HYBRINFOX at CheckThat! 2024 - Task 2: Enriching BERT Models with the Expert System VAGO for Subjectivity Detection

Notebook for the HYBRINFOX Team at CheckThat! 2024 - Task 2

Morgane Casanova[1,†], Julien Chanson[2,†], Benjamin Icard[3,†], Géraud Faye[5,6], Guillaume Gadek[5], Guillaume Gravier[1,†] and Paul Égré[4,†]

[1]Université de Rennes, CNRS, Inria, IRISA, France
[2]Mondeca, France
[3]LIP6, Sorbonne Université, CNRS, France
[4]Institut Jean-Nicod, CNRS, ENS-PSL, EHESS, France
[5]Airbus Defence and Space, France
[6]Université Paris-Saclay, CentraleSupélec, MICS, France

Abstract

This paper presents the HYBRINFOX method used to solve Task 2 of Subjectivity detection of the CLEF 2024 CheckThat! competition. The specificity of the method is to use a hybrid system, combining a RoBERTa model, fine-tuned for subjectivity detection, a frozen sentence-BERT (sBERT) model to capture semantics, and several scores calculated by the English version of the expert system VAGO, developed independently of this task to measure vagueness and subjectivity in texts based on the lexicon. In English, the HYBRINFOX method ranked 1st with a macro F1 score of 0.7442 on the evaluation data. For the other languages, the method used a translation step into English, producing more mixed results (ranking 1st in Multilingual and 2nd in Italian over the baseline, but under the baseline in Bulgarian, German, and Arabic). We explain the principles of our hybrid approach, and outline ways in which the method could be improved for other languages besides English.

Keywords

Subjectivity, Objectivity, Vagueness, Hybrid AI, VAGO, Large Language Models

CEUR-WS.org/Vol-3740/paper-33.pdf

1. Introduction

Detecting subjectivity in natural language is an important task for a number of applications in the domain of news and communication, with strong ties to disinformation and propaganda that constitute the playground of the HYBRINFOX project. Indeed, objective statements in news can be defined as statements that are open to verification by limiting bias and interpretive disagreement. While an objective statement may turn out false, the information it conveys is generally taken to be trustworthy, because it is prone to independent confirmation and fact-checking.

On the contrary, subjective statements convey personal feelings and opinions. By definition, they are prone to inter-personal disagreement and do not obey the same norms of justification and verification. While subjectivity may appear in explicit opinion papers that are not necessarily considered as propaganda or as manipulation, it is also widely used implicitly in conjunction with false objective statements, or just to bias true objective information.

Task 2 of the CLEF 2024 CheckThat! benchmark [1, 2, 3, 4], running for the second time since 2023, aimed at detecting subjective utterances and thus met the objectives of the HYBRINFOX project that seek to develop hybrid methods for the identification of vague information likely to introduce

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

† MC, JC and BI share first authorship on this paper. GGr and PE are co-lead authors.

morgane.casanova@irisa.fr (M. Casanova); julien.chanson@mondeca.com (J. Chanson); benjamin.icard@lip6.fr (B. Icard); geraud.faye@centralesupelec.fr (G. Faye); guillaume.gadek@airbus.com (G. Gadek); guillaume.gravier@irisa.fr (G. Gravier); paul.egre@ens.psl.eu (P. Égré)

0009-0005-4530-5646 (B. Icard); 0000-0002-2985-5964 (G. Faye); 0000-0002-9114-7686 (P. Égré)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

