Eevvgg at CheckThat! 2024: Evaluative Terms, Pronouns and Modal Verbs as Markers of Subjectivity in Text

Notebook for the CheckThat! Lab at CLEF 2024

Ewelina Gajewska[1,*,†]

[1]Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland

Abstract

This work tests performance of simple machine learning algorithms against large language models (LLMs) utilising transfer learning for a binary detection of subjectivity in news articles. Second, the influence of feature normalisation on classification performance is examined. Third, the work measures impact of training data size on subjectivity extraction. The proposed **BERTd** model that makes use of additional information about stance markers in news articles was placed **8th** in the official ranking of the CLEF 2024 CheckThat! lab Task 2: Subjectivity in News Articles competition for English data, achieving 0.70 macro-averaged F_1. Models that distinguish subjective opinions from objective facts could be utilised in studies on information verification (detection of fake news, understood as a mixture of subjective opinion and facts).

Keywords

stance, subjectivity, fake news, text classification, information extraction, opinion mining

1. Introduction

Subjectivity is inherently encoded in language and involves expressions of the speaker's position, attitude, and feelings towards the uttered message [17]. Thus, identification of articles written from a subjective perspective of the author involves detection of stance markers: words that express some form of evaluation or judgement (e.g. words denoting emotional valence), pronouns or modal verbs and passive constructions [16]. This work makes use of syntactic and semantic features that are fed to machine learning algorithms and large language models (LLMs) for a binary detection of news articles written from a subjective versus objective perspective of the author [4]. The datasets have been provided by the organizers of the CLEF 2024 CheckThat! lab [5], which is the international contest on challenging classification and retrieval problems. The aim of this competition is to advance the field of information retrieval from text. To this end, a new approach to subjectivity detection is proposed: a Transformer-based model that makes use of both text content and additional meta features derived from articles. Such an approach shows more consistent results than simple transfer learning (TL; adding a classification layer on top of BERT encoder) fed with textual content only. The current work describes the approach by the *eevvgg* team for Task 2: Subjectivity in News Articles of the CLEF 2024 CheckThat! lab in English news articles.

It is a common approach in information retrieval research to experiment with different text representation models, as in [6] where conversion of text with TF-IDF (Term Frequency Inverse Document Frequency) algorithm outperformed models with so-called Count Vectorizer, that is, a simple frequency count of particular terms in each text sample. Influence of preprocessing techniques on classification performance of deep learning models was tested in [3]. Machine learning methods are among fundamental methods used in the field of natural language processing [2]. They are used, for example, in web search engines for information retrieval [7]. Thereby, a user looking for specific information gets results

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

*Corresponding author.

†These authors contributed equally.

ewelina.gajewska.dokt@pw.edu.pl (E. Gajewska)

https://ewelina04.github.io/ (E. Gajewska)

https://orcid.org/0009-0006-6012-4787 (E. Gajewska)

relevant to the searched topic. Previous studies investigated performance of several machine learning models: logistic regression, SVM, and Naive Bayes for text classification tasks such as recognition of political affiliation of the US presidential candidates from their presidential campaign speeches [1]. Contribution of this work is three-fold: first, examination of machine learning vs. transfer learning approaches for information extraction; second, impact of feature normalisation on classification performance; third, influence of training data size on detection performance. Models for detection of subjective opinions versus objective facts could be utilised, for example, in studies on information verification (detection of fake news, which are a mixture of subjective opinion and facts [8]).

