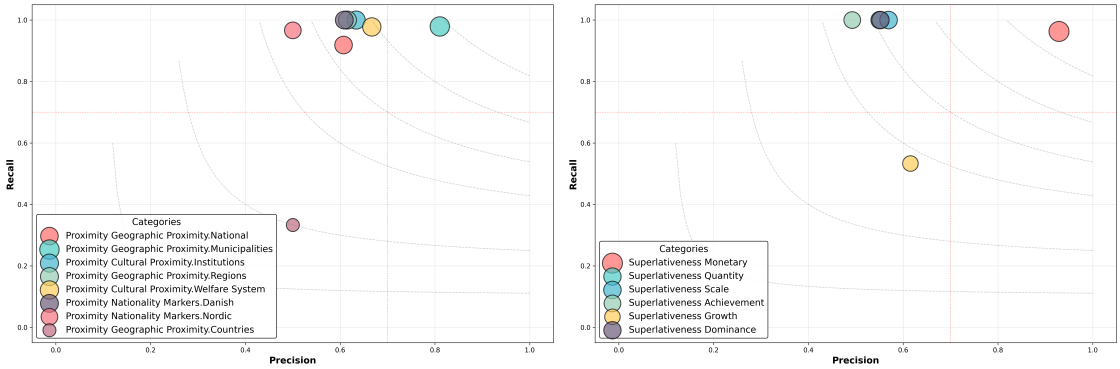

Figure 4: Precision-recall plots for proximity and superlativeness subclasses.

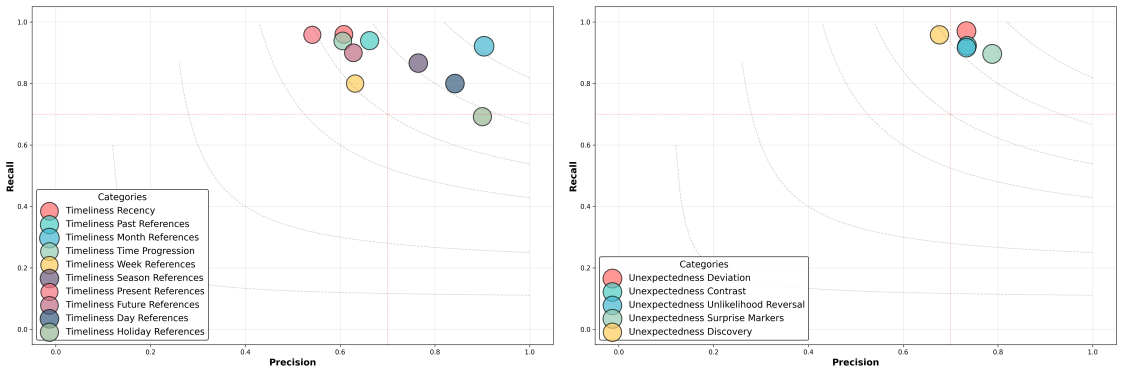

Figure 5: Precision-recall plots for timeliness and unexpectedness subclasses.

Various trends are easily identified. For some classifiers, all subclasses perform reasonably well (unexpectedness, impact), some perform slightly worse and with some variability between the subclasses (consonance, personalization, timeliness), and some classifiers have one or more subclasses that perform badly (superlativeness, proximity, positivity, negativity, eliteness). Figure 6 below visualizes these trends by presenting a hierarchy of classification difficulty measured by the sum of how far the model is from perfect F1 and how inconsistent the model is (coefficient of variation (CV) score).

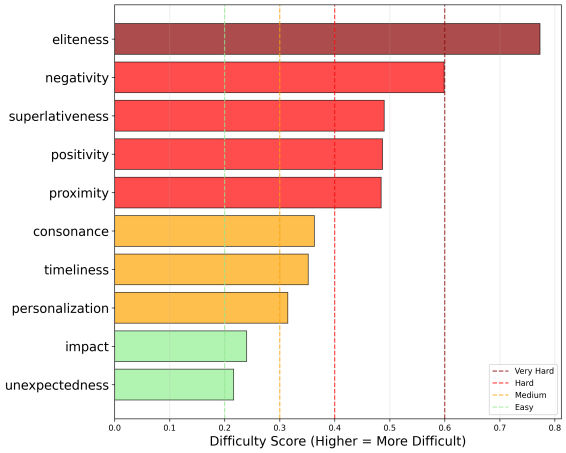

Figure 6: Classification difficulty hierarchy.