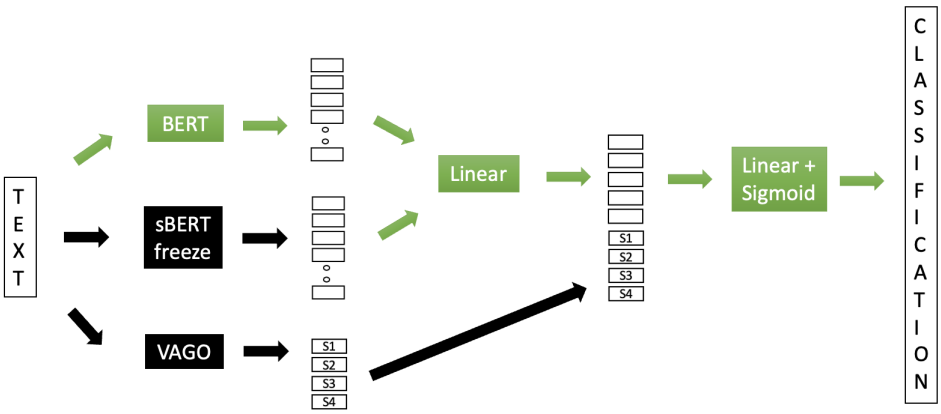

Figure 1: HYBRINFOX model combining BERT, sBERT and the four scores (S1, S2, S3, S4) calculated by the expert system VAGO for the task of objectivity versus subjectivity classification. The green arrows indicate the elements being trained.

or encourage bias (subjectivity, evaluativity). In particular, HYBRINFOX aims to develop tools for measuring linguistic vagueness in texts, taking advantage of a symbolic AI method, VAGO, to improve the performance of deep learning models [5, 6]. The project also explores the boundary between truthful and untruthful uses of linguistic vagueness in discourse [7, 8].

Subjectivity, vagueness, uncertainty, speculations, expressions of opinions are rather difficult to detect automatically in texts as these concepts involve several pragmatic and rhetorical aspects that go beyond the lexical semantics for which most NLP models are designed. The difficulty also translates in the definition of subjectivity, where annotation guidelines have evolved over time – see, e.g., [9, 10, 2]. There is, however, a vast literature on these topics, with a number of recent contributions as part of previous editions of the CheckThat! benchmark [11].

The HYBRINFOX method that competed explores the augmentation of an LLM-based system with vagueness scores given by an expert system primarily based on lexical analysis as illustrated in Figure 1. The method in fact combines a RoBERTa model, fine-tuned for subjectivity detection, a frozen sentence-BERT (sBERT) model to capture semantics, and the scores calculated by the English version of the expert system VAGO [12].[1] Dimensionality reduction operates on the concatenation of the BERT and sentence-BERT embeddings before combining with the vagueness scores as input to a classification linear layer. We justify the selection of this package by comparison with other, less efficient combinations, on the developmental data (see Table 2 below).

The hybrid system was primarily developed for English, and then adapted to other languages, relying on automatic translation into English for the other test languages. The method ranked 1st in English on the evaluation data (with a macro F1 score of 0.7442), but in other languages it produced more mixed results (ranking 1st on Multilingual and 2nd in Italian over the baseline, but 3rd in Bulgarian, 4th in German, and 6th in Arabic under the baseline), likely due to loss in accuracy caused by the translation step.

The paper is organized as follows. We start with a brief state of the art in Section 2. In Section 3 we introduce the expert system VAGO designed to produce vagueness and subjectivity scores. Section 4 reports on experiments conducted towards the choice of an adequate hybrid system, and presents comparative results on the development set. Section 5 analyzes the results obtained in the evaluation phase, including post hoc analyses. Finally, Section 6 concludes with suggestions on how to extend and improve the current results.

[1] In Checkthat! 2024 Task 1, our team explored a distinct but similarly hybrid approach for checkworthiness estimation, see [13] for details.

2. State of the art

Subjectivity and vagueness detection has been addressed with expert systems, often relying on lexical analysis as well as on patterns, viz. [14, 15, 12]. To some extent, [14] also makes use of linguistic heuristics to determine the uncertainty scope within an utterance. More recent work, however, relies on statistical models, such as [16, 17], and all systems at the 2023 benchmark were based on large language models (LLMs) such as BERT or GPT [18, 19, 20, 21, 22, 23, 24]. Most systems explored fine-tuning different language models for subjectivity detection, either in a multilingual or in a language-specific setting. Some systems added components to cope with data imbalance and scarcity, here again leveraging large language models including generative ones for data augmentation [18, 20]. While statistical systems do perform well, they lack explicit features of expert systems that make them explainable and that we believe can make them more efficient. Conversely, expert systems make very limited use of contextual features. The gist of our approach, therefore, is to define a method drawing on both approaches and combining their respective strengths.

3. Symbolic scoring using VAGO

VAGO relies on a symbolic approach to assign scores of vagueness, subjectivity, detail, and objectivity to sentences. Regarding the measure of detail and objectivity of a text, the detection is based in part on identifying named entities (including people, locations, temporal indications, institutions, and numbers), using the open-source library for Natural Language Processing spaCy.[2] Our underlying assumption is that such entities ground the information reported in specific objects and generally leave very limited room for variable interpretation. The more named entities, the more detailed the sentence is likely to be. Given a sentence ϕ, we say that its category is NE if it names an entity. The class NE is not closed, as its members are determined by named entity recognition.