2. Related Work

Differentiation between subjective opinions and objective facts compose a basis of proper journalism. Subjective reporting of news might reflect bias of the author which automatic solutions can identify and tag properly to reduce the spread of fake news, for example. Several work have experimented with subjectivity detection methods from text. Unified method for detection of subjectivity in multilingual text content was proposed in [12]. Fine-tuned ELECTRA large model outperformed other large language models (such as BERT and RoBERTa) for subjectivity analysis in text in the context of fake news detection, achieving 0.983 accuracy [20]. Such analyses have proved to be useful for detecting fake news with a lexicon-based approach [11] as well as in opinion mining using deep learning techniques [18]. The current work extends these studies by combining previous approaches and testing the proposed model in several experimental settings. Team eevvgg proposes a deep learning model that combines the fine-tuned BERT encoder [9] and lexicon-based method for extracting linguistic markers of subjectivity. The proposed architecture is called **BERTd** and is tested against other BERT-based models and machine learning algorithms in two experimental conditions: on a smaller vs. a larger set of data (n=667 vs. n=1511 training samples, and n=166 vs. n=484 test samples, respectively); and on raw vs. normalised values of linguistic meta features.

3. Methodology

3.1. Material

This work deals with binary detection of subjectivity from text material: whether a sentence expresses a subjective view of the author (SUBJ) or presents an objective view on the topic (OBJ). The paper proposes solutions for subjectivity extraction for English data. Initial training data comprises 833 text samples (data units) from newspaper articles: short pieces of text of up to 100 words. Development of a text classification system starts with pre-processing of the data, then splitting it into training and testing sets, extraction of features, model training and model validation. Then, the dataset was divided into two splits: 80% for training purposes and 20% for testing. Results of training and evaluation on this set (called **small set**) are reported in Table 1. Evaluation and training is conducted also on a bigger set - official test set - released by the CheckThat! organizers after submission deadline and training data comprising all available train and dev sets - 1511 in total; evaluation is conducted on 484 samples from the official test. Results of training and evaluation on this set (called **big set**) are reported in Table 2.

3.2. Text Preprocessing

Usually, the first step in information retrieval tasks is to represent the text using a certain model. A common approach is to represent a document as a vector of features - most simple representations of text include Bag-of-Words (BOW) models. Regarding machine learning algorithms TF-IDF method is employed (with max_df=0.75, min_df=2). Transfer learning approaches utilise BERT encoder (BERT base uncased) as a text representation method. In order to improve the predictive performance of these models, several meta features were constructed from the textual content of news articles using the concept of feature engineering [14]. It involves the application of transformation functions on given

features to generate new ones. In order to extract such features from text, it needs to be normalised. Data cleaning involved 3 steps: conversion of text into lowercase, removal of stop-words and punctuation symbols. Then, the text was lemmatised, that is, words were converted to their dictionary forms. Finally, linguistic features were extracted from the clean and lemmatised text of articles. In total, 16 syntactic features (stance markers) were extracted from text samples, specifically, frequency of occurrence of each category of terms in a given text sample. Specific terms that belong to each category are specified below:

1. Subject pronouns: I, you, he, she, it, we, you, they;
2. Object pronouns: me, you, him, her, it, us, you, them;
3. Possessive pronouns: mine, yours, his, hers, its, ours, yours, theirs;
4. Demonstrative pronouns: this, these, that, those
5. Interrogative pronouns: who, whom, which, what
6. Relative pronouns: who, whom, that, which, whoever, whichever, whomever;
7. Indefinite pronouns: all, another, any, anybody, anyone, anything, each, everybody, everyone, everything, few, many, nobody, none, one, several, some, somebody, someone;
8. Reflexive pronouns: myself, yourself, himself, herself, ourselves, yourselves, themselves
9. Modal verbs: must, shall, will, should, would, can, could, may, might;
10. Obligation verbs: need, have to, must, might, may, has to, shall;
11. Frequency adverbs: hardly, ever, rarely, scarcely, seldom, never, sometimes, often, always, usually, normally;
12. Comparison adverbs: bad, badly, worse, worst, good, better, well, best, far, farther, further, farthest, furthest, little, less, least, few, somehow;
13. Reporting verbs: advise, agree, challenge, claim, decide, demand, encourage, invite, offer, persuade, promise, refuse, remind, say;
14. Pronouns: a sum of features 1-8;
15. Emotive words: words associated with an expression of emotions marked as such in the lexicon of emotion-laden terms [13];
16. Polarising words: words associated with inducing social polarisation (dividing the society into ‘us’ versus ‘them’ groups) marked as such in the lexicon of polarising terms [19];

The summary of preprocessing procedure is illustrated in Figure 1.