This hierarchy suggests that some news value classes are more difficult to learn and/or that there may be issues related to how some of them have been conceptualized when implementing the multilabel BERT classifiers. Unexpecteness and impact, expressing newness, surprise, expressions of unusuality [25] and referring to consequences [6] respectively are the easiest for the model to learn, which may be explained by the fact that such conceptualization may recurrently be realized linguistically through the same specific words and phrases or with a meaning clearly captured through the word embeddings [10] analyzed by the model even when expressed with different surface material. Personalization, timeliness and consonance make out the middle group, suggesting that there may be minor issues, which could suggest issues with class imbalance or that the phenomena captured by the classifier is context-dependent, but learnable, a situation where increasing quality and amount of annotated data for training could lead to significant improvements. For the struggling news value classifiers, proximity, positivity, superlativeness, negativity and eliteness more severe issues seem to be at play, with either extreme class imbalance or heavy difficulties

with conceptualization of the subclasses, possibly requiring contextual knowledge beyond what is available to the model through word embeddings of isolated sentences. To qualify this further, we will inspect the best and worst performing classifiers further, analyzing which linguistic material the classifiers associate with the subclasses. We will also inspect more general trends and explanations, considering support issues, news value complexity and consistency.

4.2 The best performing classifier: Linguistic clarity and simplicity

Unexpectedness is the best performing classifier. Unexpectedness involves constructing an event as rare and unusual [30]. The five unexpectedness subcategories the data has been annotated according to are: ”deviation”, ”contrast”, ”unlikelihood, reversal”, ”surprise markers”, ”discovery”. Figure 7 below shows metrics for each subcategory”™s F1 score and mean confidence.

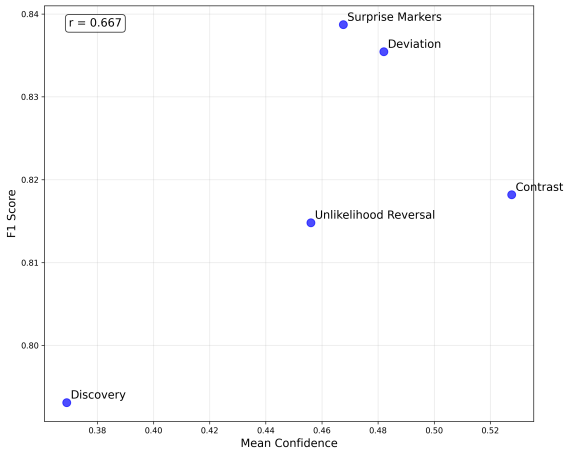

Figure 7: Mean confidence-F1 score plot for unexpectedness subclasses.

This relationship is telling: A relatively higher confidence than F1 may suggest a high degree of certainty when predicting, which is assumed to be attributed to clear linguistic patterns, whereas a relatively higher F1 performance score may suggest variability in linguistic patterns, and the specific metrics may be telling in terms of the extent to which patterns are learnable from the data. ”Contrast” has a relatively high mean confidence as compared to its F1 score. Inspecting the top 10 sentences predicted as expressing contrast using the unexpectedness classifier in a randomly sampled corpus consisting of 10.000 sentences from Danish news discourse from 2022 [13] give some indication of the linguistic patterns clearly associated with this subcategory. The top ranked sentence with a confidence of 0.81 for contrast is: ”Omvendt kan man også spørge sig selv, om det, dengang man lavede aftalen, simpelthen var for naivt at regne med så få fly, og at vi helt kunne melde fra til at bidrage til Natos missioner” (Conversely, one can also ask oneself whether, when the agreement was made, it was simply too naive to count on so few aircraft and that we could completely opt out of contributing to NATO missions). To provide interpretability of model decisions, a perturbation-based word importance visualization was implemented. For each word, the change in news value predictions was measured when that word was masked, then words were highlighted according to how the news value category was affected by their removal, with highlighting intensity reflecting the magnitude of impact. Figure 8 shows words contributing to predicting ”contrast” in the sentence.

