BEYOND THE OBVIOUS: DETECTING IRONY IN DUTCH NEWS HEADLINES

NOA MOLLEE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2072369

COMMITTEE

First reader: Harm Brouwer
Second reader: Julija Vaitonyte

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2024

WORD COUNT

7576

BEYOND THE OBVIOUS: DETECTING IRONY IN DUTCH NEWS HEADLINES

NOA MOLLEE

Abstract

One limiting factor for sentiment analysis is the presence of irony in datasets. To improve sentiment detection, robust sarcasm detection algorithms are necessary. A larger body of research has already been conducted to detect sarcastic utterances in English data, achieving high performance with both traditional and newer machine learning tools. However, the focus has been on the accuracy rather than the explainability of the models, and very few studies have applied these methods to Dutch datasets.

This paper explores applying explainable AI (XAI) techniques to sarcasm detection in Dutch, focusing on a dataset containing Dutch news headers. This study compares the performance of a traditional SVM model with the state-of-the-art BERTje architecture, addressing key questions about the transferability of English sarcasm detection methods to Dutch, the comparative performance of SVM versus BERTje, and the insights revealed through a local explainability model (SHAP) and global introspection methods (ALE and LIG).

In previous work on Dutch social media data, SVM outperformed BERTje. However, for the SVM and BERTje models created for this thesis, the opposite was true: BERTje was better than SVM in every aspect. Although this does not fit with the body of work on Dutch data, it is in line with the English sarcasm detection research.

SVM was analyzed both locally and globally using SHAP and ALE plots, while BERTje was inspected with Layer Integrated Gradients. The SVM model mostly relies on the number of nouns, proper nouns, coordinating conjunction, as well as the polarity differences. It tends to classify short sentences with changes in sentiment as sarcastic. BERTje classifies most tokens related to politics and violence as genuine, whereas more everyday words are generally seen as sarcastic.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The sarcastic headline dataset has been acquired from the Kaggle. This dataset is publically available, and all headlines contained in this dataset are also publically available on nu.nl or speld.nl. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. Most code is my original work using well-known Python libraries, such as numpy (Harris et al., 2020) and pandas (pandas development team, 2020), as well as the documentation of the libraries I used, such as SHAP (Lundberg & Lee, 2017) and ALE (Apley & Zhu, 2020). External sources, such as Github repositories or ChatGPT, were also used to write parts of the code. Where this is applicable, this is explicitly stated in the code. For checking of spelling and grammar, the free version of Grammarly was used. A generative language model, ChatGPT, was used to suggest a title, as well as to proofread this paper. However, aside from the title, no text or code was artificially generated. For managing references, Zotero was used. The typesetting tool used was LaTeX in the Overleaf editor. No other tools were used to assist with this work.

2 INTRODUCTION

Irony is used freely in our everyday speech (Amante, 1981). The terms 'irony' and 'sarcasm' are often regarded as synonyms, as both irony and sarcasm are a form of 'metaphorical speech': words or sentences that are meant to convey the opposite meaning (*Merriam-Webster*, n.d.). However, there are slight differences between the two terms. As opposed to irony, sarcasm can be used as an indirect form of verbal conflict (Bowes & Katz, 2011), making it biting and antagonistic. However, sarcasm is also used to refer to general, non-aggressive verbal irony (Filatova, 2012): metaphorical speech used in spoken or written language. On the other hand, the term irony encapsulates both verbal and situational irony. In situational irony, there is a clash between the expected outcome and the real situation (Lucariello, 1994), like a fire station burning down, or a chef failing to boil an egg. This can be seen as the inherent irony within an event, rather than irony within a sentiment. *Irony* and *sarcasm* are often inflated to both mean general metaphorical speech and are therefore often used interchangeably, both colloquially as well as within the scientific literature. For the rest of this paper, *irony* and *sarcasm* will be used to refer to verbal irony. When specifically discussing situational irony, this will be clearly stated.

As irony is prevalent within society, its automated detection is becoming an increasingly popular field of research (Yacoub, Slim, & Aboutabl, 2024). This is because it closely ties in with a related field of research: sentiment analysis. In traditional sentiment analysis, the sentiment is classified by either machine learning, or using a lexicon containing sentiment markers (Medhat, Hassan, & Korashy, 2014). Most datasets used to train these classification models are publicly sourced, such as user reviews/feedback, web forums, and stories. Because of the prevalence of irony in human speech, both genuine and sarcastic utterances are present within these datasets. However, as Medhat et al. (2014) state, both the machine learning and lexicon-based approaches struggle to distinguish between ironic and non-ironic data points, when not specifically trained to do so. Although sarcasm detection models being developed already have high performances, they lack understanding of nuances such as tonality and speed of speech, as well as the differences of these nuances across languages (NOS, 2024). By implementing a robust detection of irony, we can improve real-life applications such as sentiment analysis of reviews (Maynard & Greenwood, 2014), hate speech detection (Tiwari, 2024), and helping neuro-divergent people recognize irony (NOS, 2024).

Furthermore, English is the main language used in many studies, although some other languages such as Arabic (Rahma, Azab, & Mohammed, 2023) and Hindi (Kulkarni & Rodd., 2021) are also more commonly utilized. This field of research for Dutch data is still in its infancy. Creating an explainable model trained on Dutch data would further our understanding of the inner workings of current sarcasm detection algorithms.

Given the lack of research on Dutch data, this paper aims to answer the following questions.

RQ1 *Are the results obtained from Dutch SVM and BERTje sarcasm detection models comparable to those obtained in English?*

SVM and BERT have been proven to be effective sarcasm detection tools for the English language. However, since the nuances of irony differ across languages (NOS, 2024), it is relevant to discuss whether well-founded methodologies for the English language can be used on Dutch data.

RQ2 *Is the performance of the SVM model comparable to that of BERTje?*

With the rise of transformers and LLMs, these models are considered state-of-the-art. They recognize patterns that are not noticeable by humans. However, our human insight into the mechanisms of irony could aid in improving models. Therefore, analyzing whether these

complex models using emergent features outperform a "traditional" feature-engineering approach is of interest.

Furthermore, despite the demand for automatic sentiment classification, we need to be aware of the risks involved with the implementation and usage of these models. These algorithms can be used maliciously in politics and the commercial sphere (Matos, 2021). According to Matos (2021), as AI engineers, we cannot fully prevent ill-natured use of our algorithms. However, we can prevent internal manipulation to some extent, by understanding the flaws within the system and thus preventing bad actors from abusing them.

Traditional machine learning techniques, like linear regression, SVM, and decision trees, are easier to understand due to their limited logic (Deng, 2018) compared to neural networks. These models rely mostly on feature engineering, and the models themselves are relatively simple. The recent focus within AI research has been on complex models such as transformer models (Alqahtani, Alhenaki, & Alsheddi, 2023). However, these state-of-the-art models are opaque due to their complexity, making it difficult to understand their decision-making. More insight into the inner workings of a model should allow us to mitigate bias and potential legal and security risks, as well as simplify the prediction of how a model will perform on real-life data (*What is explainable AI? | IBM*, n.d.). To solve this, XAI (Explainable Artificial Intelligence) techniques can be used to understand the reasoning of a model.

Not only does explainability have social relevance, but it is also an underdeveloped topic within sarcasm detection research. Research focuses on making a high-performing model, and rarely implements interpretability (Johnson, Hakobyan, & Drimalla, 2023). Even when explainability is used, it is often used in error analysis. Furthermore, Johnson et al. (2023) point out that prior research hardly ever implements global explainability methods, which allow insight into the decision-making processes of the model, nor are multiple explainability methods used, which can help make the explainability more robust.

Therefore, this paper will also delve into the explainability of the two models. This leads to the following research question.

RQ3 *When analyzed using both global and local explainability methods, what are the underlying mechanisms behind the predictions by SVM and BERTje?*

Explainability is often overlooked. Most papers, like ((Maladry, Lefever, Van Hee, & Hoste, 2023a)), utilize manual error analysis. Using this method, you can analyze patterns in the data, to see whether you can find any common flaws. However, this does not

give us insight into the actual mechanisms the models used to arrive at these decisions. Inspecting these mechanisms should allow us to further our understanding of the internal flaws of the models. This way, these flaws can be improved in future iterations of irony detection models.

3 RELATED WORK

Sarcasm detection is an established field of research. This section reviews prior work in irony detection, both as a whole and in Dutch, as well as in the area of explainability. These papers will form the foundation of this thesis, even though the specific combination of using Dutch news headlines and adding explanatory methods is a topic that has not yet been explored.

3.1 *Irony Detection*

For real-life interactions, various modalities can be used to signify sarcasm. A difference in pitch can identify ironic sentiments: Dutch sarcastic utterances have a longer duration, lower intensity, and less vocal noise compared to sincere speech (Jansen & Chen, 2020). Facial expression and body language (Attardo, Eiserhold, Hay, & Poggi, 2003) can be physical cues of insincerity. Additionally, certain cues in language usage such as interjections (R. Kreuz & Caucci, 2007) and polarity between sentiments (Mladenović, Krstev, Mitrović, & Stanković, 2017) can indicate sarcasm. These combined auditory, visual, and contextual cues allow us to distinguish ironic sentiments smoothly.

In textual sarcasm detection, you are limited to only one of these modalities. However, as people still recognize written sarcasm without explicit markers (like *sarcasm*) (Ghosh & Muresan, 2018), this shows only lexical, semantic, and contextual clues allow for sarcasm detection, without having to include multiple modalities. Prior research in irony and sarcasm detection has employed a variety of methodologies to tackle this complex linguistic challenge (Alqahtani et al., 2023).

According to a review done in 2020 on sarcasm detection for Twitter data[1] (Sarsam, Al-Samarraie, Alzahrani, & Wright, 2020), SVM is one of the most popular sarcasm detection tools, with a performance between 50.93% and 91.8%. The disparity between these accuracies is quite drastic. According to Sarsam et al. (2020), this is caused by differences in feature

[1] Since July 2023, Twitter has been rebranded to X. However, for clarity and to use the same nomenclature as the literature, the terms 'Twitter' and 'tweet' will be used instead of 'X' and 'posts on X'

engineering. Therefore, when using SVM, the model needs to be trained on the right features to ensure the validity and reliability of the predictions.

However, data from social media is different from news data. A study on Dutch data (Burgers, Mulken, & Schellens, 2012) compared the lexical and semantic features of sarcasm within different genres: commercial and noncommercial advertisements, columns, cartoons, letters to the editor, book and film reviews. They analyzed what different sarcasm factors and markers occurred in each genre. Differences in intention (informative vs social) and the target audience could influence tropes, syntax, and typography marking the sarcasm, as well as the type of sarcasm used.

3.2 *Explainability*

Many AI models are inherently opaque; their decision-making cannot be easily understood. To clear up this opacity, Explainable AI (XAI) methods can be used. Within this field of research, the terms explainability and interpretability are both used as synonyms (Doran, Schulz, & Besold, 2017). These terms have slight differences in nuance, with interpretability focussing on the model itself, while explainability is more focused on the predictions. However, the goal of the field of XAI as a whole is to enable users to understand the decisions of any model.

Not every field within AI requires explainability. Whether explainability should be implemented mainly depends on one factor: the risk of *incompleteness* (Burkart & Huber, 2021). Incompleteness refers to the uncertainty of a model due to lacking knowledge. This can be due to hard-to-define issues such as ethics, unpredictability like human behavior, or problems that change over time. When a system is concerned with issues that are easy to predict, like aircraft collision avoidance systems, explainability has a lower priority.

Explainability in a system increases trust and solves issues such as fairness and accountability. XAI not only enhances these but there are also several legal frameworks in place imposing interpretability measures, such as the GDPR (European Commission, 2016) and the recent AI Act (European Commission, 2022). These European regulations are meant to protect the data, privacy, and well-being of users. The GPDR (European Commission, 2016) forces automated systems to be explicable: the algorithm needs to be transparent, and its decisions explainable to those affected both directly and indirectly. The AI Act of 2021 (European Commission, 2022) specified at-risk sectors, such as transport, education, employment, migration, justice, and health care. These high-risk sectors have to adhere to strict standards of safety, data management, and interpretability. All other sections, among which is also the automatic analysis of text, should

follow transparency obligations to foster trust between algorithms and their users. Therefore, in light of these regulations, explainability is essential in most applications.

When applying XAI to your AI model, you can use two different approaches (Zhang, Tiňo, Leonardis, & Tang, 2021): either *local* or *global* explainability. With local explainability, you explain the decision-making of one specific predicted value. With global explainability, you want to shed light on the entire decision-making process of the model. Some methods are also considered *semi-local*; they provide explanations for groups of instances, rather than individual instances.

Some explainability methods are model-specific, meaning they can only be used for a specific type of model. An example of this is Layer Integrated Gradients (Sundararajan, Taly, & Yan, 2017), which can be used for deep learning models with multiple layers. When comparing the explanations behind multiple models, a model-specific method might not work, as the models you are comparing might have drastically different architectures. In this case, you might want to use a model-agnostic method, such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), Partial Dependence Plot (PDP), or Accumulated Local Effects (ALE) plots.

3.2.1 *Local Explainability*

LIME (Local Interpretable Model-agnostic Explanations), originally proposed in Ribeiro, Singh, and Guestrin (2016) shows the feature importance for the classification of a single instance. LIME can be used on many kinds of data, such as tables, imagery, and text. Another commonly used local explainability method is SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), which works similarly to LIME as it also calculates the feature importances.

One big advantage SHAP has over LIME is that its outputs are more stable (Kalai & Samet, 1987). LIME is more sensitive to small changes in the data, and is not as robust to noise. There can also be more than one possible output for LIME, whereas SHAP only provides one explanation for every instance. Because of this, SHAP is considered the gold standard for local explainability within NLP (Mosca, Szigeti, Tragianni, Gallagher, & Groh, 2022).

Using analysis through SHAP regression values, the contribution each feature to the predictions made by a machine learning predictor model can be approached ((Silva, Keller, & Hardin, 2022)). This is calculated by creating an "explanation model", which evaluates the prediction of the model as the sum of the contributions of each input feature and the mean predicted value (Molnar, 2022).