By contrast, the detection of vagueness and subjectivity relies on a closed but evolving lexical database, which consisted of 1,614 terms in English at the time of the CheckThat! 2024 experiment [25]. Derived from a typology of vagueness proposed in [7], this database provides an inventory of lexical items distributed in four categories for vagueness (approximation vagueness (V_A), generality vagueness (V_G), degree vagueness (V_D), combinatorial vagueness (V_C), and an additional category of explicit markers of subjectivity (E_S), not counted as vague.

Expressions of approximation vagueness include modifiers like “approximately”, which relax the truth conditions of the modified expression. Generality vagueness includes determiners like “some” and modifiers such as “at most”. The category of expressions related to degree vagueness and combinatorial vagueness [26] mainly consists of one-dimensional gradable adjectives (such as “tall” and “old”) and multidimensional gradable adjectives, including a number of evaluative adjectives (like “beautiful”, “intelligent”, “good”, or “skilled”). These expressions, unlike expressions of generality vagueness, lack precise truth-conditions, and they leave room for inter-personal disagreement and subjectivity [27, 28, 29, 30]. Finally, explicit markers of subjectivity include separate expressions such as exclamation marks (“!”), first-person pronouns (“I/we”), and some expressive adverbs (“ever”, “of course”). As seen in Table 1, the category V_C is much bigger than the others, and as we will see below it also plays a determinant role in the detection of subjectivity.

Expressions of type NE, V_A, V_G, are treated as factual or *objective*, whereas expressions of type E_S, V_D, V_C, are treated as *subjective*. Table 1 provides the exact numbers of items per category available in the English VAGO database used for the CheckThat! 2024 competition (version of May 2024).

For each measurement, relevant markers are detected and scored from words to sentences. For a given sentence ϕ, its *vagueness score* is calculated as the ratio between the sum of vague words in the

[2] https://spacy.io/

Table 1
Number of entries in the English VAGO database used in CheckThat! 2024.

VAGO categories	Labelling	Number of entries
Approximation	V_A	9
Generality	V_G	35
Degree vagueness	V_D	57
Combinatorial vagueness	V_C	1,500
Explicit subjectivity	E_S	13
All categories		1,614

sentence, $|V|_\phi$, and the total number of words in the sentence, notated N_ϕ. That is:

$$R_{vagueness}(\phi) = \frac{\overbrace{|V_D|_\phi + |V_C|_\phi}^{\text{subjective}} + \overbrace{|V_A|_\phi + |V_G|_\phi}^{\text{objective}}}{N_\phi} = \frac{|V|_\phi}{N_\phi} \tag{1}$$

where $|V_A|_\phi$, $|V_G|_\phi$, $|V_D|_\phi$, and $|V_C|_\phi$ represent the number of terms in ϕ belonging to each of the four vagueness categories.

The *subjectivity score* of a sentence ϕ is calculated as the ratio between the subjective expressions in ϕ, either vague or explicit markers, and the total number of words in ϕ. That is, letting $|S|_\phi = |E_S|_\phi + |V_D|_\phi + |V_C|_\phi$:

$$R_{subjectivity}(\phi) = \frac{|S|_\phi}{N_\phi} \tag{2}$$

The *detail-vs-vagueness score* of a sentence ϕ is defined as the relative proportion of named entities $|NE|_\phi$ in the sentence, compared to the number of vague terms in ϕ (across all categories) $|V|_\phi$:

$$R_{detail/vagueness}(\phi) = \frac{|NE|_\phi}{|NE|_\phi + |V|_\phi} \tag{3}$$

The *objectivity-vs-subjectivity score* of a sentence is defined as the relative proportion of objective expressions in ϕ (objective vagueness and named entities), compared to the number of subjective terms in ϕ. That is, letting $|O|_\phi = |NE|_\phi + |V_A|_\phi + |V_G|_\phi$:

$$R_{objectivity/subjectivity}(\phi) = \frac{|O|_\phi}{|O|_\phi + |S|_\phi} \tag{4}$$

In summary, the symbolic method encodes objectivity in terms of two main dimensions: named entities (NE) and objective vague expressions. And likewise it encodes subjectivity in terms of subjective vague expressions and explicit markers of subjectivity. Hence, vagueness and subjectivity overlap, but neither includes the other, and vagueness does not rule out objectivity. Expressions of detail are all markers of objectivity, but objectivity is a broader category.