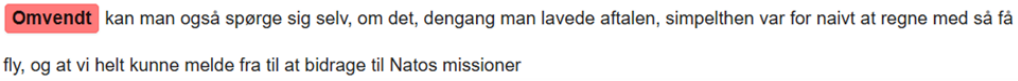

Figure 8: Word contribution for top ranked sentence for the ”contrast” unexpectedness subclass.

Having filtered out words contributing to predicting ”contrast” with scores below 0.05 for the visualization, ”omvendt” (conversely) contributes significantly to the prediction of this sentence as expressing contrast with high confidence with a score of 2.07. Summed with other low-scoring words, the sentence gets a predicted rating of 4.25 of 5 for ”contrast”. For top 2-10 sentences, words such as ”anderledes” (different, 1.96; 1.59), ”modsatte” (opposite, 1.52; 2.33), ”omvendt” (conversely, 2.09; 2.05; 2.26), ”modsætning” (opposition, 2.03) contribute significantly to the predictions of the sentences as expressing contrast. That is, it is primarily single words, especially adverbs and nouns that denote aspects of contrast that contribute to the prediction. Only in one case of the top ten ranking in terms of high confidence, several words contribute to a high score.

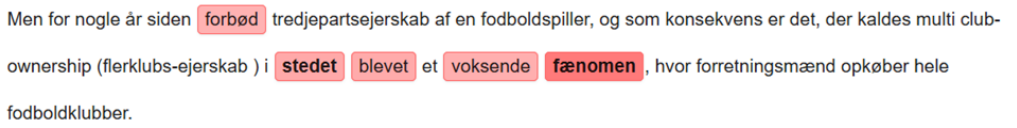

Figure 9: Word contribution for top ranked sentence for the ”contrast” unexpectedness subclass.

This sentence, which can be translated to ”But a few years ago, third-party ownership of a football player was banned, and as a consequence, what is called multi-club ownership has instead become a growing phenomenon, with businessmen buying up entire football clubs”, is rated as 4.11 out of 5 for contrast, with ”fænomen” (phenomenon), ”stedet” (instead), ”blevet” (become), ”voksende” (growing), ”forbød” (banned) contributing the most. Seemingly, the model picks up on a contrast between a past and resent practice of having businessmen buying clubs rather than football players.

For the unexpectedness subcategories that perform with the highest F1, ”surprise markers” and ”deviation (see Figure 11 above), the model is not quite as confident when making predictions. Inspecting the top 10 sentences for surprise markers shows that expressions of surprise using different word classes dominate ”overraskelse” (surprise, 3.06; 2.90; 3.13; 2.91), ”overraskende” (surprising, 2.98; 2.89), ”chok” (shock, 2.89; 2.91), ”overrasket” (surprised, 3.10), ”pludselig” (suddenly, 3.03). The top ten sentences give some indication of why predicting this subclass causes some trouble despite high F1 scores. Firstly, there is some variability in the type of sudden change being expressed, and thus potentially the types of contexts unexpectedness are being expressed in, with words with positive (surprise), negative (shock) and more neutral (suddenly) implications. Secondly, as exemplified in two of the sentences, the expression of being surprised may be negated and not caught by the model, as in ”Tsitsi Dangarembga var ikke overrasket over dommen, fortalte hun til BBC efter retsmødet” (Tsitsi Dangarembga was not surprised by the verdict, she told the BBC after the hearing). This means that the context is a factor that plays into whether a word realizes a specific news value or not, which the model is struggling with in this case and may play a part in the more conservative approach to predicting surprise.

For the top 10 sentences predicted to express ”deviation”, some of the same words as for ”surprise” contribute the most: ”overrasket” (surprised, 2.49; 2.76), ”pludseligt” (suddenly, 2.3), ”pludselig” (suddenly, 2.55; 2.48, 2.96; 1.20) ”chok” (2.54), but we also find words with slightly different meaning, expressing that a phenomenon is surprising in the sense of being unusual: ”uset” (unprecedented, 1.70), ”utrolige” (incredible, 2.59). It also characterizes the sentences rated as expressing deviance that quite a few words feature above the threshold of 0.05, meaning that several

words contribute more significantly. The words above the threshold are: ”sygdommen” (the illness), ”natten” (the night), ”familiemennesket” (the family person), ”giver” (makes), ”mening” (sense), ”oplevede” (experienced), ”aflyser” (cancel), ”ringer” (call), ”hun” (she), ”resulteret” (resulted), ”den” (it), ”ekstrem” (extreme), ”var” (was), ”hidtil” (until now), ”omløb” (circulation), ”opstod” (arose), ”fordi” (because). To some extent this word list points at how deviance may be related to positioning persons or groups as experiencing phenomena or doing actions that are framed as abnormal which may be realized using a range of linguistic means.