The prediction value, also known as the "SHAP regression value" or "SHAP value", is based on game theory (Molnar, 2022). These values are computationally expensive to calculate. However, the SHAP library in Python allows for swifter versions of SHAP (Mosca et al., 2022), meaning it is computationally feasible to use on models with a high-dimensional feature set as well as complex neural networks.

SHAP can be used to inspect individual instances and groups of instances within your dataset, making it suitable for both local and semi-local explainability. But, it can also be used to inspect the model globally (Chromik, 2021). You can do this using a beeswarm plot (Lundberg & Lee, 2017), which shows how the top features in a dataset impact the model's output. SHAP can be used for both feature-based models as well as neural networks such as BERTje, although is met with some issues (Kokalj, Škrlj, Lavrač, Pollak, & Robnik-Šikonja, 2021). BERTje utilizes sequential information to classify input data, whilst SHAP calculates the predictions for individual values separately and sums these values, disregarding context. Kokalj et al. (2021) proposes a solution for this issue: TransSHAP, a new way of visualizing SHAP values, specifically targeted toward transformer models. In the Kokalj et al. (2021) paper, BERT was used to test the TransSHAP method, demonstrating it works well for this model.

3.2.2 *Global Explainability*

Aside from SHAP, there are many other global explainability methods. A well-known example is PDP (Partial Dependence Plot), which is a low-dimensional plot of a model that portrays the estimated relationship between the output and different features (Greenwell, Boehmke, & McCarthy, 2018). Because of their low-dimensionality, interpreting them is straightforward.

As Molnar (2022) also points out, the partial dependence function not only plots the effects of the feature itself but also its interactions with other features, it cannot be used on dependent data. This independence assumption is a significant limiting factor, as features of real-life data are often interconnected. A method that deals with this is ALE (Accumulated Local Effects) plots (Apley & Zhu, 2020), which also describe the influence of the features of the output, but without the independence assumption.

The ALE method calculates the prediction differences when the feature of interest is replaced with grid values in a certain instance (Molnar, 2022). The effect of changing the features with other variables is measured and averaged. This is done for all intervals, and the effects are accumulated. Because you locally measure the effects and later accumulate them, you can

be sure you are measuring the feature of interest, and not the correlation with other variables.

When plotting the global feature importances with SHAP, it simply calculates the feature importances for every individual feature, and adds these together into one plot (Lundberg & Lee, 2017). Each instance is represented by a dot. Therefore, the most important difference between ALE and SHAP is that ALE calculates feature importances over a group of instances and aggregates these, while SHAP calculates this for individual instances before aggregating. These methods can result in slightly differing explanations (Liang, Cai, & Su, 2022). Therefore, both techniques will be utilized for robustness.

Both PDP and ALE rely on pre-existing features within the dataset. However, when a model uses built-in feature extraction like BERTje, these methods will not be feasible. A feature-less model requires a model-specific explainabilitity solution like Layer-Integrated Gradients (Maladry et al., 2023a; Nayak & Timmapathini, 2021; Sundararajan et al., 2017) This method lowers feature values until it finds the threshold that arouses important gradient changes. Because it relies on gradient changes, it cannot be used on models that are not neural networks.

3.3 *Dutch Sarcasm detection*

Like many low-resource languages, not much research has been done in the field of Dutch sarcasm detection. Nonetheless, it is not left completely undiscovered. In an often-cited study from 2013 (Liebrecht, Kunneman, & van den Bosch, 2013) researchers at Radboud University trained a Linguistic Classification System on 3.3 million Dutch tweets, of which only 135 were sarcastic. Their algorithm correctly classified 101 out of 135 tweets. But, when training their model on the 250 tweets it had classified as most likely to be sarcastic, it only achieved 30% accuracy. Liebrecht et al. (2013) concludes their model could not distinguish between sarcastic tweets and similar (but non-sarcastic) tweets, making it ineffectual in real-life sarcasm detection.

One research group from Ghent University has written multiple papers on Dutch sarcasm detection. In a recent study (Maladry, Lefever, Van Hee, & Hoste, 2022), they used an SVM classifier on 5,566 annotated Dutch tweets, and compared in to the Dutch version of BERT (BERTje) (de Vries et al., 2019). They trained multiple SVC models: a baseline model and 4 SVC models with different clash features. These SVC models were compared to a baseline BERTje model, which was trained to distinguish between ironic and genuine tweets. Notably, all SVC models, even the baseline model, achieved higher accuracies than the BERTje baseline.

This same group of researchers (Van Hee, De Clercq, & Hoste, 2021) also applied the same methods to classify the sentiment of news texts. This paper focuses on the comparison between different lexicon-based models to traditional machine-learning models, based on how similar their classification is to manual annotation. For the lexicon-based classifiers, two different methods were used. For one, they classified the sentiment of each word in the text based on different sentiment lexicons. This method was compared to a pre-trained model called SentiNET (Chou, Tramèr, Pellegrino, & Boneh, 2018). These sentiment classifiers were compared to three different machine learning models: SVM, as well as two Dutch BERT classifiers, BERTje (de Vries et al., 2019) and RobBERT (Delobelle, Winters, & Berendt, 2020). All the machine learning models significantly outperformed the sentiment-based classifiers. However, unlike in the 2022 study, BERTje and RoBERT outperform the SVM model in the weighted average F1-score. This difference is interesting: it either points toward the newer methodology of manually selecting features being better, or an inherent difference in news-based data compared to social media data.

Continuing from their 2022 study, this same group of researchers delves into the interpretability of their transformer model (Maladry et al., 2023a). This was done by tweaking certain features and words, as well as using Layer Integrated Gradients, Discretized Integrated Gradients, and Layer-wise Relevance Propagation, with the ultimate goal of detecting bias. They note that strongly expressed sentiments are more likely to be classified as ironic, as are all positive sentiments. These sentiments are therefore most at risk of being wrongly classified as ironic. As this research is done on Dutch tweets, it might be interesting to see if similar biases are detected when training on news headers, as well as comparing them to biases in an SVC model.

4 METHOD

4.1 *Data*

The data used in this research project is a public dataset taken from Kaggle (Harrotuin, n.d.). It consists of 13262 Dutch news headlines from 11/10/2007 - 12/05/2020. Of these headlines, 5001 (37.7%) are sarcastic and 8261 (62.3%) are non-sarcastic entries. The non-sarcastic headlines are taken from *nu.nl* ((*NU.nl*, n.d.)), which is a renowned news site. The sarcastic articles are created by *De Speld* (*speld.nl*, n.d.), which is the Dutch equivalent of the fake-news site The Onion.

This dataset does not distinguish between verbal irony (a clash between literal and hidden meaning) and situational irony (irony through a mis-

match in expected reality compared to actual outcome). Therefore, this distinction will also not be made during the classification.

Aside from the headline and its corresponding binary rating of sarcasm, the dataset also includes the link to the article, the source (either speld.nl or nu.nl), as well as three boolean variables denoting the subject of the news piece: *is_binnenland* (domestic Dutch news), *is_buitenland* (foreign news), *is_politiek* (political news). There is a correlation of -0.64 between *is_binnenland* and *is_buitenland*, a correlation of -0.47 between *is_binnenland* and *is_poltiek*, and a correlation of -0.37 between *is_poltiek* and *is_buitenland*. This indicates a small chance of any of these subjects occurring for the same headline, but there are some co-occurrences. Of the headlines, 45.2% covered domestic news, 33.4% was foreign news, and 21.5% was political; all headlines have at least 1 subject. There is no strong correlation between one of the three subjects, and whether the headline is sarcastic: a correlation of 0.2 for domestic news, -0.15 for foreign news, and -0.07 for political news. So, there should be no bias toward a single subject. As the subjects do not help predict whether a headline is sarcastic, these will be excluded from the models. However, the subjects may help us find out whether either of the models is biased toward a certain subject matter.

4.2 *SVM*

To train an Support-Vector Classification (SVC) algorithm, you have to extract features. The preprocessing and feature extraction is largely based on the prior research in Van Hee (2017), Van Hee et al. (2021), and Maladry et al. (2022).

To ensure validity, training was done across multiple seeds (Qian et al., 2021). This ensures a higher performance is not due to the random seed, to improve the reproducibility of the study (Baker, 2016).

4.2.1 *Preprocessing*

Since the training and cross-validation were done over multiple different seeds, the pre-processing and feature engineering were applied prior to the data splitting. This is usually considered bad practice, as data from the training set can leak into the test set, and therefore artificially inflate the results. However, in this specific case, the features of each headline are calculated completely separately from each other. Therefore, data leakage should be no issue.

In a study on the implicit sentiment portrayed in news lines (Van Hee et al., 2021) an SVM and BERTje model were trained, and compared to different lexicon-based approaches. To train the SVC, they used only n-

gram features, with automated feature extraction. However, as they note in their discussion, the SVM model could be included by extended feature extraction.

Therefore, the preprocessing will be largely taken from their research on Twitter data. Both Maladry et al. (2022) and Van Hee (2017) use the same feature extraction techniques. Preprocessing consists of tokenization, Part-of-Speech tagging, lemmatization, and named entity recognition (NER), as well as additional cleaning of the data. Using this preprocessed data, several lexical, syntactic, and semantic features were extracted.

A key difference between social media data and news headline data is that tweets are user-generated. This means a substantial number of features that mark sarcasm in tweets, especially lexical features, do not translate well to news headlines. These features include unusual spelling and punctuation usage, such as character or punctuation flooding and capitalization, as well as tweet-specific features, such as hashtags and emoticons. Therefore, the methods used for Twitter data cannot be directly duplicated for classifying news headlines.

The same preprocessing methods as described in Van Hee (2017) were utilized, except for cleaning tweet-related characters such as hashtags and emoticons. For the preprocessing, the Dutch spaCy library (Honnibal & Montani, 2017) was applied, which is trained on a lexicon containing news data.

For the syntactic feature extraction, the method of Van Hee (2017) was followed, extracting several PoS and NER features. The 2017 study also utilizes temporal clash, using the LeTs Preprocess part-of-speech tagger (Van de Kauter, Marjan and Coorman, Geert and Lefever, Els and Desmet, Bart and Macken, Lieve and Hoste, Veronique, 2013). However, due to lack of public access to this tagger, as well as this method not being used in other SVC models for irony, this particular feature was omitted from the preprocessing pipeline.

For the semantic features, a Word2Vec model based on an existing set of Dutch embeddings (Tulkens, Emmery, & Daelemans, 2016) was used. The embeddings used were based on a combination of Wikipedia, Roularta, and Sonar500 data. The 320-dimension version was chosen, as these outperform the 160-dimensional embeddings according to the authors. Using these embeddings, a vector was created of the tokens in each headline; tokens not included in the embeddings were left out. The word vectors of each headline were averaged, and this was used as the semantic feature.

Van Hee et al. (2021) compared two different methods for classifying sentiment in Twitter data: SentiNET, and a combination of 4 different sentiment lexicons. This latter method outperformed the prior quite significantly. However, one of the lexicons cited was an "in-house" lexicon that

was not publicly available. Therefore, the approach used for the sentiment feature consisted of combining the remaining lexicons: Pattern (Smedt & Daelemans, 2012), DuOMAn (Jijkoun & Hofmann, 2009), and NRC Emotion Lexicon (Mohammad & Turney, 2013). The sentiment is determined by these lexicons consecutively; if Pattern cannot classify the sentiment, then look at DuOMAn, finally at NRC. If none of these lexicons can classify a word, it is marked as 'neutral'. From this, the average sentiment of each headline was determined, along with three other sentiment features marking the changes in sentiment within the sentence.

After this preprocessing, the final feature set consisted of 79 features. A list of these features can be found in Appendix A (page 33).

4.2.2 *Training*

After preprocessing, the data was first divided into a feature set and their corresponding labels. These features were scaled using the StandardScalar from sklearn (Pedregosa et al., 2011). This scaled data was split into a training, validation, and test set with a 80-10-10 split, using stratified sampling to ensure an even distribution of sarcastic and non-sarcastic data in each set.

Firstly, the hyperparameters need to be defined. To find these, 3-fold cross-validation was used to find the best C, gamma, and kernel. The values included in this cross-validation were partially based on the best values from Van Hee 2021 and Maladry 2022. The hyperparameters with the highest accuracy were selected for the training of the final model.

For training the SVM model, the LibSVM library was used (Chang & Lin, 2011). This training was done over 10 different seeds. These seeds all used the best hyperparameters[2]. The performance (accuracy, precision, recall, and F1-score) was evaluated on the test set for each seed and averaged. The best-performing model on the test set was saved using pickle, so it can be analyzed later using local and global explainability methods.

4.3 *BERTje*

The methods for BERTje were strongly based on the methods of Van Hee et al. (2021) and Maladry et al. (2022). These two papers both use the same preprocessing, fine-tuning, and training methodologies, to classify fine-grained news events, as well as Twitter data.

[2] C: 5, γ : 0.001, *kernel* : *RBF*

4.3.1 *Preprocessing*

The data was split before preprocessing into a training (80%), validation (10%), and test set (10%). Then, each set was tokenized with a pre-trained Dutch tokenizer, padded, and turned into a tensor. These preprocessed sets of tensors are used to train the BERTje model.

4.3.2 *Training*

This preprocessed dataset was used to train a BERTje model. The model utilized is from the official documentation (de Vries et al., 2019) and has already been fine-tuned to be used in Dutch sequence classification. Like other transformer models, BERTje can be used for both classification and extracting embeddings. For classification tasks, the model uses its final classification layer. By removing this final layer, BERTje can also provide word vectors, making it versatile for various NLP tasks. For this specific application, all layers are kept, so it is suitable for classification.

The model was first trained on the training set. The hyperparameters[3] were previously proven to be effective in sarcasm classification tasks ((Maladry et al., 2022; Van Hee et al., 2021). For each epoch, the performance was tested on the validation set. If the accuracy of the model on the validation set exceeds the previous validation performance, this model is saved. By selecting the best model on the validation set, rather than the training accuracy, it is assured the model is not overfit.

Finally, the best BERTje model is evaluated on the test set, using accuracy, precision, recall and F1-score.