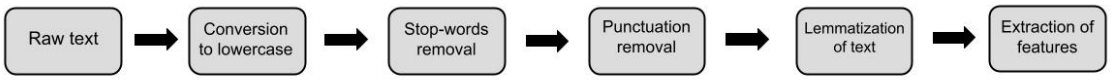

Figure 1: Pipeline for extraction of meta features (linguisitc markers of subjectivity).

3.3. Tools

In the light of available algorithms, scikit-learn map of estimators is followed in order to chose four of them: Naive Bayes (NB), logistic regression (LR), decision trees (DT) and decision forests (DF). The Naive Bayes is one of the simplest and most popular models in the field of supervised machine learning [2] and amongst the most efficient and effective classifiers [10]. Classification with the use of this estimator is based on the calculated probabilities: probability of a certain label for a given data point is estimated through multiplying the probability of this label by a sum of probabilities of all features describing this data point given the label [7]. The NB algorithm learns probabilities based on prior distribution across classes from the training data following the assumption that all features are independent. SVM models learn to categorise data into separate classes by building a margin in the feature space that

minimizes the distance between each class and that margin [21]. The logistic regression algorithm calculates the log-odds (converted into probabilities) of an event as a linear combination of one or more independent variables. The Decision Tree model predicts the value of a target variable by learning decision rules, which are inferred from the training data. DT builds a tree-like structure from these rules by splitting data into subsets based on the values of particular features until a stopping criterion is met. DTs have two main advantages of being simple to understand and interpret [15]. A decision tree forest is an ensemble learning method that combines outputs of multiple decision trees to reach the final result. Finally, the suitability of large language models for subjectivity detection is investigated. Specifically, BERT uncased model[1]) that produces contextualised text embeddings and achieves state of the art results in most information retrieval tasks. Transfer learning paradigm is utilised for training BERT-based models. Tensorflow and transformers libraries are employed for their implementation.

3.4. Experimental Settings

Machine Learning. Default settings of hyper-parameters set by the authors of the scikit-learn library are employed in machine learning estimators. TF-IDF method is utilised as a text representation method. Textual features are combined with (normalised) meta features described in Section 2.1 and fed to these algorithms.

Transfer Learning. Regarding TL with LLMs, BERT is utilised for as a text encoder, specifically the CLS token. Then, additional layers are added on top of embeddings returned by the BERT-encoder. Two BERT-based models are proposed: **BERTs** comprises of two fully-connected layers[2] of size 128 and a 0.5 dropout rate between them; in **BERTd** two fully-connected layers of size 246 and 32, separated by a 0.4 dropout layer are attached to the BERT encoder. In addition **BERTb** is utilised a baseline model with only a classification layer added on top of the BERT encoder. Cross-entropy is employed as the loss function and rectified linear unit (ReLU) as the activation function in all hidden layers. Classification layer comprises of two units, which in combination with the softmax function returns probabilities that a given text sample belong to class "OBJ" and "SUBJ". Learning rate is set to 5e-5 as advised by the authors of the transformers library. BERT-based models are trained for either 2, 3 or 4 epochs and the best result is reported. Tensorflow implementation of these networks and functions is employed.

Evaluation metrics. Three metrics are employed for evaluation purposes: weighted F_1 (Eq. 2), macro-averaged F_1 (Eq. 3) and accuracy (Eq. 4).

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

$$WeightedF1 = \frac{\sum_{i=1}^{n} \times w_i \times F1_i}{n} \tag{2}$$

$$MacroF1 = \frac{\sum_{i=1}^{n} \times F1_i}{n} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where TP: true positives, FP: false positives, TN: true negatives, FN: false negatives.