4.3 The worst performing classifier: Linguistic complexity and data imbalance

This linguistic clarity contrasts sharply with news values that require complex contextual understanding, exemplified with the worst performing classifier is eliteness. Eliteness is ”discursively constructed as of high status or fame in the eyes of the target audience” [25] and has been conceptualized for classification purposes in the six subclasses ”media elite”, ”expert elite”, ”cultural elite”, ”business elite”, ”political elite” and ”institutional status”. For most subclasses, eliteness is accomplished by mentioning people or groups of people within various areas in ways that indicate or imply having high status. For ”institutional status”, organizations and institutions associated with status may be referred to. The metrics for each subcategory”™s F1 score and mean confidence can be found in Figure 10 below.

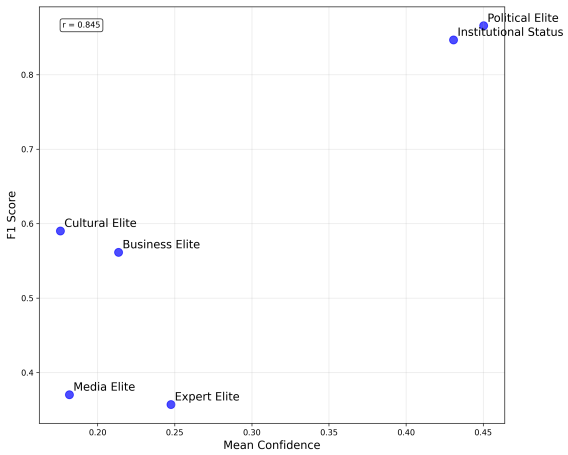

Figure 10: Mean confidence-F1 score plot for eliteness subclasses.

The low performance of some subclasses is clearly related to support with media elite, expert elite, cultural elite and business elite having support of 12, 14, 24 and 26 as compared to institutional status and political elite having support of 65 and 63. But this problem has its source in previous conceptualizations, sampling and annotation of the classifier. Initially when implementing the classifiers using rules, eliteness was defined using 27 subclasses divided into eight groups, which were later merged into six subclasses. Taking ”˜political elite”™ and ”˜media elite”™ as examples, from the outset political individuals were defined as a class with subgroups of political eliteness types, including expressions and titles related to the government, royalty and international politics. Media eliteness on the contrary was a class with much fewer word patterns associated with it, such as editors in chief, commenters and influencers. Thus, the linguistic pointers indexing these two classes differ greatly in frequency, in how eliteness is marked in relation to politics and media topics, thus creating a definition or boundary-case problem, and they will differ in variability of contexts in which the eliteness types may be realized. When classes are imbalanced from the outset, merging classes and sampling more data to annotate on this basis for model implementation will perpetuate data bias and imbalance.

Inspecting the top 10 sentences for ”˜political elite”™ and ”˜media elite”™, we can get more insight into what linguistic information the model presently uses to predict these subclasses. Interestingly, for the ”˜political elite”™ subclass several words consistently contribute to the ratings when making predictions. The sentence attributed to political elite with the highest confidence, translated as ”Watch as Lars Løkke Rasmussen presented his new party, the Moderates, on Constitution Day 2021 in the video below”, exemplifies this as five words are marked as contributing to the ranking of the sentence as indexing political elite.

Figure 11: Word contribution for top ranked sentence for the ”political elite” eliteness subclass.