VAGO does not consist solely of a lexicon and scoring rules, but it also includes expert rules of vagueness-cancellation (viz. the measure phrase “180cm” in “Mary is 180cm tall” cancels the vagueness and subjectivity of the adjective “tall” when it occurs unmodified as in “Mary is tall”). Quotation marks are also handled as cancelling the subjectivity of terms occurring within the marks. This choice agrees with with the annotation guide proposed in [2], in which reported speech is taken to be conveying objective information on views that may themselves be subjective.

Table 2
Performance of different BERT-based systems tested on the development data (English).

Systems	Threshold	Macro F1	SUBJ F1	
RoBERTa	0.1	0.786	0.7851	
RoBERTa + sBERT	0.15	0.8025	0.8033	
RoBERTa + VAGO Terms	0.3	0.802	0.8108	
RoBERTa + VAGO Scores	0.05	0.817	0.8258	
RoBERTa + sBERT + VAGO Scores	0.1	0.8173	0.8346	◁
RoBERTa + sBERT + VAGO Terms + VAGO Scores	0.2	0.8144	0.8099	

4. Development phase: defining an optimal hybrid system

During the development phase, we tested and compared six variants of our system on the English data. Two variants are purely machine-learning based and serve as a baseline to measure the contribution of VAGO to a hybrid system. The other four variants address different ways of integrating VAGO features to a machine-learning approach. Each system was fine-tuned on the English training data over 30 epochs, with a batch size of 6 and a learning rate of 10e-6. Results for the different systems are reported in Table 2. For languages other than English, we used the English system on an initial translation into English using the DeepL translator[3]. We thus only report results on English for the development phase.

As an initial baseline, a first pure machine-learning system consists in fine-tuning a BERT model for document classification, specifically the "RoBERTa-base" model, with a single value output 1 for subjective, and 0 for objective. We chose RoBERTa based on its good performances on text classification and for reasons of familiarity, leaving open whether other models could be used instead.

The optimization criterion at training is the binary cross-entropy. At test time, we compare the score given by the system to a threshold, utterances with a score above the threshold being deemed as subjective.[4] By varying the threshold, we obtain different trade-offs in terms of misses and false alarms for subjective utterances, yielding ROC curves as reported in Figure 2, where the red curve corresponds to the baseline. We defined the optimal threshold on the test set of the development data, searching for the threshold maximizing the official macro F1 metric. Optimal thresholds found on the validation data along with the corresponding scores are reported in Table 2, where the ◁ symbol shows the system retained.

Due to the fine-tuning of all the parameters, the resulting model is likely to encode mostly lexical information dedicated to the task at hand, disregarding semantics. We thus tested a variant combining the BERT model with a sentence-BERT model whose parameters are frozen. The sentence-BERT model is a BERT model trained within a siamese architecture to yield sentence embeddings that are close one to another in the embedded space for utterances having similar meanings. These models, trained on paraphrasing and on natural language inference tasks, are known to encode semantics to some extent, thus providing our system with an intuition of the meaning to facilitate classification. In the experiments, we used the "distilbert-base-nli-mean-tokens" checkpoint and did not re-estimate the parameters at training time to keep general semantics. Adding a semantic description as input to the classifier improved results as reported in Table 2.

Making use of the expert system VAGO in a BERT-based approach can rely on one of two features: (i) the terms detected by VAGO for the five categories (V_A, V_G, V_D, V_C, E_S) in an utterance, referred to as "VAGO Terms"; (ii) the four vagueness scores described in the previous section, referred to as "VAGO Scores".

In the first case, a straightforward approach simply consists in augmenting the input utterance with the VAGO terms before training a BERT classifier. The input to the BERT classifier thus consists of the utterance to classify, followed by a list of VAGO terms separated from the utterance with a [SEP] token.

[3] https://www.deepl.com/
[4] Note that training is thus done with an implicit threshold equal to 0.5.