To understand the difference in performance between the SVM and BERTje models, the instances were identified wherein the classification between the models differed. For a list of these differently classified headlines, see Appendix B (page 34) and Appendix C (page 41).

4.4 *SHAP, ALE and LIG*

Previous research ((Maladry et al., 2023a)) used model-specific explainability methods like Layer-Integrated Gradients (LIG) to interpret their BERTje results. However, since SVM is not a deep-learning model, these same methods cannot be applied. Therefore, the mechanisms behind SVM were investigated using SHAP and ALE.

SHAP is used to analyze both the SVM and BERTje models. For each model, different functions were used from the SHAP library (Mosca et al., 2022). KernelSHAP, a completely model-agnostic method, was applied to SVM.

[3] optimization: Adam; learning rate: 5e05, batch size: 64; number of epochs: 3.

First, the SHAP explanations for SVM were created, using both these differently classified instances, as well as the correctly identified instances. Aside from the incorrect and correct classifications, the entire test set was also used to create SHAP explanations. From these three types of explanations, beeswarm plots were created to visualize the SHAP values.

To do this, a SHAP explainer was created for the SVM model trained on the headline data. This SHAP explanation model was used to calculate SHAP values for 100 randomly sampled values from each of the three datasets. By only using a smaller subset of instances, the computational complexity of the SHAP value calculation was limited, while still being able to find the emergent patterns from the data, as is discussed in the results.

For the SHAP analysis of BERTje, the TransSHAP method was used, as per the methods of Kokalj et al. (2021). This is a SHAP explanation model and visualizer specifically for the transformers. Using their method, 10 correctly and 10 incorrectly identified instances were analyzed. Unlike with the SHAP plot, these visualizations do not show the importance of features, but of individual words.

ALE (Apley & Zhu, 2020) creates a plot for every feature of the SVM feature set, showing its importance to the overall output of the model. Using this method, the most important features according to the SVM SHAP beeswarm plots can be analyzed more in-depth. On top of the ALE plots of single features, 2D ALE plots show the interactions between two features. This will allow to expose more complex patterns within the model.

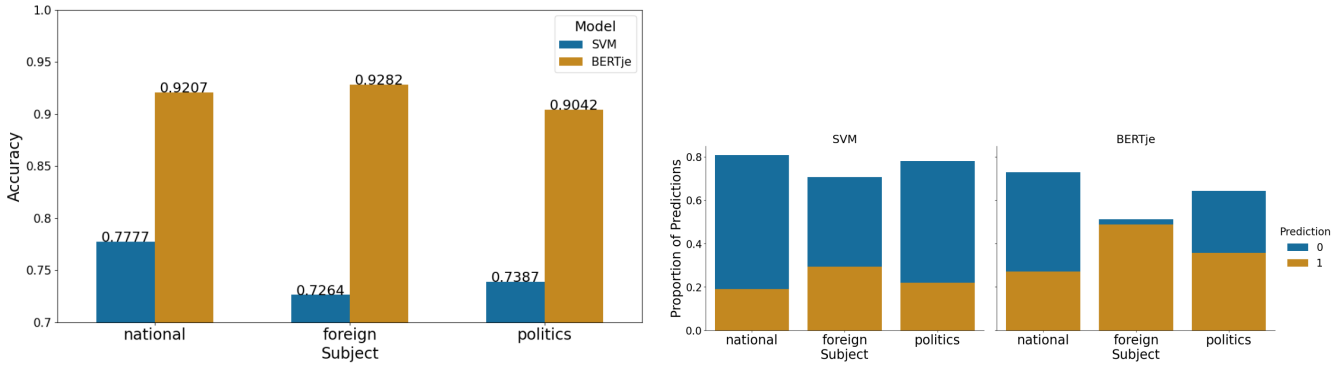
Since the SVM model was trained on normalized values, these were used to create the ALE plot. However, to make the plots easier to understand, the corresponding non-normalized values were used on the axis of the visualization, following the methods of the Jomar (2023) documentation.

Since BERTje is not a feature-based model, ALE cannot be applied. Therefore, one of the explainability methods proven effective in Maladry et al. (2023a) was to understand this model better globally: Layer-Integrated Gradients (LIG). A LIG explainer was trained on the entire BERTje model using the transformers-interpret library (Pierse, 2024). Then, 20 correctly and incorrectly classified headlines were analyzed; these were randomly sampled and consisted of the same amount of sarcastic and non-sarcastic instances. For these sentences, the LIG model returned the word importance for the classification, which can be used to make inferences about the model.

5 RESULTS

The results (Table 1) reveal BERTje outperforms SVM in all evaluation metrics. The most notable difference is the recall: BERTje has a recall of 0.91, while SVM has a recall of 0.79. This points towards the SVM model being biased toward the majority class of non-sarcastic utterances.

Before applying SHAP and ALE, some simple analysis was done by plotting the predictions per subject. The subject of each news header was left out of the training datasets. As mentioned previously, none of the subjects had a significant correlation with sarcasm. However, the subjects were not equally represented in the data: Dutch domestic news was almost half of the data, whereas only about 20% of the headlines covered politics. Therefore, the majority subject matter might be more easily classified than the minority class.



(a) Accuracies of SVM and BERTje per Subject

(b) Classifications of SVM and BERTje per Subject

Figure 1: Bar plots of the accuracy (a) and classifications (b) per subject for SVM and BERTje

From Figure 1a, it is clear that BERTje outperforms SVM in every subject. The performance of each category for SVM seems to reflect the amount of data for each subject, *national* has the highest performance and amount of data, followed by *foreign*, with *politics* having both the worst performance and smallest amount of data. Politics was also the subject with the lowest performance for BERTje. However, the classification of

Table 1: Scores obtained by the machine learning approach on the held-out test set

Model	Test Accuracy	Test Precision	Test Recall	Test F1 (macro)
BERTje	**0.9205**	**0.8849**	**0.9070**	**0.9157**
SVM	0.8726	0.8576	0.7942	0.8623

foreign headlines was better than national headlines, despite not being the majority class.

The classifications of each model per class were also plotted (see Figure 1a). The headlines are more often classified as non-sarcastic, as the majority of the data is non-sarcastic. Of the headlines concerning national news in the dataset, 48.3% was sarcastic. For foreign news, this was only 27.5%, and the political headlines were 31.2% sarcastic. Of the proportions in Figure 1b, the number of sarcastic classifications of political headlines for both models came the closest to the real proportion. The amount of national headlines classified as sarcastic was too low, whereas the sarcastic classifications for foreign headlines were too high. It is interesting that politics had the lowest accuracy for both models, but the proportion of sarcastic to non-sarcastic classifications seems to be the best representation of the original data. These results were further investigated with the other explainability tools.

5.1 *Explainability of SVM*

SHAP was used to create feature importance beeswarm plots for the correctly classified classes (Figure 2), the incorrectly classified classes (Figure 3), as well as all predictions for SVM. The plots show the 20 features that influence the model decision most on a local level, according to their mean absolute SHAP value. Each dot corresponds to one news headline. The beeswarm plots show how the 20 most important features influence the classification of each news headline. Positive SHAP values indicate the headline is more likely to be ironic.

The beeswarm plot based on the correctly classified data points (Figure 2) shows clear clusters of data points. This means that there is a relatively clear correlation between the values of certain feature values and their influence on the model. This is demonstrated by the most important feature: the absolute count of nouns, which are scored 0, 1, 2, or >2. The SHAP value is high when this feature is a low value, indicating a lower amount of nouns is important in classifying a sentence as sarcastic.

For the incorrectly classified headlines (3), there are no clear clusters. This means that you cannot draw clear conclusions about the value of the feature, and the importance it plays in the classification of sarcasm globally.

A similar set of features is the most important for both the correctly and incorrectly classified values, as well as the model overall (Figure 4). This means that the model tends to rely on certain features, despite them not always being predictors of sarcasm.

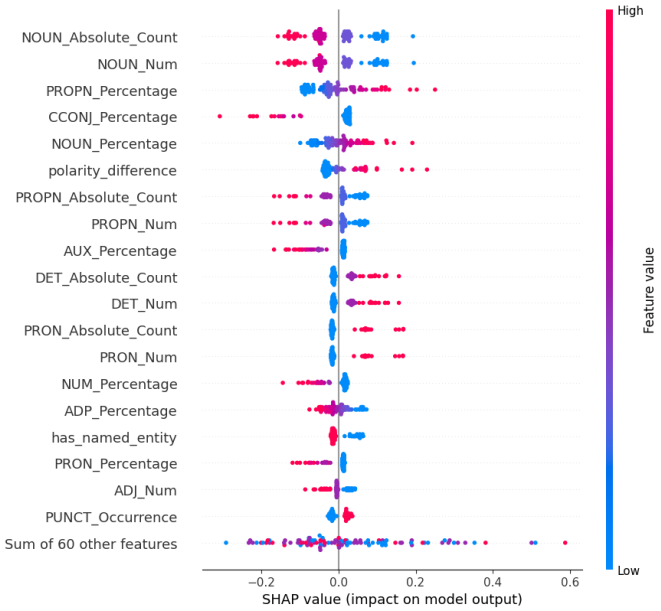


Figure 2: SHAP beeswarm of correctly classified instances (SVM)

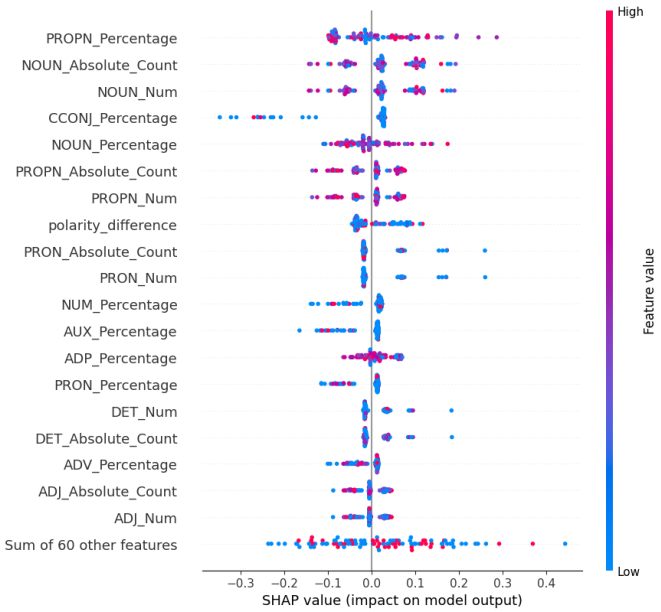


Figure 3: SHAP beeswarm of incorrectly classified instances (SVM)

The most important types of features were: the amount of nouns, proper nouns (which are names of people, things, or places), and coordinating conjunctions (words connecting larger sentences, like *and*, *or*, and *but*), as well as the polarity differences.

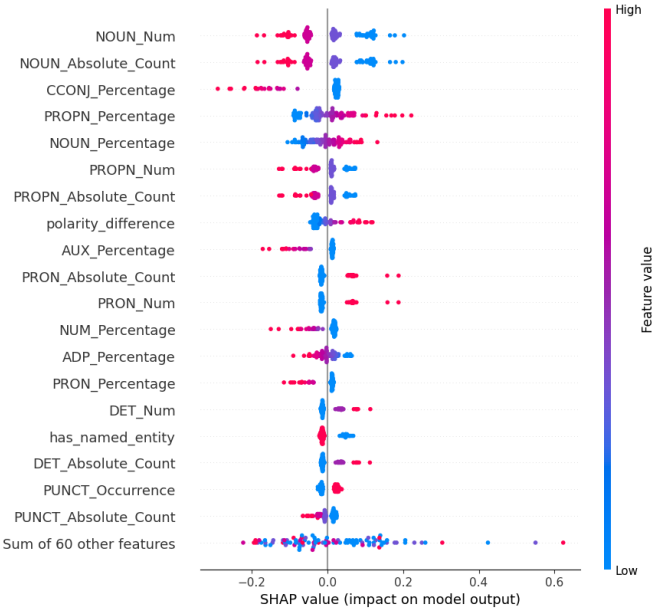
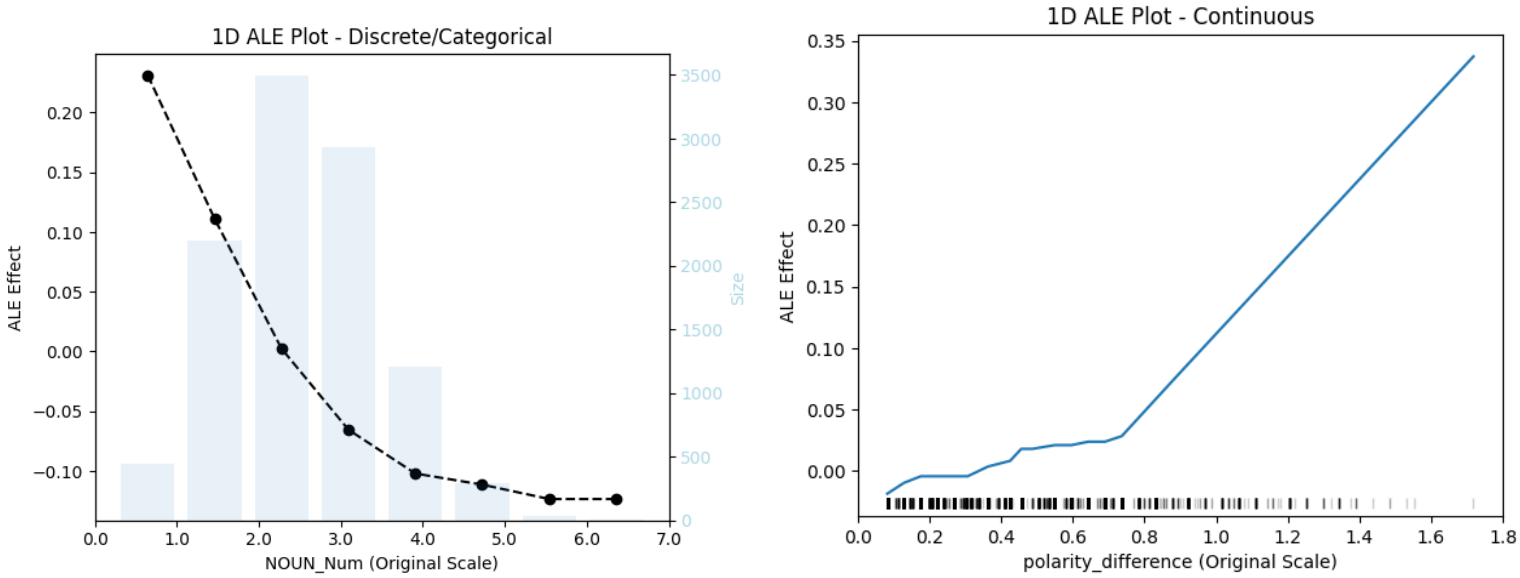


Figure 4: SHAP beeswarm of all instances (SVM)

To reiterate, SHAP is not a global method: each dot represents a single instance. To analyze the impact individual features or combinations of two features have on the SVM model globally, we can use ALE. The same top twenty features generated by the SHAP model were analyzed, both individually and paired. Two of the highest features according to the SHAP plots, *NOUN_Num* and *polarity_difference*, will be further analyzed. These features were selected because they represent two different types of features (syntactic and semantic), and because they tend to skew sarcasm models toward incorrect decisions (Maladry, Lefever, Van Hee, & Hoste, 2023b).