In fact, a total of 58 words in the 10 sentences the model is most confident about attributing to political eliteness are above the threshold of 0.05. The words include names of political party leaders, political figures, country names, words associated with election, the parliament, and political organization, but also verbs such as ”lyder”, (sounds) ”skal” (must), ”behandles” (handled), ”flyve” (fly) and nouns such as ”synspunkter” (points of view) and ”sag” (case) that may be associated with political processes and phenomena depending on the context. It may be somewhat surprising that a variety of linguistic elements recurrently contribute to attributing a sentence to political eliteness as eliteness mainly has to do with people. However, on the other hand, it points to how eliteness is not simply a matter of mentioning specific people or institutions, but with how status is attributed to them, by e.g. giving them titles, using attributes and describing them as accomplishing matters and acting in specific contexts. This may suggest that another NLP approach, Named Entity Recognition [28], may not be the right way to go for identifying eliteness subcategories, or that this approach should be combined with an approach as the one pursued in this paper where a sophisticated LLM ”˜learns”™ to predict phenomena based on a large set of annotated sentences.

For the subclass ”media elite” not fewer than 77 words in the ten sentences which the model is most confident about predicting to be associated with this subclass are above the threshold of 0.05. However, while some of the words might be associated with eliteness broadly, including ”seniorforsker” (senior researcher), ”ekspert” (expert), ”vicedirektør” (deputy director), ”folkekirken” (the national church), very few have a direct link to media. Those words are ”dokumentar” (documentary), ”rapport” (report), ”forum”, and even these are boundary cases. In fact, it is difficult to see the link to media eliteness in the top ranked sentence below which translates to ”˜Senior researcher on sensational murder: ”It’s obvious that the Ukrainians are behind it”.

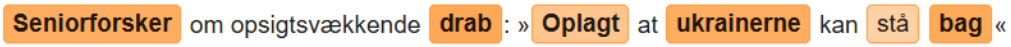

Figure 12: Word contribution for top ranked sentence for the ”media elite” eliteness subclass.

The words the model uses for predicting media eliteness clearly show that the model does not have a clear conceptualization of media eliteness, but that the model is capturing relevant and more general aspects of attributing status to people. This is not surprising given the low support when training the model. One way forward could be to use active learning [26], sampling more data focused on the classes with low support. Another way forward could be to reconsider the conceptualizations of the eliteness classes as such, for example merging media elite and cultural elite before sampling and annotating data for further training.

4.4 Support issues, news value complexity and consistency

Support issues reveal deeper conceptual problems. Subclasses with very low support systematically perform badly. Besides the badly performing eliteness subclasses, the negativity, positivity, superlativeness, and proximity news value classifier all have a subcategory that stands out with poor performance coupled with a low support of 16 or below. The proximity subclass ”geographic proximity, countries” is an extreme case with a support of only 3 and an F1 score of 0.4. The extreme low support for ”geographic proximity, countries” may in part be explained by the fact that this news value has been defined such that the mentioning of countries must be judged to have relevance to Denmark when annotating, presumably making it very difficult for the model to recognize a pattern and to sample data for annotation. As country mentionings would not be the most obvious manner of marking proximity in news discourse, as proximity is discursively realized in cases where the ”event/issue is discursively constructed as geographically or culturally near the target audience” [25], it could be considered to remove this subclass altogether.

For the mentioned news values with subclasses with very low support, it is relevant to investigate the relationship between class imbalance and performance further. Figure 13 below shows the relationship between F1 CV and support CV.

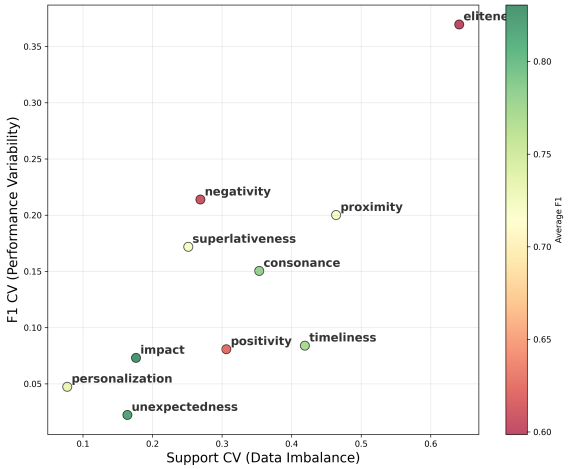

Figure 13: Support CV-F1 CV for the 10 news value classifiers.