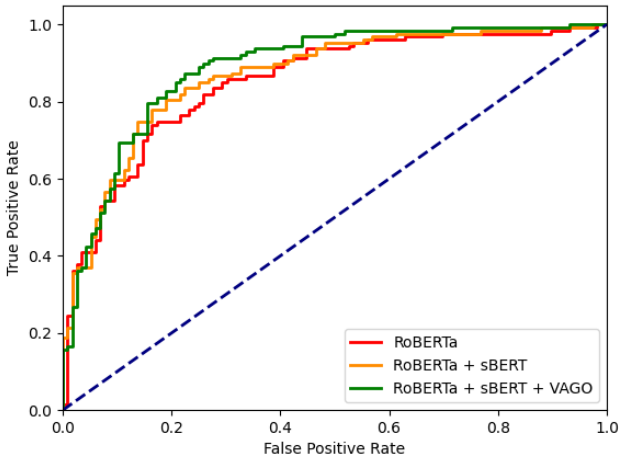

Figure 2: Receiver Operating Characteristic (ROC) Curve for the "RoBERTa", the "RoBERTa + sBERT" and the "RoBERTa + sBERT + VAGO Scores" Systems Predictions.

The idea of this approach, designated as "RoBERTa + VAGO Terms", is to reinforce the decision by emphasizing the terms deemed of interest by VAGO. In the case of VAGO scores, the idea is to combine the BERT utterance embedding with the VAGO scores before the classification. To prevent the four VAGO scores from being overwhelmed by the 768 dimensions of the BERT utterance embedding, we first reduced the dimension of the latter to 5 before concatenation with the VAGO scores. The resulting 9 dimensional feature vector constitutes the input to the classification head. Both strategies improve performance over the BERT baseline, with VAGO scores being slightly more efficient than VAGO terms.

We also combined these last two hybrid approaches with the sentence-BERT embeddings as discussed previously. The combination of BERT and sentence-BERT embeddings with the VAGO Scores defines the official HYBRINFOX system that competed in CLEF, achieving the best results on the development set. It is illustrated in Figure 1 and marked with the symbol ◁ in Table 2. The BERT and sentence-BERT embeddings are concatenated before reducing the 2x768 dimensional vector to 5 with a linear projection. As previously, this last feature vector is concatenated with the VAGO Scores before entering classification layer. We also tested the addition of the VAGO Terms to the BERT encoder in this last architecture, however with no success, yielding a reduction of the subjective F1 score. Interestingly, the ROC curves show a strong benefit for the "RoBERTa + sBERT + VAGO Scores" over the baseline, where a significant improvement of the false positive rate of 0.25 is observed for a fixed true positive rate of 90 %.

5. Evaluation phase: results and analyses

Table 3 presents the evaluation results with the selected "RoBERTa + sBERT + VAGO Scores" system, namely the performance of our method on the evaluation data provided at the time of the contest. Overall, the results were competitive and outperformed the baseline used by the organizers (a logistic regression model) in half of the cases.

In English, our main target and our pivot language, it came out first (out of 15 teams, with a macro-F1 of 0.7442 for Hybrinfox vs 0.6346 for the baseline). It also obtained good scores in Italian (0.7838 vs 0.6503, rank=2/5), and in the Multilingual task (0.6849 vs 0.6697, rank=1/3). The scores were less good and under the baseline in Bulgarian (0.7147 vs 0.7531, rank=3/5), German (0.6968 vs 0.6994, rank=4/4), and Arabic (0.4551 vs 0.4908, rank=6/7).

Notably, results in Arabic were much lower than for other languages, though our scores were of the same order of magnitude as those of other participants. This suggests that the Arabic dataset presents specific properties compared to others, making them worthy of delving further into the data

Table 3
Results of the "RoBERTa + sBERT + VAGO Scores" system on the evaluation phase datasets. Left column: settings at the time of CheckThat!2024; Right column: optimal settings calculated post hoc.

Language	Threshold	Macro F1	SUBJ F1	Optimal Threshold	Macro F1	SUBJ F1
English	0.1	0.7442	0.6	0.1	0.7442	0.6
Italian	0.1	0.7838	0.68	0.1	0.7838	0.68
Arabic	0.1	0.4551	0.27	0.05	0.4924	0.39
Bulgarian	0.1	0.7147	0.65	0.1	0.7147	0.65
German	0.1	0.6968	0.57	0.05	0.7708	0.69
Multilingual	0.1	0.6849	0.63	0.05	0.7100	0.70

and annotations.