There are two different types of 1D ALE plots: discrete and continuous (Apley & Zhu, 2020). Discrete ALE plots consist of bars (representing the number of samples within that bin), and a line showing the ALE values of each bin. Continuous ALE plots are a line plot, with the line representing the ALE values, as well as showing the distribution of the samples using black markings right above the x-axis. Higher ALE values (y-axis), either positive or negative, show the feature has a larger feature importance compared to lower ALE values.

The discrete 1D plot of the number of nouns (Figure 5a) portrays that a low number of nouns tends to be positively correlated to sarcasm within the SVM model. The continuous ALE plot of polarity difference (Figure 5b), the higher the polarity difference, the more likely it is to be classified as sarcastic. Both of these figures are in line with the SHAP plots.



(a) Amount of Nouns

(b) Sentiment changes

Figure 5: One-Dimensional ALE Plots

2D ALE plots are heat maps, with colors representing the effect of the combined features on the prediction. As you can see within Figure 6, the combination of a large number of nouns and a high polarity difference has a negative effect on the prediction. On its own, high polarity differences have a positive effect on sarcasm prediction. However, when combined with a large amount of nouns, this is actually the opposite: when both features have high values, the headline is more likely to be classified as non-ironic. This dichotomy highlights the importance of model interpretability with varied methodologies.

5.2 *Explainability of BERTje*

Ten of the correctly and incorrectly classified instances from BERTje were also analyzed using SHAP. An example of both a correct and incorrect explanation can be found in Figure 7.

In these explanations, the impact on the model (y-axis) is shown per word. Green (up) means it is correlated to sarcasm, whereas red (down)

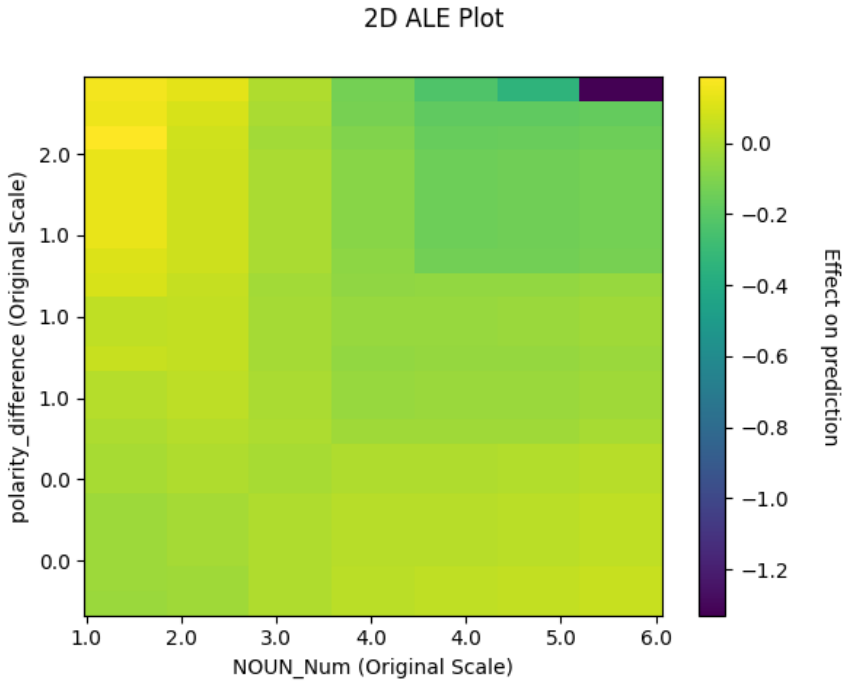


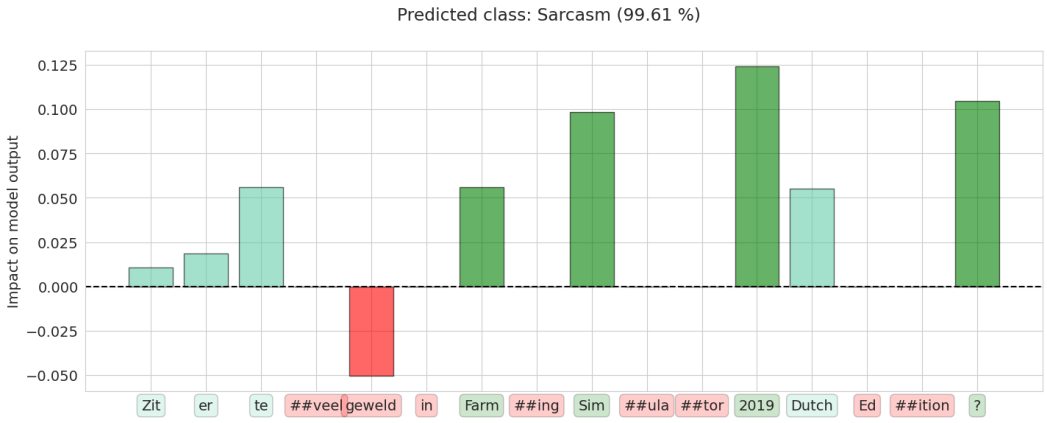
Figure 6: Two-Dimensional ALE Plot of the Number of Nouns and Polarity

is correlated to genuine statements. The height of the bars portrays the magnitude of the impact. In the correctly classified example (Figure 7a), the tokens "2019", "?", "Farm", and "Sim" indicate the sentence is sarcastic, whereas "violence" was characterized as non-sarcastic. In the incorrectly identified example (Figure 7b), the first part of "majority" as well as "smoking ban" hinted toward it being a geniune statement, while the rest of the utterance, especially the second part of "majority" and "the Dutch" led the model to classify it as sarcastic.

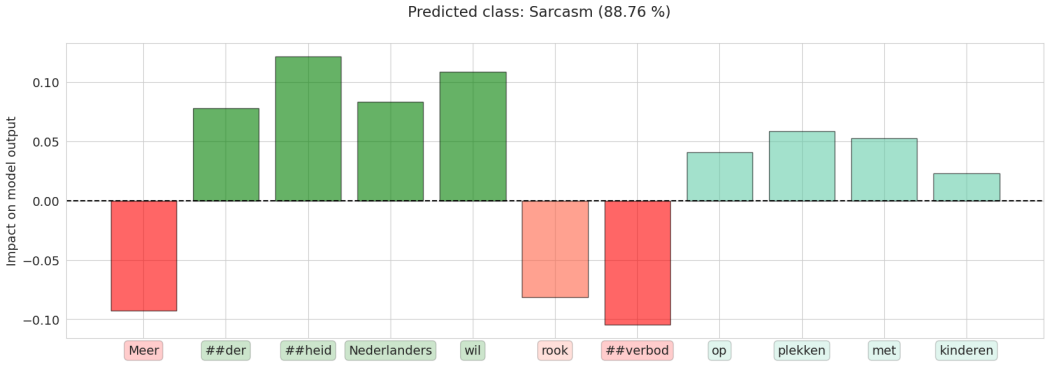
In the rest of the SHAP explanations for BERTje, a similar pattern seems to emerge. Named entities liked to politics and governance (such as "Trump", "Lincoln", "alderman", "ban", "scandal", and "prevention") seem to be classified as non-sarcastic. Words like "majority" and country names ("the Dutch" and "Belgium") also fit with this theme of bureaucracy.

More everyday words, such as "which", "with", and "due to", as well as years and punctuation like question marks and quotation marks are indicators of sarcasm. Interestingly, as opposed to most punctuation marks, the comma was strongly correlated to genuine news headlines. This checks out: commas tend to occur more in longer sentences, which are more often non-ironic according to the SVM SHAP plots.

In addition to SHAP, a Layer-Integrated Gradient interpretation created the BERTje model. This LIG explainer was used to analyze 10 sarcastic,



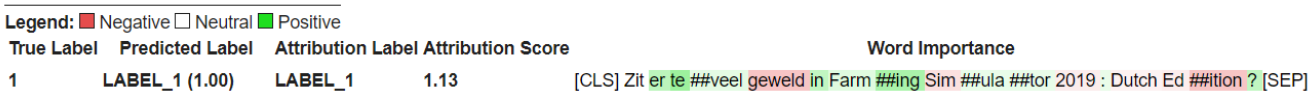
(a) Correct Classification: "Is there too much violence in Farming Simulator 2019 Dutch Edition?"



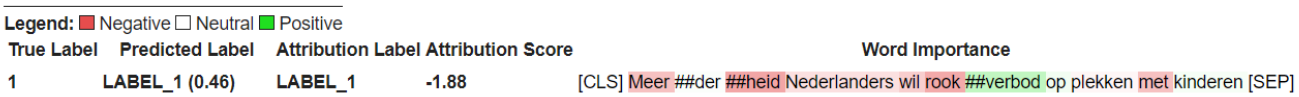
(b) Incorrect Classification: "Majority of the Dutch want a smoking ban in places with children."

Figure 7: SHAP Explanations of a) a sarcastic instance incorrectly classified as sarcastic and b) a non-sarcastic instance incorrectly classified as sarcastic

and 10 non-sarcastic headlines. One example of these explanations can be seen in Figure 8. This figure uses the same examples as Figure 7.



(a) Correct Classification: "Is there too much violence in Farming Simulator 2019 Dutch Edition?"



(b) Incorrect Classification: "Majority of the Dutch want a smoking ban in places with children."

Figure 8: LIG Explanations of a) a sarcastic instance incorrectly classified as sarcastic and b) a non-sarcastic instance incorrectly classified as sarcastic

Green (positive) are all the sarcastic words, whereas red (negative) are all the genuine tokens. Similarly to the SHAP explanations, in the correctly classified LIG example (Figure 8a), the tokens "Farming", and "?" indicate the sentence is sarcastic. However, according to this model, not only "violence", but also "Sim", "2019" and "Dutch Edition" are indicators of sincere statements.

In the incorrectly identified example (Figure 8b), only "ban" indicates sarcasm. The other words in the sentence are either neutral, or lean toward non-sarcasm. However, because "ban" has a relatively high magnitude, the sentence was still classified as sarcastic.

Overall, similar patterns to SHAP emerge. Names of politicians and political parties, companies, as well as words relating to violence ("war", "nuclear", "police", "disaster") are indicators of sarcasm. All quotation marks seem to be somewhat correlated to irony; even a comma, depending on the context. There do not seem to be any major discrepancies between the SHAP and LIG explanations.

6 DISCUSSION

According to a review done in 2020 on sarcasm detection for Twitter data (Sarsam et al., 2020), SVM is one of the most popular sarcasm detection tools, with a performance between 50.93% and 91.8%. This difference in accuracy is due to different feature engineering techniques. After 2020, SVM was still a popular choice for sarcasm detection, but newer models like LSTM and BERT quickly garnered more traction as well (Alqahtani et al., 2023). More than half of the reviewed articles used a deep learning

classification method rather than a traditional machine learning approach. Deep learning is becoming more popular in NLP as a whole; not just sarcasm detection. These complex neural networks have outperformed SVM in most recent studies for the English language. The BERTje model trained on the Dutch news data set outperformed the SVM model. This is in line with the body of research done on English data.

On the contrary, in the study Maladry et al. (2022), comparing the performance of a few different SVM models to BERTje for prediction sarcasm in Dutch tweets, the SVM models outperformed BERTje. Although the data used in this study was different, the methods followed were largely the same. However, in the results of this thesis, these results did not translate: BERTje scored better SVM in every performance measure. This is in line with the studies on English data (Alqahtani et al., 2023), but not the Maladry et al. (2022) study of Dutch data.

The differences between our results and those of Maladry et al. (2022) might be due to the differences in preprocessing, or due to either the inherent characteristics of the data, like the absence of easily recognizable sarcasm markers such as character flooding and emoticon usage. The differences could possibly also be due to different amounts of verbal irony compared to situational irony, as well as tweets being fully textual, while headlines are often multi-model, consisting of both text and an accompanying image. When regarding the incorrectly classified headlines in Appendix B on page 7 and Appendix C on page 7, it is hard to tell which is sarcastic and which is not without the context of an image, and contextual knowledge of the situation written about. For future research, it might be interesting to investigate the differences in sarcasm markers between headlines and tweets, akin to Burgers et al. (2012). Alternatively, letting human participants manually classify the headline data might help in understanding which headlines are clearly sarcastic, and which are harder to classify. Analysis of these hard-to-classify instances might give us pointers on how to improve the classification.