The plot shows that for eliteness both values are large. For news values such as negativity and superlativeness, F1 CV is quite high compared to the support CV, potentially suggesting issues related to several subclasses or even the conceptualization of the news values. For timeliness and positivity, there is a relatively low F1 CV compared to the support CV scores which may suggest an issue linked to low support related to a single or few subclasses.

It is also worth considering whether there is an optimal number of subclasses to achieve a consistently performing multilabel classifier. For the news value classifiers as conceptualized here, it turns out not to be the case. Figure 14 shows the relationship between average CV and the number of subclasses each news value has.

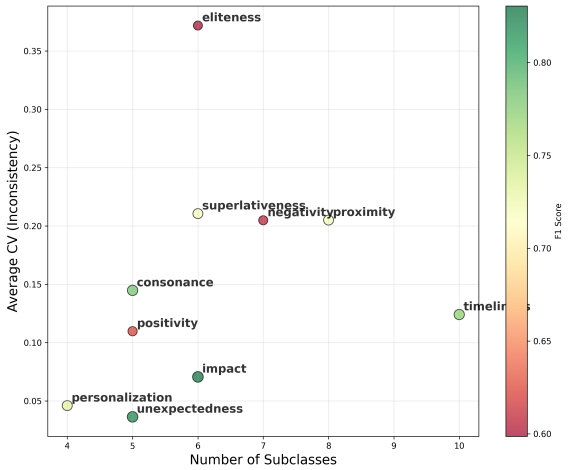

Figure 14: Plot of the relationship between number of subclasses and average CV for the 10 news value classifiers.

As the figure illustrates, timeliness which has 10 subclasses performs quite consistently, whereas eliteness and superlativeness with six subclasses, negativity with seven subclasses and proximity with eight subclasses have a higher CV on average. This pattern may be explained by the differences in what the news value classifiers conceptualize, with timeliness subclasses having to do with references to recent or upcoming events using indicators such as ”yesterday”, ”on Monday” etc., and with e.g. eliteness subcategories referencing more complex phenomena as already discussed.

Plotting the relationship between average F1 score for each news value classifier and its consistency score measured as the inverse of its average CV as in Figure 15 below gives some more indications of the potential causes for why some models struggle.

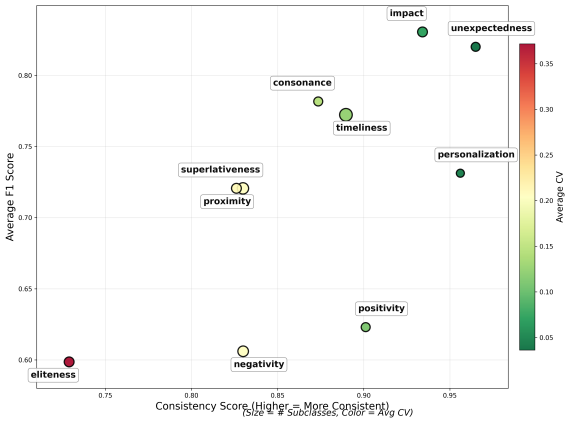

Figure 15: Consistency score-average F1 score plot for the 10 news value classifiers.

Figure 15 confirms some previous observations about news value classes that tend to perform relatively good or badly, but here we can notice that even with reasonable consistency, similar to other news values, negativity and positivity have quite low average F1 scores. This could suggest some more fundamental issues with the conceptualization of these news values, implying that the concepts expected to link the subclasses together are semantically complex or rare. Ways of handling these issues could be to add sentiment scores as information to help the classifiers recognize patterns related to negativity or positivity or simply substitute the two classifiers with

sentiment scores.

5 Discussion: Implications for computational journalism

The analysis shows that a well-established set of journalistic concepts known to be communicated in journalistic discourse is realized and analyzed very differently by a BERT-based transformer model. Some news values are predominantly realized using consistent linguistic markers, whereas others are more context -dependent and some may even require pragmatic reasoning and contextual knowledge outside of what is being expressed linguistically.

Traditional corpus linguistic methods are insufficient to identify complex patterns of linguistic markers associated with news values and their intensity, having previously focused on top 200 frequent lemmas to judge their potential as news value markers through collocation analysis [25], identifying semantic clusters in frequent words [19], or comparing news value distributions in annotated corpora of 100 articles [30].