We also conducted a post-evaluation analysis of the optimal decision threshold, which was empirically set at 0.1 on the development data. Results are reported in the right-most columns of Table 3. Overall, the threshold of 0.1 appears to be stable across languages, however it is suboptimal for German, Arabic and for the Multilingual dataset. For these three datasets, better performance would have been achieved by lowering the threshold to 0.05. In other words, with a threshold of 0.1, the miss rate for subjective utterances was probably too high to yield a good compromise between recall and precision for an optimal macro F1 score. This observation calls for further work on score normalization towards better stability across languages and datasets.

Upon further analysis, we discovered that our results for Bulgarian could have been improved through additional corpus cleaning. Specifically, square bracket symbols were inadvertently included in the translations into English, which we did not address during the evaluation phase. By removing these brackets, our results for Bulgarian improved significantly, achieving a Macro F1 score of 0.7561 and a SUBJ F1 score of 0.7122 with an optimal threshold of 0.1. This issue with square brackets was also present in other languages, but did not affect the results as significantly as it did for Bulgarian.

6. Conclusion and future work

To solve Task 2 of Subjectivity Detection, we used a hybrid approach combining the rigid symbolic AI system VAGO rules with flexible BERT predictions taking context into account. Our main target was English, one of the languages of VAGO, for which we obtained good results in the development and in the evaluation phases. For other languages, we used the English system on automatic translations obtained by DeepL.

In order to improve our approach, we believe that separate VAGO lexicons should be developed for additional languages besides English and French, in order to get rid of the translation step into English, and to be able to use appropriate BERT and sBERT models for each language. The development of an Italian lexicon was under way at the time of the evaluation, but not sufficiently advanced yet to produce satisfactory results. Alternatively, we could aim for a better control of the quality of the translation into English, our assumption being that the lexicon of objectivity and subjectivity ought to be preserved under translation.

Finally, we stress that our approach of subjectivity detection was developed independently of the task. During the training phase, we noticed that some expressions that VAGO classifies as subjective, like the vague determiner "many", were not systematically associated with the label "subjective" (31 out of 42 sentences including "many" are labelled as objective in the development set provided ahead of the evaluation phase). We did not try to modify VAGO's classification principles on this or other specific entries. Instead, we left it in place in order to see better if our understanding of subjectivity and the understanding of subjectivity proposed in [2] would agree. We were pleased to see that they mostly do, because this lends support to the hypothesis that the expression of subjectivity is characterized in part by the use of a specific lexicon and by specific rhetorical markers.

Acknowledgements

We thank two anonymous referees for helpful comments. This work was supported by the programs HYBRINFOX (ANR-21-ASIA-0003), FRONTCOG (ANR-17-EURE-0017), and THEMIS (n°DOS0222794/00 and n° DOS0222795/00). PE thanks Monash University for hosting him during the writing of this paper, in the context of the program PLEXUS (Marie Skłodowska-Curie Action, Horizon Europe Research and Innovation Programme, grant n°101086295).

References

[1] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. M. Struß, F. Antici, M. Hasanain, J. Köhler, et al., Overview of the CLEF-2023 CheckThat! Lab: Task 2 on subjectivity in news articles, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 236–249.

[2] F. Ruggeri, F. Antici, A. Galassi, K. Korre, A. Muti, A. Barrón-Cedeño, On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection., Text2Story @ European Coference on Information Retrieval 3370 (2023) 103–111.

[3] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[4] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, M. Siegel, M. Wiegand, Overview of the CLEF-2024 CheckThat! Lab Task 2 on subjectivity in news articles, in: [31], 2024.

[5] P. Guélorget, B. Icard, G. Gadek, S. Gahbiche, S. Gatepaille, G. Atemezing, P. Égré, Combining vagueness detection with deep learning to identify fake news, in: IEEE International Conference on Information Fusion, 2021, pp. 1–8.

[6] B. Icard, V. Claveau, G. Atemezing, P. Égré, Measuring vagueness and subjectivity in texts: from symbolic to neural VAGO, in: IEEE International Conference on Web Intelligence and Intelligent Agent Technology, 2023, pp. 395–401.