Generally, sarcasm detection has a few major obstacles to overcome still (Maladry et al., 2023b). One of these is that sarcasm detection often relies on cues of intense sentiment, which means that any strongly expressed utterance will get classified as sarcastic, while it might be a genuine, non-ironic statement. This does not seem to be the case with either of the models, as neither SVM nor BERTje is skewed toward high polarity. This could be due to a different kind of irony between the data used in this thesis, and the twitter data used by Maladry et al. (2023b). As discussed, irony can be verbal, characterized by hyperbole, or situational, based on the clash between expected and real outcomes (R. J. Kreuz & Roberts, 1995; Lucariello, 1994). The latter requires an understanding of the real

world, which is notably lacking in sarcasm models according to Maladry et al. (2023b). They propose a solution of a common-sense unit, based on syntactic patterns. This will allow a model to recognize irregular syntactic patterns, which are more likely to be ironic.

The models trained on news headers do not show a bias toward sentiment clashes. However, they did perform worse on political headlines, compared to the others. From the model introspection, it did not quite become clear what the cause of this performance discrepancy was. It would be interesting to explore whether integrating a common-sense unit could enhance the models' understanding of syntax. By improving situational irony recognition, this might make the model more generalizable across different subjects.

7 CONCLUSION

This paper explored sarcasm detection for Dutch news headlines with two different models: a feature-based SVM model, and the state-of-the-art transformer model BERTje. The BERTje model outperformed SVM in every metric. This is in line with the recent work on English data (Alqahtani et al., 2023). However, it does not fit the studies done on Dutch twitter data (Maladry et al., 2022; Van Hee et al., 2021), wherein the SVM models performed better than BERTje and RobBERT.

Furthermore, these two models were analyzed by a local explainability model, as well as two different global interpretability methods, in an attempt to expose the mechanisms behind these models. SVM was inspected using SHAP (local) and ALE plots (global). These had some conflicting results, leading to no clear definitive answer as to which features indicated sarcasm. BERTje, on the other hand, did not encounter such conflict. Both SHAP (local) and LIG (global) returned similar results, with named entities, as well as terms related to politics and violence, tend to be correlated to genuine statements. All miscellaneous or common words tend to be classified as sarcasm by BERTje.

REFERENCES

Alqahtani, A., Alhenaki, L., & Alsheddi, A. (2023, 2023). Text-based sarcasm detection on social networks: A systematic review. *International Journal of Advanced Computer Science and Applications*, *14*(3). Retrieved from https://tilburguniversity.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/text-based-sarcasm-detection-on-social-networks/docview/2807222534/se-2 (Copyright - © 2023. This work is licensed under

http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-11-27)

Amante, D. J. (1981). The Theory of Ironic Speech Acts. *Poetics Today*, *2*(2), 77–96. Retrieved 2024-05-02, from https://www.jstor.org/stable/1772191 (Publisher: [Duke University Press, Porter Institute for Poetics and Semiotics]) doi: 10.2307/1772191

Apley, D. W., & Zhu, J. (2020, June). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086. Retrieved from https://doi.org/10.1111/rssb.12377 (_eprint: https://academic.oup.com/jrsssb/article-pdf/82/4/1059/49323845/jrsssb_82_4_1059.pdf) doi: 10.1111/rssb.12377

Attardo, S., Eiserhold, J., Hay, J., & Poggi, I. (2003, June). Multimodal markers of irony and sarcasm. , *16*(2), 243–260. Retrieved 2024-05-07, from https://www.degruyter.com/document/doi/10.1515/humr.2003.012/html (Publisher: De Gruyter Mouton Section: HUMOR) doi: 10.1515/humr.2003.012

Baker, M. (2016, May). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. Retrieved from https://doi.org/10.1038/533452a doi: 10.1038/533452a

Bowes, A., & Katz, A. (2011, April). When Sarcasm Stings. *Discourse Processes*, *48*(4), 215–236. Retrieved 2024-04-28, from https://doi.org/10.1080/0163853X.2010.532757 (Publisher: Routledge _eprint: https://doi.org/10.1080/0163853X.2010.532757) doi: 10.1080/0163853X.2010.532757

Burgers, C., Mulken, M. v., & Schellens, P. J. (2012). Verbal Irony: Differences in Usage Across Written Genres. *Journal of Language and Social Psychology*, *31*(3), 290–310. Retrieved from https://doi.org/10.1177/0261927X12446596 (_eprint: https://doi.org/10.1177/0261927X12446596) doi: 10.1177/0261927X12446596

Burkart, N., & Huber, M. F. (2021, January). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. Retrieved 2024-05-04, from https://jair.org/index.php/jair/article/view/12228 doi: 10.1613/jair.1.12228

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27:1–27:27. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

Chou, E., Tramèr, F., Pellegrino, G., & Boneh, D. (2018). Sentinet: Detecting physical attacks against deep learning systems. *CoRR*, *abs/1812.00292*. Retrieved from http://arxiv.org/abs/1812.00292

Chromik, M. (2021). Making shap rap: Bridging local and global insights through interaction and narratives. In C. Ardito et al. (Eds.), *Human-computer interaction – interact 2021* (pp. 641–651). Cham: Springer International Publishing.

Delobelle, P., Winters, T., & Berendt, B. (2020, November). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3255–3265). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.findings-emnlp.292 doi: 10.18653/v1/2020.findings-emnlp.292

Deng, L. (2018, January). Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook [Perspectives]. *IEEE Signal Processing Magazine*, *35*(1), 180–177. Retrieved 2024-05-07, from http://ieeexplore.ieee.org/abstract/document/8253597?casa_token=ozs46PnUPQUAAAAA:bnhbRPZuFhSzW8YAtsCj6PJoezkq6QveXal_IyVSB1DfXTISAPAUNVQtdPVnbPj9rA4oKhut0YsowA (Conference Name: IEEE Signal Processing Magazine) doi: 10.1109/MSP.2017.2762725

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). *Bertje: A dutch bert model.* Retrieved from http://arxiv.org/abs/1912.09582 (cite arxiv:1912.09582)

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. Retrieved 2024-05-04, from https://arxiv.org/abs/1710.00794 (Publisher: [object Object] Version Number: 1) doi: 10.48550/ARXIV.1710.00794

European Commission. (2016). *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).* Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Commission. (2022). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act).* Retrieved from https://digital-strategy.ec.europa.eu/en/policies/ai-act

Filatova, E. (2012, May). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In N. Calzolari et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evalua-*

tion (LREC'12) (pp. 392–398). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved 2024-04-29, from http://www.lrec-conf.org/proceedings/lrec2012/pdf/661_Paper.pdf

Ghosh, D., & Muresan, S. (2018, Jun.). ' with 1 follower i must be awesome :p. '; exploring the role of irony markers in irony recognition. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15080 doi: 10.1609/icwsm.v12i1.15080

Greenwell, B. M., Boehmke, B. C., & McCarthy, A. J. (2018). *A simple and effective model-based variable importance measure.*

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Harrotuin. (n.d.). *Dutch news headlines.* Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/harrotuin/dutch-news-headlines

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* (To appear)

Jansen, N., & Chen, A. (2020, May). Prosodic encoding of sarcasm at the sentence level in Dutch. In (pp. 409–413). doi: 10.21437/SpeechProsody.2020-84

Jijkoun, V., & Hofmann, K. (2009, March). Generating a non-English subjectivity lexicon: Relations that matter. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)* (pp. 398–405). Athens, Greece: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E09-1046

Johnson, D. S., Hakobyan, O., & Drimalla, H. (2023, June). Towards interpretability in audio and visual affective machine learning: A review. *arXiv.org*. doi: 2306.08933v1

Jomar, D. (2023). *Pyale: Python accumulated local effects library.* Retrieved from https://github.com/DanaJomar/PyALE (Version 0.1.0)

Kalai, E., & Samet, D. (1987, September). On weighted Shapley values. *International Journal of Game Theory*, *16*(3), 205–222. Retrieved from https://doi.org/10.1007/BF01756292 doi: 10.1007/BF01756292

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., & Robnik-Šikonja, M. (2021, April). BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In H. Toivonen & M. Boggia (Eds.), *Proceedings of the eacl hackashop on news media content analysis and automated report gener-*

ation (pp. 16–21). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.hackashop-1.3

Kreuz, R., & Caucci, G. (2007, April). Lexical Influences on the Perception of Sarcasm. In A. Feldman & X. Lu (Eds.), *Proceedings of the Workshop on Computational Approaches to Figurative Language* (pp. 1–4). Rochester, New York: Association for Computational Linguistics. Retrieved 2024-05-07, from https://aclanthology.org/W07-0101

Kreuz, R. J., & Roberts, R. M. (1995). Two Cues for Verbal Irony: Hyperbole and the Ironic Tone of Voice. *Metaphor and Symbolic Activity*, *10*(1), 21–31. Retrieved from https://doi.org/10.1207/s15327868ms1001_3 (Publisher: Routledge _eprint: https://doi.org/10.1207/s15327868ms1001_3) doi: 10.1207/s15327868ms1001_3

Kulkarni, D. S., & Rodd., S. S. (2021, November). Sentiment Analysis in Hindi—A Survey on the State-of-the-art Techniques | ACM Transactions on Asian and Low-Resource Language Information Processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1–46. Retrieved 2024-03-06, from https://dl-acm-org.tilburguniversity.idm.oclc.org/doi/full/10.1145/3469722?casa_token=IQO8gJYeIfkAAAAA%3Axk2GJOAZeNQQS_evb1DWM8px6_6Wj7Sx3TBZQ1wUtaVlBxIcEGL6A5lEZ9F0TMRubLzhtx5CFTja2Q doi: https://doi.org/10.1145/3469722

Liang, L., Cai, X., & Su, Z. (2022). What is Decisive in Forecasting P2p Credit Loan Defaults? An Interpretable Analysis of Stacking Models Based on Shap and Ale Methods. *SSRN Electronic Journal*. Retrieved 2024-05-18, from https://www.ssrn.com/abstract=4052363 doi: 10.2139/ssrn.4052363

Liebrecht, C., Kunneman, F., & van den Bosch, A. (2013, June). The perfect solution for detecting sarcasm in tweets #not. In A. Balahur, E. van der Goot, & A. Montoyo (Eds.), *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 29–37). Atlanta, Georgia: Association for Computational Linguistics.

Lucariello, J. (1994, June). Situational irony: A concept of events gone awry: Journal of Experimental Psychology: General. *Journal of Experimental Psychology: General*, *123*(2), 129–145. Retrieved 2024-04-29, from https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1994-32940-001&site=ehost-live (Publisher: American Psychological Association) doi: 10.1037/0096-3445.123.2.129

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural infor-*

mation processing systems (Vol. 30). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maladry, A., Lefever, E., Van Hee, C., & Hoste, V. (2022, May). Irony Detection for Dutch: a Venture into the Implicit. In J. Barnes et al. (Eds.), *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (pp. 172–181). Dublin, Ireland: Association for Computational Linguistics. Retrieved 2024-03-06, from https://aclanthology.org/2022.wassa-1.16 doi: 10.18653/v1/2022.wassa-1.16

Maladry, A., Lefever, E., Van Hee, C., & Hoste, V. (2023a, July). A Fine Line Between Irony and Sincerity: Identifying Bias in Transformer Models for Irony Detection. In J. Barnes, O. De Clercq, & R. Klinger (Eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 315–324). Toronto, Canada: Association for Computational Linguistics. Retrieved 2024-03-06, from https://aclanthology.org/2023.wassa-1.28 doi: 10.18653/v1/2023.wassa-1.28

Maladry, A., Lefever, E., Van Hee, C., & Hoste, V. (2023b, 07). The limitations of irony detection in dutch social media. *Language Resources and Evaluation*, 1-32. doi: 10.1007/s10579-023-09656-1

Matos, P. F. (2021). *ECIAIR 2021 3rd European Conference on the Impact of Artificial Intelligence and Robotics*. Academic Conferences and publishing limited. (Google-Books-ID: PZBTEAAAQBAJ)

Maynard, D. G., & Greenwood, M. A. (2014, March). *Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis* [Proceedings Paper]. (Conference Name: Language Resources and Evaluation Conference (LREC) ISBN: 9782951740884 Meeting Name: Language Resources and Evaluation Conference (LREC) Place: Reykjavik, Iceland Publisher: ELRA)

Medhat, W., Hassan, A., & Korashy, H. (2014, December). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. Retrieved 2024-04-30, from https://www.sciencedirect.com/science/article/pii/S2090447914000550 doi: 10.1016/j.asej.2014.04.011

Merriam-webster. (n.d.). https://www.merriam-webster.com/dictionary/metaphor. (Retrieved [Access Date], from https://www.merriam-webster.com/dictionary/metaphor)

Mladenović, M., Krstev, C., Mitrović, J., & Stanković, R. (2017, September). Using Lexical Resources for Irony and Sarcasm Classification. In *Proceedings of the 8th Balkan Conference in Informatics* (pp. 1–8). New York, NY, USA: Association for Computing Machinery. Retrieved 2024-05-

07, from https://dl.acm.org/doi/10.1145/3136273.3136298 doi: 10.1145/3136273.3136298

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

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022, October). SHAP-based explanation methods: A review for NLP interpretability. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2022.coling-1.406

Nayak, A., & Timmapathini, H. (2021). Using integrated gradients to explain linguistic acceptability learnt by BERT. *CoRR*, *abs/2106.07349*. Retrieved from https://arxiv.org/abs/2106.07349

NOS. (2024). *Computer herkent sarcasme steeds beter.* Retrieved from https://nos.nl/artikel/2520728-computer-herkent-sarcasme-steeds-beter (Accessed: 2024-05-17)

Nu.nl. (n.d.). https://www.nu.nl/.

pandas development team, T. (2020, February). *pandas-dev/pandas: Pandas.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3509134 doi: 10.5281/zenodo.3509134

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pierse, C. (2024). *transformers-interpret: Explainability for huggingface transformers.* https://github.com/cdpierse/transformers-interpret. GitHub.