The perturbation-based analysis implemented after training a BERT-based model on annotated data as explored in this paper shows how the trained classifier models make decisions on the extent to which a sentence should be associated with a news value, even on subclass level. This step bridges computational methods and qualitative discourse analysis, providing explainability which is in demand both in research communities and in relation to practical implementation [9].

The 10 models conceptualize different phenomena in news discourse using linguistic means and have shown different degrees and types of learnability struggles. For the best performing models, simply adding more annotated data using active learning may improve performance further. In some cases, the main issue seems to be single subclasses most of which might even not be key to grasping the news value concept. In those cases, the solution could be to remove the subclass or merge it with another subclass in case of semantic overlap. For some classifiers, a more fundamental reconceptualization of the news value and the methods used for identifying it could be considered, e.g. adding sentiment scores and Named Entity Recognition (NER) as supplementary information for the classifier to have access to for making predictions. A hybrid approach could also be considered, e.g. implementing topic modelling (e.g. BERTopic [15] or similar) to model the complex relationships between linguistic news value markers and topics. If implementing such an approach, the news value conceptualizations of identified subclasses could be reconsidered, as some news value classes in part identify topics, i.e., the eliteness subclasses ”business elite”, ”political elite” etc. Thus, methodological pluralism, i.e. combining different NLP techniques depending on the phenomenon to be classified, may be a valuable way forward to address the specific theoretical and methodological problems encountered in the case of transformer model-based news value classification.

6 Conclusion

This analysis reveals that journalistic concepts exist on a learnability hierarchy determined by their linguistic complexity, with significant implications for computational journalism and discourse analysis. Comparing 10 BERT-based multilabel classifiers trained on LLM-annotated sentences (rated 1-5 for news value subcategories' fit) reveals unequal learning across concepts. This paper analyzed performance differences, including class imbalance, consistency, and linguistic feature attribution patterns. Best-performing models attribute importance to single words when predicting, while struggling models distribute importance across many words per sentence. These findings illuminate how news values are realized linguistically and provide pathways for increasing explainability in large language model text classification.

References

[1] Adelakun, Najeem Olawale and Baale, Adebisi Abimbola. “Sentiment analysis of financial news using the bert model”. In: *ITEGAM-JETIA 10*, no. 48 (2024), pp. 21–27.

[2] Ahmed, Jeelani and Ahmed, Muqeem. “Online news classification using machine learning techniques”. In: *IIUM Engineering Journal 22*, no. 2 (2021), pp. 210–225.

[3] Al Ghamdi, Mohammed A, Bhatti, Muhammad Shahid, Saeed, Atif, Gillani, Zeeshan, and Almotiri, Sultan H. “A fusion of BERT, machine learning and manual approach for fake news detection”. In: *Multimedia Tools and Applications 83*, no. 10 (2024), pp. 30095–30112.

[4] Alizadeh, Meysam, Kubli, Maël, Samei, Zeynab, Dehghani, Shirin, Zahedivafa, Mohammadmasiha, Bermeo, Juan D, Korobeynikova, Maria, and Gilardi, Fabrizio. “Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning”. In: *Journal of Computational Social Science 8*, no. 1 (2025), pp. 1–25.

[5] Bandari, Roja, Asur, Sitaram, and Huberman, Bernardo. “The pulse of news in social media: Forecasting popularity”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 6. 1. 2012, pp. 26–33.

[6] Bednarek, Monika. “Investigating evaluation and news values in news items that are shared through social media”. In: *Corpora 11*, no. 2 (2016), pp. 227–257.

[7] Burscher, Björn, Odijk, Daan, Vliegenthart, Rens, De Rijke, Maarten, and De Vreese, Claes H. “Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis”. In: *Communication Methods and Measures 8*, no. 3 (2014), pp. 190–206.

[8] Chen, Xinying, Cong, Peimin, and Lv, Shuo. “A long-text classification method of Chinese news based on BERT and CNN”. In: *IEEE Access 10* (2022), pp. 34046–34057.

[9] Deck, Luca, Schoeffer, Jakob, De-Arteaga, Maria, and Kühl, Niklas. “A critical survey on fairness benefits of explainable AI”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024, pp. 1579–1595.

[10] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. “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)*. 2019, pp. 4171–4186.