4. Results

Results on a small train set, reported in Table 1, indicate that **BERTb**, a simple transfer learning approach, outperforms all other classifiers. Nonetheless, all tested models perform above baseline using

[1] https://huggingface.co/google-bert/bert-base-uncased
[2] https://keras.io/api/layers/core_layers/dense/

Table 1
Results: **small set**. Classification performance of machine learning (ML) and transfer learning (TL) models. A single asterisk signals models that perform above baseline in all three metrics (accuracy, weighted (w-F_1) and macro (m-F_1) F_1 scores). Double asterisk signals best model in each category of machine learning (ML) and transfer learning (TL) models. In bold the best performing model is marked.

Category	Model	Meta features	w-F_1	m-F_1	Accuracy
Baseline	Stratified	-	0.54	0.51	0.54
ML	NB*	1-16	0.56	0.51	0.63
	LR*	1-16	0.57	0.53	0.61
	DT*	1-16	0.57	0.55	0.57
	RF*	1-16	0.53	0.48	0.59
	NB*	normalised 1-16	0.68	0.57	0.68
	LR**	normalised 1-16	0.70	0.60	0.71
	DT*	normalised 1-16	0.62	0.52	0.61
	RF*	normalised 1-16	0.70	0.57	0.72
TL	**BERTb****	-	**0.78**	**0.76**	**0.78**
	BERTs*	1-16	0.76	0.75	0.76
	BERTd*	1-16	0.75	0.73	0.75
	BERTs*	normalised 1-16	0.76	0.75	0.76
	BERTd*	normalised 1-16	0.74	0.73	0.74

prior label distribution. Normalisation of additional features boosts performance for ML algorithms (NB, LR, DT, RB), although BERT-based models note minor differences in performance. All models developed with a TL approach outperform all ML algorithms achieving, on average, 27% higher results (23% for normalised meta features and 30% for models without normalisation or meta features) in terms of macro F_1.

The proposed **BERTd** model (officially submitted to the task), consisting of two hidden layers and fed with data **without normalisation of meta features**, outperforms other solutions in turn on the bigger set (see Table 2). The difference in performance between ML and TL approaches decreases to 20% - due to lower macro F_1 of two TL models: BERTb and BERTs. Compared to BERTb and BERTs, **BERTd** notes smaller differences in performance between small and big training sets. Thus, its performance is more stable across data than BERTb and BERTs. Nonetheless, all models outperform a baseline classifier using class prior distribution. All **TL models and the LR classifier** achieve also higher macro F_1 than the baseline provided by the CheckThat! organisers.

Table 2
Results: **big set**. Classification performance of machine learning (ML) and transfer learning (TL) models. A single asterisk signals models that perform above baseline in all three metrics (accuracy, weighted (w-F_1) and macro (m-F_1) F_1 scores). Double asterisk signals best model in each category of machine learning (ML) and transfer learning (TL) models. A plus indicates models that achieve higher results than the baseline provider by CheckThat! organizers. In bold the best performing model is marked.

Category	Model	Meta features	w-F_1	m-F_1	Accuracy
Baseline	Stratified	-	0.60	0.51	0.57
	LR Organizers	-	n/a	0.64	n/a
ML	NB*	normalised 1-16	0.65	0.55	0.64
	LR**+	normalised 1-16	0.71	0.61	0.71
	DT*	normalised 1-16	0.64	0.54	0.62
	RF*	normalised 1-16	0.68	0.55	0.69
TL	BERTb*+	-	0.78	0.70	0.78
	BERTs*+	1-16	0.78	0.69	0.80
	BERTd**+	**1-16**	**0.79**	**0.71**	**0.79**
	BERTs*+	normalised 1-16	0.79	0.70	0.80
	BERTd*+	normalised 1-16	0.78	0.71	0.78

5. Conclusion

Once again, **transfer learning outperformed** simpler machine learning approaches for information extraction from text. All BERT-based models (with or without syntactic features) achieved substantially higher results for a binary detection of subjectivity than 4 ML algorithms. **BERTd** fed with additional features (syntactic features of marker) shows more consistent performance than a baseline BERTb model comprising of BERT encoder and a classification layer. Normalisation of meta features was found to boost performance for ML models. Increase of training data size had almost no impact on prediction performance for ML models and a negative influence for TL solutions in terms of macro F_1 metric.