Qian, S., Pham, V. H., Lutellier, T., Hu, Z., Kim, J., Tan, L., ... Shah, S. (2021). Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 30211–30227). Curran Associates, Inc. Retrieved 2024-05-07, from https://proceedings.neurips.cc/paper/2021/hash/fdda6e957f1e5ee2f3b311fe4f145ae1-Abstract.html

Rahma, A., Azab, S. S., & Mohammed, A. (2023). A Comprehensive Survey on Arabic Sarcasm Detection: Approaches, Challenges and Future Trends. *IEEE Access*, *11*, 18261–18280. Retrieved 2024-03-06, from http://ieeexplore.ieee.org/abstract/document/10049545 (Conference Name: IEEE Access) doi: 10.1109/ACCESS.2023.3247427

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, *abs/1602.04938*. Retrieved from http://arxiv.org/abs/1602.04938

Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020, September). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. *International Journal of Market Research*, *62*(5), 578–598. (Publisher: SAGE Publications) doi: 10.1177/1470785320921779

Silva, S. J., Keller, C. A., & Hardin, J. (2022). Using an Explainable Machine Learning Approach to Characterize Earth System Model Errors: Application of SHAP Analysis to Modeling Lightning Flash Occurrence. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002881. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002881 (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002881) doi: https://doi.org/10.1029/2021MS002881

Smedt, T. D., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, *13*(66), 2063–2067. Retrieved from http://jmlr.org/papers/v13/desmedt12a.html

speld.nl. (n.d.). https://www.speld.nl/.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *CoRR*, *abs/1703.01365*. Retrieved from http://arxiv.org/abs/1703.01365

Tiwari, R. S. (2024). Hate speech detection using LSTM and explanation by LIME (local interpretable model-agnostic explanations). In *Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications* (pp. 93–110). Elsevier. Retrieved 2024-02-19, from https://linkinghub.elsevier.com/retrieve/pii/B9780443220098000057 doi: 10.1016/B978-0-443-22009-8.00005-7

Tulkens, S., Emmery, C., & Daelemans, W. (2016, may). Evaluating unsupervised dutch word embeddings as a linguistic resource. In N. C. C. Chair) et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)*. Paris, France: European Language Resources Association (ELRA).

Van de Kauter, Marjan and Coorman, Geert and Lefever, Els and Desmet, Bart and Macken, Lieve and Hoste, Veronique. (2013). LeTs preprocess: the multilingual LT3 linguistic preprocessing toolkit. *COMPUTATIONAL LINGUISTICS IN THE NETHERLANDS JOURNAL*, *3*, 103–120.

Van Hee, C. (2017). *Can machines sense irony? : exploring automatic irony detection on social media* (dissertation, Ghent University). Retrieved

2024-04-15, from http://hdl.handle.net/1854/LU-8531569

Van Hee, C., De Clercq, O., & Hoste, V. (2021, April). Exploring Implicit Sentiment Evoked by Fine-grained News Events. In O. De Clercq et al. (Eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 138–148). Online: Association for Computational Linguistics. Retrieved 2024-03-07, from https://aclanthology.org/2021.wassa-1.15

What is explainable AI? | IBM. (n.d.). Retrieved 2024-02-19, from https://www.ibm.com/topics/explainable-ai

Yacoub, A. D., Slim, S., & Aboutabl, A. (2024, January). A Survey of Sentiment Analysis and Sarcasm Detection: Challenges, Techniques, and Trends. *International journal of electrical and computer engineering systems*, *15*(1), 69–78. Retrieved 2024-01-29, from https://hrcak.srce.hr/313458 (Publisher: Elektrotehnički fakultet Sveučilišta J.J. Strossmayera u Osijeku) doi: 10.32985/ijeces.15.1.7

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(5), 726-742. doi: 10.1109/TETCI.2021.3100641

APPENDIX A

A list of all features used in predicting sarcasm in the SVM model ((Van Hee, 2017)):

- Part-of-speech features. For each tag, indicate:
  - Whether it occurs in the headline
  - Whether it occurs 0, 1, or 2 times
  - The frequency of the tag in absolute numbers.
  - The frequency of the tag as a percentage.
- Named entity features:
  - Binary feature, indicating the presence of a named entity in the tweet
  - The number of named entities in the text
  - The number of tokens that are part of a named entity
  - The frequency of tokens that are part of a named entity
- Semantic features:

  - A feature vector based on Dutch embeddings for each token in a sentence

- Sentiment features:
  - The number of positive, negative and neutral lexicon words averaged over text length
  - The overall tweet polarity (i.e. the sum of the values of the identified sentiment words)
  - The difference between the highest positive and lowest negative sentiment values;
  - Binary feature, indicating the presence of a polarity contrast between two lexicon words (i.e. at least one positive and one negative lexicon word are present).

APPENDIX B

Table 2: Correct Predictions by SVM (Sarcastic)

Headline	**SVM Prediction**	**BERTje Prediction**	**True Label**
Rennen voor Jan Roos in de straten van Enkhuizen	sarcastic	non-sarcastic	sarcastic
Waarom hoor je in de media niks over het gevaarlijke coronavirus?	sarcastic	non-sarcastic	sarcastic
Het nieuwe Cubaanse programma CastroTV praat met overleden revolutionairen	sarcastic	non-sarcastic	sarcastic
Vrouwen bij Korps Mariniers mogen eindelijk voor hun geslacht uitkomen	sarcastic	non-sarcastic	sarcastic
Aanklager Pistorius heeft te weinig tijd om alle argumenten uiteen te zetten	sarcastic	non-sarcastic	sarcastic
Buma voert keihard oppositie bij formatieonderhandelingen	sarcastic	non-sarcastic	sarcastic
Dit gebeurde er vannacht in Raalte	sarcastic	non-sarcastic	sarcastic
Man stelt plinten leggen uit door stikstofuitspraak	sarcastic	non-sarcastic	sarcastic

Continued on next page

Table 2 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Dit zijn de rapportcijfers van Rutte III	sarcastic	non-sarcastic	sarcastic
Kamer plaatst vraagtekens bij studietoelage prins Bernhard	sarcastic	non-sarcastic	sarcastic
Referendum Italië: een overwinning voor tegenstanders van Italiaanse Senaatshervorming over de hele wereld	sarcastic	non-sarcastic	sarcastic
Opinie: Mag het hellend-vlak-argument straks ook al niet meer?	sarcastic	non-sarcastic	sarcastic
Emmen wil binnenstad opknappen met oefeninterland Engeland-Rusland	sarcastic	non-sarcastic	sarcastic
Bijleveld vindt dat Kamer burgerdoden nu eens rust moet gunnen	sarcastic	non-sarcastic	sarcastic
Brussel verenigt zich in 19 gemeentes, 6 politiezones en 2 talen	sarcastic	non-sarcastic	sarcastic
Belediging staatshoofden voortaan geregeld via modelcontracten	sarcastic	non-sarcastic	sarcastic
Biden probeert opzichtig vrouwonvriendelijk succes van Trump te kopiëren	sarcastic	non-sarcastic	sarcastic
Supermarkt in Indonesië verkoopt pakjes uit de Hollandse keuken	sarcastic	non-sarcastic	sarcastic
Toeristen op besmet cruiseschip balen dat ze al weken op een boot zitten	sarcastic	non-sarcastic	sarcastic
Moet Nederland inlichtingenhulp aan arme landen als de VS afbouwen?	sarcastic	non-sarcastic	sarcastic
Erdogan laat printer arresteren wegens verspreiden van kritisch artikel	sarcastic	non-sarcastic	sarcastic
Nieuwe partij Zeus Leeft! spreekt veel conservatieve kiezers aan	sarcastic	non-sarcastic	sarcastic
Extra beveiliging voor Kabouter Buttplug in Rotterdam	sarcastic	non-sarcastic	sarcastic

Continued on next page

Table 2 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Nederland onthoudt zich van stemming over zwangerschap Kate Middleton	sarcastic	non-sarcastic	sarcastic
Republikeinen overwegen eigen partij op te richten	sarcastic	non-sarcastic	sarcastic
Op NS hartkloppingen kunnen treinreizigers hun hart- en vaatproblemen delen	sarcastic	non-sarcastic	sarcastic

Table 3: Correct Predictions by SVM (Non-Sarcastic)

Headline	SVM Prediction	BERTje Prediction	True Label
Slob wil ook bestuur Haagse hindoeschool wegsturen	non-sarcastic	sarcastic	non-sarcastic
Stofstorm hult Phoenix in duisternis	non-sarcastic	sarcastic	non-sarcastic
Wandelaar treft 4 meter lange python aan in sloot in Etten-Leur	non-sarcastic	sarcastic	non-sarcastic
Minister Blok: Rusland blijft het Westen confronteren	non-sarcastic	sarcastic	non-sarcastic
Dinsdag weer een coronapersconferentie: hier staan we nu	non-sarcastic	sarcastic	non-sarcastic
Pakistanen zoeken in sneeuw naar slachtoffers lawine	non-sarcastic	sarcastic	non-sarcastic
Marianne Thieme volgende week terug in de Tweede Kamer na ziekte	non-sarcastic	sarcastic	non-sarcastic
Omstreden Palmen opnieuw wethouder in Brunssum	non-sarcastic	sarcastic	non-sarcastic
Van cowboys tot allemaal bruin: sigarettenpakjes door de jaren heen	non-sarcastic	sarcastic	non-sarcastic
Nek-aan-nekrace in VS kan nog alle kanten op door telling poststemmen	non-sarcastic	sarcastic	non-sarcastic
Jouw instapgids voor de Amerikaanse presidentsverkiezingen	non-sarcastic	sarcastic	non-sarcastic

Continued on next page

Table 3 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Coronagezant Sijbesma begrijpt niet dat mensen zich niet laten testen	non-sarcastic	sarcastic	non-sarcastic
F-35 (JSF) kampt weer met problemen: voorlopig aan de grond bij onweer	non-sarcastic	sarcastic	non-sarcastic
Intimidatie en mishandeling: waarom de boa meer wapens eist	non-sarcastic	sarcastic	non-sarcastic
Desi Bouterse eist hertelling van stemmen Suriname	non-sarcastic	sarcastic	non-sarcastic
Gesprekken over pensioenakkoord na ultieme poging toch mislukt	non-sarcastic	sarcastic	non-sarcastic
KNMI meet warmste week ooit: gemiddeld 33,1 graden	non-sarcastic	sarcastic	non-sarcastic
Coronagevolgen in de zorg: agressieve jongeren, ouderen stoppen met eten	non-sarcastic	sarcastic	non-sarcastic
Onderzoekscommissie VK: Moskou poogde Schots referendum te beinvloeden	non-sarcastic	sarcastic	non-sarcastic
Opkomst Amsterdam 9,3 procent, meeste stemmers in Utrecht	non-sarcastic	sarcastic	non-sarcastic
Nederland heeft 25 gemeenten minder vanaf 2019	non-sarcastic	sarcastic	non-sarcastic
App waarschuwt voor teken: dit zijn de risicogebieden	non-sarcastic	sarcastic	non-sarcastic
Drone roept Russen in afgelegen gebieden op om thuis te blijven	non-sarcastic	sarcastic	non-sarcastic
Klusjesman Ruinerwold dreigt met hongerstaking tot dood	non-sarcastic	sarcastic	non-sarcastic
Voortaan altijd celstraf voor geweld tegen hulpverlener	non-sarcastic	sarcastic	non-sarcastic
Meerderheid Nederlanders wil rookverbod op plekken met kinderen	non-sarcastic	sarcastic	non-sarcastic
Vrijgelaten Wit-Russische gevangenen tonen gevolgen mishandeling	non-sarcastic	sarcastic	non-sarcastic

Continued on next page

Table 3 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Opnieuw recordaantal positieve tests afgenomen: bijna 2.000 binnen één dag	non-sarcastic	sarcastic	non-sarcastic
D66 en PVV verliezen flink, PvdA verrassend de grootste	non-sarcastic	sarcastic	non-sarcastic
EU weert Pakistaanse vliegmaatschappijen half jaar om dubieuze brevetten	non-sarcastic	sarcastic	non-sarcastic
Muteknop moet laatste debat tussen Trump en Biden in goede banen leiden	non-sarcastic	sarcastic	non-sarcastic
Halbe Zijlstra trekt zich terug uit race om post Wereldbank	non-sarcastic	sarcastic	non-sarcastic
Gemeente Den Haag: Situatie Malieveld beheersbaar gebleven	non-sarcastic	sarcastic	non-sarcastic
Verhoren toeslagenaffaire week 1: Eindelijk spreken de hoge ambtenaren	non-sarcastic	sarcastic	non-sarcastic
Wetenschappelijk onderzoekscentrum minder afhankelijk van ministerie	non-sarcastic	sarcastic	non-sarcastic
Raad van State kraakt nieuwe Arbeidswet	non-sarcastic	sarcastic	non-sarcastic
Koranverbranding wakkert protesten in Malmö aan	non-sarcastic	sarcastic	non-sarcastic
Lege schappen en doodse stilte: Nederlander zit vast in afgesloten Wuhan	non-sarcastic	sarcastic	non-sarcastic
Honderden gestrande Nederlanders teruggehaald uit Peru	non-sarcastic	sarcastic	non-sarcastic
Te koop aangeboden: lok haar met bloed van Amerikaanse president Lincoln	non-sarcastic	sarcastic	non-sarcastic
Betogers wereldwijd over waarom zij demonstreren tegen racisme	non-sarcastic	sarcastic	non-sarcastic
Oproep aan Noord-Hollanders: Ga niet langs de deuren met Sint-Maarten	non-sarcastic	sarcastic	non-sarcastic