[11] Fruttaldo, Antonio, Venuti, Marco, et al. “A cross-cultural discursive approach to news values in the press in the US, the UK and Italy: The case of the supreme court ruling on same-sex marriage”. In: *ESP Across Cultures 14* (2017), pp. 81–97.

[12] Galtung, Johan and Ruge, Mari Holmboe. “The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers”. In: *Journal of peace research 2*, no. 1 (1965), pp. 64–90.

[13] Goldhahn, Dirk, Eckart, Thomas, Quasthoff, Uwe, et al. “Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages.” In: *LREC*. Vol. 29. 2012, pp. 31–43.

[14] González-Carvajal, Santiago and Garrido-Merchán, Eduardo C. “Comparing BERT against traditional machine learning text classification”. In: *arXiv preprint arXiv:2005.13012* (2020).

[15] Grootendorst, Maarten. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv preprint arXiv:2203.05794* (2022).

[16] Guo, Jingxuan, Mast, Jelle, and Vosters, Rik. “When socialism meets terrorism: A computer-assisted discursive news values analysis of Chinese newspapers’ coverage of domestic and international terrorist attacks”. In: *Mass Communication and Society 27*, no. 6 (2024), pp. 1495–1528.

[17] Huan, Changpeng. “Leaders or readers, whom to please? News values in the transition of the Chinese press”. In: *Discourse, Context & Media 13* (2016), pp. 114–121.

[18] Huan, Changpeng. “Politicized or popularized? News values and news voices in China’s and Australia’s media discourse of climate change”. In: *Critical Discourse Studies 21*, no. 2 (2024), pp. 200–217.

[19] Javadinejad, Arash et al. “A corpus-assisted approach to discursive news values analysis”. In: *Research in Corpus Linguistics (RiCL) 12*, no. 1 (2024), pp. 1–29.

[20] Kaliyar, Rohit Kumar, Goswami, Anurag, and Narang, Pratik. “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach”. In: *Multimedia tools and applications 80*, no. 8 (2021), pp. 11765–11788.

[21] Lin, Szu-Yin, Kung, Yun-Ching, and Leu, Fang-Yie. “Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis”. In: *Information Processing & Management 59*, no. 2 (2022), p. 102872.

[22] Makki, Mohammad. “‘Discursive news values analysis’ of Iranian crime news reports: Perspectives from the culture”. In: *Discourse & Communication 13*, no. 4 (2019), pp. 437–460.

[23] Makki, Mohammad. “The role of ‘culture’in the construction of news values: a discourse analysis of Iranian hard news reports”. In: *Journal of Multicultural Discourses 15*, no. 3 (2020), pp. 308–324.

[24] Ndama, Oussama, Bensassi, Ismail, et al. “The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news”. In: *Bulletin of Electrical Engineering and Informatics 14*, no. 2 (2025), pp. 1231–1240.

[25] Potts, Amanda, Bednarek, Monika, and Caple, Helen. “How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina”. In: *Discourse & Communication 9*, no. 2 (2015), pp. 149–172.

[26] Prabhu, Sumanth, Mohamed, Moosa, and Misra, Hemant. “Multi-class text classification using bert-based active learning”. In: *arXiv preprint arXiv:2104.14289* (2021).

[27] Rybinski, Krzysztof. “Content still matters. A machine learning model for predicting news longevity from textual and context features”. In: *Information Processing & Management 60*, no. 4 (2023), p. 103398.

[28] Süerdem, Ahmet K and Gümüş, Samet. “Named Entity Recognition for Classifying Techno-scientific Persons: Combining Pre-trained Language Models and Silver Standard Datasets”. In: *Digital Humanities Looking at the World: Exploring Innovative Approaches and Contributions to Society*. Springer, 2024, pp. 211–228.

[29] Tan, Zhen, Li, Dawei, Wang, Song, Beigi, Alimohammad, Jiang, Bohan, Bhattacharjee, Amrita, Karami, Mansooreh, Li, Jundong, Cheng, Lu, and Liu, Huan. “Large language models for data annotation and synthesis: A survey”. In: *arXiv preprint arXiv:2402.13446* (2024).

[30] Yu, Hailing and Zhu, Yuzhi. “The Making of Good News: Discursive Construction of Good News Through News Values”. In: *Journalism Studies 26*, no. 2 (2025), pp. 141–160.