[7] P. Égré, B. Icard, Lying and vagueness, in: J. Meibauer (Ed.), Oxford Handbook of Lying, OUP, 2018.

[8] P. Égré, B. Spector, A. Mortier, S. Verheyen, On the optimality of vagueness: "around", "between" and the Gricean maxims, Linguistics and Philosophy 46 (2023) 1075–1130.

[9] T. Wilson, J. Wiebe, Annotating opinions in the world press, in: Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue, 2003, pp. 13–22.

[10] C. Kennedy, Two kinds of subjectivity, in: C. Meier, J. van Wijnbergen-Huitink (Eds.), Subjective meaning: Alternatives to relativism, Walter de Gruyter GmbH & Co KG, 2016.

[11] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. S. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouani, Overview of the CLEF–2023 CheckThat! Lab on checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, p. 251–275.

[12] B. Icard, G. Atemezing, P. Égré, VAGO: un outil en ligne de mesure du vague et de la subjectivité, in: Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, 2022, pp. 68–71.

[13] G. Faye, M. Casanova, B. Icard, J. Chanson, G. Gadek, G. Gravier, P. Égré, HYBRINFOX at CheckThat! 2024: Enhancing language models with structured information for checkworthiness estimation, in: [31], 2024.

[14] H. Yang, A. De Roeck, V. Gervasi, A. Willis, B. Nuseibeh, Speculative requirements: Automatic

detection of uncertainty in natural language requirements, in: IEEE International Requirements Engineering Conference, 2012, pp. 11–20.
[15] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: Conference on Empirical Methods in Natural Language Processing, 2003, pp. 105–112.
[16] H. Huo, M. Iwaihara, Utilizing bert pretrained models with various fine-tune methods for subjectivity detection, in: Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data, 2020, pp. 270–284.
[17] S. Sagnika, B. S. P. Mishra, S. K. Meher, An attention-based CNN-LSTM model for subjectivity detection in opinion-mining, Neural Computing and Applications 33 (2021) 17425–17438.
[18] I. Baris Schlicht, L. Khellaf, D. Altiok, DWReCO at CheckThat! 2023: Enhancing subjectivity detection through style-based data sampling, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 306–317.
[19] K. Dey, P. Tarannum, M. A. Hasan, S. R. H. Noori, NN at CheckThat! 2023: Subjectivity in news articles classification with transformer based models, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 318–328.
[20] R. A. Frick, Fraunhofer sit at CheckThat! 2023: can LLMs be used for data augmentation & few-shot classification? detecting subjectivity in text using chatGPT, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 329–336.
[21] F. A. Leistra, T. Caselli, Thesis Titan at CheckThat! 2023: Language-specific fine-tuning of mDeBERTaV3 for subjectivity detection, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 351–359.
[22] G. Pachov, D. Dimitrov, I. Koychev, P. Nakov, 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, 2023, pp. 404=–412.
[23] H. T. Sadouk, F. Sebbak, H. E. Zekiri, ES-VRAI at CheckThat! 2023: Enhancing model performance for subjectivity detection through multilingual data aggregation, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 423–429.
[24] S. Tran, P. Rodrigues, B. Strauss, E. Williams, Accenture at CheckThat! 2023: Impacts of back-translation on subjectivity detection, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2023, pp. 507–517.
[25] G. Atemezing, B. Icard, P. Égré, Vague Terms in SKOS to detect vagueness in textual documents, 2022. URL: https://doi.org/10.5281/zenodo.4718530. doi:10.5281/zenodo.4718530.
[26] W. P. Alston, Philosophy of Language, Prentice Hall, 1964.
[27] C. Wright, The epistemic conception of vagueness, The Southern journal of philosophy 33 (1995) 133.
[28] C. Kennedy, Two sources of subjectivity: Qualitative assessment and dimensional uncertainty, Inquiry 56 (2013) 258–277.
[29] S. Verheyen, S. Dewil, P. Égré, Subjectivity in gradable adjectives: The case of *tall* and *heavy*, Mind & Language 33 (2018) 460–479.
[30] S. Solt, Multidimensionality, subjectivity and scales: Experimental evidence, in: E. Castroviejo, L. McNally, G. Sassoon (Eds.), The Semantics of Gradability, Vagueness, and Scale Structure, Springer, 2018, pp. 59–91.
[31] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.