Continued on next page

Table 3 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Aangestoken en blusbaar: zes misverstanden over de Australische branden	non-sarcastic	sarcastic	non-sarcastic
Bloomberg stak al bijna half miljard dollar uit eigen zak in campagne	non-sarcastic	sarcastic	non-sarcastic
Versoberde Pride gestart met toespraak zonder publiek bij Homomonument	non-sarcastic	sarcastic	non-sarcastic
Opnieuw wapenstilstand in conflict om Nagorno-Karabach	non-sarcastic	sarcastic	non-sarcastic
Ontsnapte man opgepakt die Belgische gevangenis ansichtkaart stuurde	non-sarcastic	sarcastic	non-sarcastic
Fraude ten gunste van grijze kiwi ontdekt bij Vogel van het Jaarverkiezing	non-sarcastic	sarcastic	non-sarcastic
VVD wil toch geen uitbreiding verkoopverbod vuurwerk	non-sarcastic	sarcastic	non-sarcastic
Menselijk strottenhoofd gevonden tussen spullen bij Zwolse kringloop	non-sarcastic	sarcastic	non-sarcastic
Groei van de luchtvaart is steeds minder sexy in Den Haag	non-sarcastic	sarcastic	non-sarcastic
Nederland vecht pulsvisverbod aan bij Europees Hof	non-sarcastic	sarcastic	non-sarcastic
Trump wil dat iedereen die New York ontvlucht in zelfquarantaine gaat	non-sarcastic	sarcastic	non-sarcastic
Hoogste militair VS heeft er spijt van dat hij in uniform naast Trump liep	non-sarcastic	sarcastic	non-sarcastic
Trump ruziet met journalist over prijsgeven uitslag coronatest	non-sarcastic	sarcastic	non-sarcastic
Profiel: Dit was het leven van George Floyd	non-sarcastic	sarcastic	non-sarcastic
Kabinet investeert honderden miljoenen in infrastructuur Nederland	non-sarcastic	sarcastic	non-sarcastic

Continued on next page

Table 3 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Halsema: Jeugd wil geen door gemeente georganiseerd Oud en Nieuw	non-sarcastic	sarcastic	non-sarcastic
Kabinet onderzoekt inreisverbod voor extremistische predikant	non-sarcastic	sarcastic	non-sarcastic
Corona-experiment: Duitse proefpersonen wonen concert bij	non-sarcastic	sarcastic	non-sarcastic
Democraten presenteren wetsvoorstel voor politiehervorming VS	non-sarcastic	sarcastic	non-sarcastic
Verdachte dood Boldewijn trekt verklaring dat hij slachtoffer in water zag in	non-sarcastic	sarcastic	non-sarcastic
Strengere coronaregels in Belgie: winkelen moet weer alleen	non-sarcastic	sarcastic	non-sarcastic
Studenten klagen: drie dagen voor start studiejaar nog steeds geen rooster	non-sarcastic	sarcastic	non-sarcastic
Onrust op Curaçao na zeeblokkade, woede richt zich op Blok	non-sarcastic	sarcastic	non-sarcastic
Het conflict in Ethiopie escaleert: wat is er aan de hand?	non-sarcastic	sarcastic	non-sarcastic
Blok geeft 4 miljoen euro voor grensbewaking Niger	non-sarcastic	sarcastic	non-sarcastic
Studenten breken wereldduurrecord zonneracen: 924 kilometer in twaalf uur	non-sarcastic	sarcastic	non-sarcastic
Fiscus breekt belofte: ouders in toeslagenaffaire ontvangen dossier later	non-sarcastic	sarcastic	non-sarcastic
KNMI maakt stormnamen nieuw seizoen bekend: Aiden, Evert en Klaas	non-sarcastic	sarcastic	non-sarcastic
Jacob Blake verschijnt vanuit ziekenhuisbed voor rechter	non-sarcastic	sarcastic	non-sarcastic
Dijkhoff heeft geen problemen met omstreden tweets van nieuw VVD-Kamerlid	non-sarcastic	sarcastic	non-sarcastic
Advocaat van Jos B. vindt maar één uitkomst mogelijk: vrijspraak	non-sarcastic	sarcastic	non-sarcastic

Continued on next page

Table 3 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Lodeweges en Wijnaldum blikken vooruit op Oranje tegen Polen	non-sarcastic	sarcastic	non-sarcastic
Kabinet: Webwinkel moet klant minder gemakkelijk krediet verstrekken	non-sarcastic	sarcastic	non-sarcastic
Tweede dode door raadselachtig nieuw virus in China	non-sarcastic	sarcastic	non-sarcastic

APPENDIX C

Table 4: Correct Predictions by BERTje (Sarcastic)

Headline	SVM Prediction	BERTje Prediction	True Label
Doorgaan Zomergasten onzeker door politiestaking	non-sarcastic	sarcastic	sarcastic
Brazilië doneert Amazonewoud aan getroffen gebieden Australië	non-sarcastic	sarcastic	sarcastic
Basisschool-stelletje reserveert alvast kinderopvangplaats voor eerste kind	non-sarcastic	sarcastic	sarcastic
Nekvel-docent aangenomen in Tweede Kamer om Kamerleden tucht bij te brengen	non-sarcastic	sarcastic	sarcastic
Opnieuw corruptieschandaal in omkoopwereld	non-sarcastic	sarcastic	sarcastic
Deze 5 maatregelen moeten verruwing in de Kamer tegengaan	non-sarcastic	sarcastic	sarcastic
Huisjesmelkers willen steunpakket voor studenten zodat die exorbitante huren kunnen blijven betalen	non-sarcastic	sarcastic	sarcastic
Ga je naar buiten? Pas op voor de tekenprocessievleermuis!	non-sarcastic	sarcastic	sarcastic
Willem-Alexander is vanaf 1 januari 2019 mogelijk een lulhannes	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Groot gedeelte vis in de Noordzee is over de datum	non-sarcastic	sarcastic	sarcastic
Canada begint nucleair programma in hoop op betere relatie met Trump	non-sarcastic	sarcastic	sarcastic
Nieuwe digitale afdeling van Defensie onthult kantoorcamouflagepakken	non-sarcastic	sarcastic	sarcastic
President Obama verstopt nucleaire codes in exemplaar van de grondwet	non-sarcastic	sarcastic	sarcastic
Loekasjenko benoemt waterkanon tot minister van Binnenlandse Zaken	non-sarcastic	sarcastic	sarcastic
Eerste Kamerleden moeten logo's van bedrijven waar ze nevenfuncties bekleden op pakken dragen	non-sarcastic	sarcastic	sarcastic
Iraniërs opgelucht: extra economische sancties van VS welkome afleiding van onderdrukking door regime	non-sarcastic	sarcastic	sarcastic
Ook Peppa Pig haalt in nieuwe aflevering hard uit naar China	non-sarcastic	sarcastic	sarcastic
Kabinet redt toerismesector, koopt duizenden wietmutsen	non-sarcastic	sarcastic	sarcastic
Poetin presenteert nieuw bewijs Oekraïense straaljager	non-sarcastic	sarcastic	sarcastic
Strava-statistieken wijzen uit: Mark uit Den Haag loopt elke avond rondje langs hoofdkantoor Shell	non-sarcastic	sarcastic	sarcastic
De toekomst: is-ie wel te vertrouwen?	non-sarcastic	sarcastic	sarcastic
Compromis: Rotterdam geen hoofdstad, maar krijgt wel 020 als netnummer	non-sarcastic	sarcastic	sarcastic
Opnieuw een terroristische aanslag voorkomen. Waarom doet de overheid niets?	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Jaap van Dissel heeft nachtmerrie waarin mondkapjes helpen tegen het coronavirus	non-sarcastic	sarcastic	sarcastic
Voorlopige prognose: Wilders veroordeeld voor haatzaaien	non-sarcastic	sarcastic	sarcastic
Meerderheid Amerikanen redt democratische rechten van landgenoten die daar niet om geven	non-sarcastic	sarcastic	sarcastic
Wie te geloven over MH17? Onderzoeksteam en Facebooker Conny spreken elkaar tegen	non-sarcastic	sarcastic	sarcastic
Russische hackers stellen orde op zaken bij Belastingdienst	non-sarcastic	sarcastic	sarcastic
Assad: ik zat fout met gifgas	non-sarcastic	sarcastic	sarcastic
Rusland bestookt IS ook in Oekraïne	non-sarcastic	sarcastic	sarcastic
Clingendael bewijst potentieel nut pantserwagens tijdens mei 1940	non-sarcastic	sarcastic	sarcastic
Nederland heeft storm goed doorstaan	non-sarcastic	sarcastic	sarcastic
Petra's schilderijen bieden ook in tijden van corona geen troost	non-sarcastic	sarcastic	sarcastic
KNMI overweegt overstap naar Fahrenheitschaal	non-sarcastic	sarcastic	sarcastic
Fred kan na drie maanden horecasluiting eindelijk weer geen fooi geven	non-sarcastic	sarcastic	sarcastic
Bram was net van plan zichzelf een paar maanden in huis op te sluiten, en toen kwam corona...	non-sarcastic	sarcastic	sarcastic
Storing: vanochtend geen schandaal rond Trump	non-sarcastic	sarcastic	sarcastic
Winterdorp op zolder Oom Jan gaat gebukt onder kleine criminaliteit	non-sarcastic	sarcastic	sarcastic
Poetin doneert aan Rode Kruis na vergiftiging Navalny	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Belangenverstrengeling? Bert Koenders is lid van de terreurgroep die hij als minister steun gaf	non-sarcastic	sarcastic	sarcastic
Shit, Jorinde (23) leeft nog wanneer gevolgen klimaatverandering desastreus zullen zijn	non-sarcastic	sarcastic	sarcastic
Tsjechisch wonderkind (15) onthoudt 12 minuten lang naam van gesprekspartner	non-sarcastic	sarcastic	sarcastic
PvdA nu al meer dan 1720 dagen vast in kabinet	non-sarcastic	sarcastic	sarcastic
Nederlanders delen massaal selfies van terugkeer bij sekswerkers	non-sarcastic	sarcastic	sarcastic
Ridouan T. compenseert privévlucht door komende jaren sober te leven	non-sarcastic	sarcastic	sarcastic
Operaties voor bodemprijzen bij faillissementsuitverkoop ziekenhuizen	non-sarcastic	sarcastic	sarcastic
Naar het onweer kijken? Dit zijn de mooiste open plekken om het noodweer te zien!	non-sarcastic	sarcastic	sarcastic
Brits parlement niet akkoord met aftreden May	non-sarcastic	sarcastic	sarcastic
Amersfoort en Apeldoorn niet significant verschillend	non-sarcastic	sarcastic	sarcastic
Scheveningse deugbrigade probeert hele stad onder de roetvegen te krijgen	non-sarcastic	sarcastic	sarcastic
Wereldwijde kritiek op passieve Ploumen na inreisverbod Trump	non-sarcastic	sarcastic	sarcastic
Onsmakelijk: beelden van typisch Brabants gehucht uitgelekt na stalbrand	non-sarcastic	sarcastic	sarcastic
Kernraketten Volkel gecontroleerd tot ontploffing gebracht	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
DDoS-storing: Shell neemt urenlang verantwoordelijkheid voor Groningen	non-sarcastic	sarcastic	sarcastic
Vergevingsgezind Nederland geeft coronavirus tweede kans	non-sarcastic	sarcastic	sarcastic
Rutte gaat ogen dicht knijpen en heel hard hopen dat Turkije verandert	non-sarcastic	sarcastic	sarcastic
Pechtold deelt organen met D66-vlaggetjes uit in ziekenhuis	non-sarcastic	sarcastic	sarcastic
India stemt over terugdraaien onafhankelijkheid	non-sarcastic	sarcastic	sarcastic
Britten gooien grens open voor belastingvluchtelingen	non-sarcastic	sarcastic	sarcastic
Compensatie voor Zeeland: groot asielzoekerscentrum naar Vlissingen	non-sarcastic	sarcastic	sarcastic
Joris van Kraats (8) sluit zich aan bij rebellenleger Kony	non-sarcastic	sarcastic	sarcastic
Ank Bijleveld stapt op als beoogd minister van Defensie	non-sarcastic	sarcastic	sarcastic
Gemeente Amsterdam voelt na half uur nog steeds niks van zerotolerancebeleid	non-sarcastic	sarcastic	sarcastic
Alcohol op rantsoen in Den Helder en bij Telstar	non-sarcastic	sarcastic	sarcastic
Tweede Kamer van Koophandel brengt politici en ondernemers in de war	non-sarcastic	sarcastic	sarcastic
Ex-prostituees snel weer aan de slag	non-sarcastic	sarcastic	sarcastic
GroenLinks wil opheldering over identiteit fractievoorzitter	non-sarcastic	sarcastic	sarcastic
Jan Dijkgraaf hoort stemmers in z'n hoofd	non-sarcastic	sarcastic	sarcastic
Deze nieuwe gezichten moeten de VVD uit het slop trekken	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Uitgelekt: nieuwe coalitie wil geen plannen meer doorspelen naar de pers	non-sarcastic	sarcastic	sarcastic
Japanse boer vindt atoombom in rijstveld	non-sarcastic	sarcastic	sarcastic
Analyse: Clinton won op inhoud, Trump won op fascisme	non-sarcastic	sarcastic	sarcastic
Samsom: laat PvdA-beleid over aan gemeentes	non-sarcastic	sarcastic	sarcastic
Den Haag verbiedt carbidschieten: slecht nieuws voor jongeren die 50 meter hoge melkbus bouwen	non-sarcastic	sarcastic	sarcastic
Tien verschillen tussen Wilders en Hitler	non-sarcastic	sarcastic	sarcastic
Fabrikant van standbeelden van slavenhouders maakt zich zorgen over toekomst van bedrijf	non-sarcastic	sarcastic	sarcastic
Etnisch profileren Belastingdienst: multinationals met dubbele nationaliteit werden extra soepel gecontroleerd	non-sarcastic	sarcastic	sarcastic
Amsterdamse man opgeslokt door Japans tourgroepje	non-sarcastic	sarcastic	sarcastic
Frankrijk ligt dwars bij vernietiging Straatsburg	non-sarcastic	sarcastic	sarcastic
Dit iconische staafdiagram gaat de wereld over	non-sarcastic	sarcastic	sarcastic
VVD wil maximumsnelheid voor zetelverlies verlagen	non-sarcastic	sarcastic	sarcastic
Opwarming ophitserij van de Telegraaf stijgt sneller dan verwacht	non-sarcastic	sarcastic	sarcastic
WhatsApper Ronnie al drie maanden aan het typen	non-sarcastic	sarcastic	sarcastic
Open dagen kerncentrale Fukushima	non-sarcastic	sarcastic	sarcastic
Orbán: vluchtelingen vormen bedreiging voor fascistoïde identiteit Hongarije	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Brits parlement kiest voor permanente teringzooi	non-sarcastic	sarcastic	sarcastic
Maandag 7 december: de eerste reacties	non-sarcastic	sarcastic	sarcastic
Amsterdamse veerpontjes krijgen rem en roer	non-sarcastic	sarcastic	sarcastic
Jonge ouders beginnen steeds later aan oprichting politieke partij	non-sarcastic	sarcastic	sarcastic
Gemeenten zetten draaiorgels in om mensen uit parken te jagen	non-sarcastic	sarcastic	sarcastic
Dispuut twijfelt over opbrengst lustrumveiling: aids de wereld uit of op reis naar Bali	non-sarcastic	sarcastic	sarcastic
Amsterdam de komende vier jaar zonder oppositie	non-sarcastic	sarcastic	sarcastic
Weekendtips: foodtruckfestival en een nieuwe militaire dictatuur	non-sarcastic	sarcastic	sarcastic
OM: motorclubs zetten aan tot pedofilie	non-sarcastic	sarcastic	sarcastic
Oppositie schenkt coalitie 76e zetel om Van Haga te kunnen dumpen	non-sarcastic	sarcastic	sarcastic
Gelekt: appgesprek tussen Halsema en Grapperhaus op bruiloft	non-sarcastic	sarcastic	sarcastic
Vaticaan grijpt in: priesters mogen geen seks meer	non-sarcastic	sarcastic	sarcastic
Verdacht trajectje gevonden tussen Ommen en Almelo	non-sarcastic	sarcastic	sarcastic
Qatar bouwt Saoedi-Arabië, Bahrein, Egypte en Verenigde Arabische Emiraten na	non-sarcastic	sarcastic	sarcastic
Middelbareschooladvies Frank: 2 jaar gymnasium en dan wietverslaafd overstappen naar havo	non-sarcastic	sarcastic	sarcastic
Hoogwater Maas en Rijn: Nederlanders gevraagd om half uur extra te douchen	non-sarcastic	sarcastic	sarcastic
Ook schandalen stappen over van VVD naar Forum	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Ook Rutte stemt: "Fantastisch om eindelijk invloed te hebben op Europees beleid."	non-sarcastic	sarcastic	sarcastic
Regen op Koningsdag: veel stukken afdak in Amsterdam al afgetapet	non-sarcastic	sarcastic	sarcastic
Kamerlid Teeven eist opheldering over bonnetjesaffaire	non-sarcastic	sarcastic	sarcastic
Demonstranten Hongkong krijgen zesde ster	non-sarcastic	sarcastic	sarcastic
LIVEBLOG: Rick weet zich geen houding te geven tegenover vrouw in trein	non-sarcastic	sarcastic	sarcastic
Ongeïntroduceerde studenten: kunnen onze steden dat wel aan?	non-sarcastic	sarcastic	sarcastic
Kiezers voldoen niet aan eisen GroenLinks	non-sarcastic	sarcastic	sarcastic
Treiterpremier zorgt voor overlast in parlement	non-sarcastic	sarcastic	sarcastic
IS keert zich tegen bombarderen in Syrië	non-sarcastic	sarcastic	sarcastic
Vier nieuw-rechtse partijen raken slaags bij all you can eat-buffet	non-sarcastic	sarcastic	sarcastic
Malieveld vol: bouw zet extra verdieping op veld om te kunnen protesteren	non-sarcastic	sarcastic	sarcastic
Tata Steel weigert verantwoordelijkheid te nemen voor kikkerregen in Wijk aan Zee	non-sarcastic	sarcastic	sarcastic
De revolte van Mat Herben: tien jaar later	non-sarcastic	sarcastic	sarcastic
Jaap van Dissel praat Kamer 2 uur lang bij over gebarentolkwissel	non-sarcastic	sarcastic	sarcastic
Onbekende man haalt 40 miljoen binnen voor Defensie	non-sarcastic	sarcastic	sarcastic
Geen regels, geen grenzen: eerste Vrije Rijschool geopend in Amsterdam	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Nederlandse rechter overleden: geen enkel gevolg voor verkiezingen	non-sarcastic	sarcastic	sarcastic
VN uit kritiek op Iraans vrouwknikkeren	non-sarcastic	sarcastic	sarcastic
Strijdende partijen in Syrië trekken lootjes voor bondgenootschappen	non-sarcastic	sarcastic	sarcastic
Wat als deze man straks over nucleaire wapens beschikt?	non-sarcastic	sarcastic	sarcastic
Geen roze olifanten herdacht op 4 mei	non-sarcastic	sarcastic	sarcastic
Wat vinden mensen op zinkende schepen eigenlijk van de Brexit-chaos?	non-sarcastic	sarcastic	sarcastic
Orthopeden ineens superspastisch over het delen van stompfoto's	non-sarcastic	sarcastic	sarcastic
Hoe heeft Nederland alle extra vrije tijd tot nu toe besteed?	non-sarcastic	sarcastic	sarcastic
Liefhebbers mensenrechten Rusland vallen na 15 juli in zwart gat	non-sarcastic	sarcastic	sarcastic
Asscher: PvdA creëert komende periode 17 nieuwe waarden	non-sarcastic	sarcastic	sarcastic
Arie Slob was jarenlang een af en toe in beeld verschijnende politicus	non-sarcastic	sarcastic	sarcastic
Partij voor de Dieren wil stemrecht voor 4 tot 10-jarigen	non-sarcastic	sarcastic	sarcastic
Politici mogelijk niet naar WK handbal in Qatar	non-sarcastic	sarcastic	sarcastic
Toestand Joling stabiel na aanrijden van oud vrouwtje	non-sarcastic	sarcastic	sarcastic
Nederland gerustgesteld: Rutte gebruikte geen informatie voor nemen beslissing	non-sarcastic	sarcastic	sarcastic
Wilders: Moeilijk om geschikte mensen te vinden als beginnende partij	non-sarcastic	sarcastic	sarcastic