Limitations

Baselines. The current work would benefit from a more thorough analysis of results of the proposed models against other systems developed on the employed dataset.

Generalisability of performance. The proposed architecture notes satisfactory performance on the utilised dataset (ranking 8th in the official leaderboard of the competition on English data), however, its robustness is yet to be tested, for example, in scenarios with different training datasets or multilingual data.

Ablation studies. Ablation analysis in future work could measure the impact of individual features on the final performance of the proposed subjectivity detectors.

References

[1] Acharya, A., Crawford, N., & Maduabum, M. (2016). *A nation divided: Classifying presidential speeches*.

[2] Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. Springer Science & Business Media.

[3] Alshdaifat, E. A., Alshdaifat, D. A., Alsarhan, A., Hussein, F., & El-Salhi, S. M. D. F. S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, *6*(2), 11.

[4] Antici, F., Ruggeri, F., Galassi, A., Korre, K., Muti, A., Bardi, A., & Barrón-Cedeño, A. (2024, May). A Corpus for Sentence-Level Subjectivity Detection on English News Articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) (pp. 273-285).

[5] Barrón-Cedeño, A. et al. (2024). The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In: Goharian, N., et al. *Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science*, vol 14612. Springer, Cham. https://doi.org/10.1007/978-3-031-56069-9_62

[6] Basarkar, A. (2017). *Document Classification using Machine Learning*.

[7] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

[8] De Grandis, M., Pasi, G., & Viviani, M. (2019, July). Multi-criteria decision making and supervised learning for fake news detection in microblogging. In *Workshop on Reducing Online Misinformation Exposure* (pp. 1-8).

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[10] Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis, 21*(3), 267-297.

[11] Jeronimo, C. L. M., Marinho, L. B., Campelo, C. E., Veloso, A., & da Costa Melo, A. S. (2019, December). Fake news classification based on subjective language. In *Proceedings of the 21st*

International Conference on Information Integration and Web-based Applications & Services (pp. 15-24).

[12] Karimi, S., & Shakery, A. (2017). A language-model-based approach for subjectivity detection. *Journal of Information Science*, 43(3), 356-377.

[13] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3), 436-465.

[14] Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. S. (2017, August). Learning Feature Engineering for Classification. In *IJCAI* (Vol. 17, pp. 2529-2535).

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine learning research, 12*, 2825-2830. https://scikit-learn.org/stable/

[16] Reilly, J., Zamora, A., & McGivern, R. F. (2005). Acquiring perspective in English: the development of stance. *Journal of pragmatics*, 37(2), 185-208.

[17] Ruggeri, F., Antici, F., Galassi, A., Korre, K., Muti, A., & Barrón-Cedeño, A. (2023). On the Definition of Prescriptive Annotation Guidelines for Language-Agnostic Subjectivity Detection. *Text2Story at ECIR*, 3370, 103-111.

[18] Sagnika, S., Mishra, B. S. P., & Meher, S. K. (2021). An attention-based CNN-LSTM model for subjectivity detection in opinion-mining. *Neural Computing and Applications, 33*(24), 17425-17438.

[19] Simchon, A., Brady, W. J., & Van Bavel, J. J. (2022). Troll and divide: the language of online polarization. *PNAS nexus*, 1(1), pgac019.

[20] Vieira, L. L., Jeronimo, C. L. M., Campelo, C. E., & Marinho, L. B. (2020, November). Analysis of the subjectivity level in fake news fragments. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (pp. 233-240).

[21] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6), 282-292.