Continued on next page

Table 4 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Landen die wél gewoon 4Klaas Willems: route via A2 en dan binnendoor is sneller	non-sarcastic	sarcastic	sarcastic
Versoepelen bonusregels: er kán eigenlijk niks mis gaan!	non-sarcastic	sarcastic	sarcastic
Duivels dilemma Republikeinen: goed zorgplan houden, of slecht plan invoeren?	non-sarcastic	sarcastic	sarcastic
Dronken Eerste Kamerleden stelen vlag uit Tweede Kamer	non-sarcastic	sarcastic	sarcastic
Zwitserland wil in de volgende oorlog genderneutraal zijn	non-sarcastic	sarcastic	sarcastic
Wereld vreest gebrek aan marketingmedewerkers als Amsterdam in quarantaine moet	non-sarcastic	sarcastic	sarcastic
Opnieuw ophef over verbouwing Binnenhof: moeten er wel glory holes komen in de plenaire zaal?	non-sarcastic	sarcastic	sarcastic
Uitbreiding Schiphol noodzakelijk voor gigantische repen Toblerone in taxfreewinkels	non-sarcastic	sarcastic	sarcastic

Table 5: Correct Predictions by BERTje (Non-Sarcastic)

Headline	SVM Prediction	BERTje Prediction	True Label
Linkse partijen willen van belastingvoordeel Shell af	sarcastic	non-sarcastic	non-sarcastic
Waarom ook Rutte kritiek krijgt als de koning iets doet wat omstreden is	sarcastic	non-sarcastic	non-sarcastic
Weerbericht: Er staat ons dit weekend van alles te wachten	sarcastic	non-sarcastic	non-sarcastic
Gaat de opvangtoeslagaffaire Snel nu alsnog de kop kosten?	sarcastic	non-sarcastic	non-sarcastic
Man krijgt taakstraf voor fatale aanrijding bij Pinkpop	sarcastic	non-sarcastic	non-sarcastic

Continued on next page

Table 5 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Wereldleiders beloven in aanloop naar VN-top aantasting natuur te stoppen	sarcastic	non-sarcastic	non-sarcastic
Zo streng is ons coronabeleid vergeleken met dat in de landen om ons heen	sarcastic	non-sarcastic	non-sarcastic
Nederlander mogelijk betrokken bij Trumps Oekraineaffaire	sarcastic	non-sarcastic	non-sarcastic
Waarom huisartsen soms willen weten of iemand in het verleden corona had	sarcastic	non-sarcastic	non-sarcastic
Britten hamsteren massaal na bekendmaking lockdown	sarcastic	non-sarcastic	non-sarcastic
Nieuw-Zeeland krijgt meest inclusieve parlement ooit	sarcastic	non-sarcastic	non-sarcastic
Wat zijn de gevolgen van de ergste sprinkhanenplaag in 70 jaar?	sarcastic	non-sarcastic	non-sarcastic
Wie is de omstreden Johnson-adviseur Dominic Cummings?	sarcastic	non-sarcastic	non-sarcastic
Wie is WHO-baas Tedros en waarom krijgt hij kritiek?	sarcastic	non-sarcastic	non-sarcastic
Lapt de overheid de klimaatafspraken aan haar laars?	sarcastic	non-sarcastic	non-sarcastic
Onduidelijk of man die aanslag op Pride aankondigde geradicaliseerd is	sarcastic	non-sarcastic	non-sarcastic
Dit moet je weten over de steeds grimmigere BLM-demonstraties in de VS	sarcastic	non-sarcastic	non-sarcastic
ME drijft rellende jongeren uit elkaar, ditmaal in Amersfoort	sarcastic	non-sarcastic	non-sarcastic
Deze coronaregels gelden straks in jouw favoriete vakantieland	sarcastic	non-sarcastic	non-sarcastic
Belgie raadt Belgen af vanuit provincie Antwerpen naar Nederland te reizen	sarcastic	non-sarcastic	non-sarcastic

Continued on next page

Table 5 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Slimme camera gaat automobilisten betrappen op telefoongebruik achter stuur	sarcastic	non-sarcastic	non-sarcastic
Zwarte Piet-foto kost Rotterdams architectenbureau klus in VS	sarcastic	non-sarcastic	non-sarcastic
Het gaat iets beter met de uitgehongerde leeuwen in Soedan	sarcastic	non-sarcastic	non-sarcastic
Wie is de vergiftigde Russische oppositieleider Navalny?	sarcastic	non-sarcastic	non-sarcastic
Inbrekers waren deze zomer minder actief vanwege coronamaatregelen	sarcastic	non-sarcastic	non-sarcastic
Robot vervangt mens in bediening: 'Zoiets is in Nederland best gek'	sarcastic	non-sarcastic	non-sarcastic
Oplossing stikstofuitspraak nog ver weg, politiek sterk verdeeld	sarcastic	non-sarcastic	non-sarcastic
China lanceert raket naar de maan om stenen op te halen	sarcastic	non-sarcastic	non-sarcastic
Macron: 'Samuel Paty was een van die leraren die je nooit zal vergeten'	sarcastic	non-sarcastic	non-sarcastic
EU-regeringsleiders vergaderen dinsdag verder over verdeling topposities	sarcastic	non-sarcastic	non-sarcastic
Verpleeghuis Maassluis meldt 25 coronabesmettingen	sarcastic	non-sarcastic	non-sarcastic
Minister Schouten: 'Ik kan de boeren niet alles geven wat ze willen'	sarcastic	non-sarcastic	non-sarcastic
Naar Griekenland op vakantie? Je hebt een negatieve coronatest nodig	sarcastic	non-sarcastic	non-sarcastic
Waarom koren wordt geadviseerd om voortaan in zigzagformatie te zingen	sarcastic	non-sarcastic	non-sarcastic
'VVD-stemmers verdeeld over klimaatsamenwerking'	sarcastic	non-sarcastic	non-sarcastic
Dekker krijgt vertrouwen van Kamer na OVV-rapport Anne Faber	sarcastic	non-sarcastic	non-sarcastic

Continued on next page

Table 5 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Baudet weigert 'Marokkanen'-tweet op te helderen	sarcastic	non-sarcastic	non-sarcastic
VN-rapporteur bekritiseert Máxima om gesprek met Saoedische kroonprins	sarcastic	non-sarcastic	non-sarcastic
Natuurbranden Californie breiden zich uit[1]75.000*mensenmoetenevacueren*	sarcastic	non-sarcastic	non-sarcastic
Arts-microbioloog: 'Coronasneltests gaan ons niet redden'	sarcastic	non-sarcastic	non-sarcastic
Zo werkt de nieuwe coronablaastest	sarcastic	non-sarcastic	non-sarcastic
DENK krijgt toch zetel in provincie Flevoland	sarcastic	non-sarcastic	non-sarcastic
Geen aanwijzingen coronabesmetting Nederlanders op Tenerife	sarcastic	non-sarcastic	non-sarcastic
Oud-VVD-senator dreigt verhoging studierente te blokkeren	sarcastic	non-sarcastic	non-sarcastic
Tweede Kamer wil dubbele achternaam mogelijk maken	sarcastic	non-sarcastic	non-sarcastic
Verkiezingsupdate: Trump zet z'n centen op heropening van de economie	sarcastic	non-sarcastic	non-sarcastic
Sprinkhanenplaag Afrika: 'Eet ze op of laat ze elkaar opeten'	sarcastic	non-sarcastic	non-sarcastic
Rutte: Geef thuis geen feestjes meer, ontvang maximaal zes gasten	sarcastic	non-sarcastic	non-sarcastic
Schotland eerste land ter wereld dat menstruatieproducten gratis aanbiedt	sarcastic	non-sarcastic	non-sarcastic
Hoe zit het met die coronaspoedwet?	sarcastic	non-sarcastic	non-sarcastic
Spaanse carnavalsgroep biedt excuses aan na optocht met Holocaustthema	sarcastic	non-sarcastic	non-sarcastic
Dit kun je door de versoepeling wél doen met familie, vrienden en vrije tijd	sarcastic	non-sarcastic	non-sarcastic

Continued on next page

Table 5 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Topviroloog VS vreest voor 100.000 dagelijkse nieuwe coronabesmettingen	sarcastic	non-sarcastic	non-sarcastic
Advocate toeslagenaffaire: 'Ouders hebben zich nooit kunnen verdedigen'	sarcastic	non-sarcastic	non-sarcastic
GroenLinks-Europarlementariër Judith Sargentini verlaat politiek	sarcastic	non-sarcastic	non-sarcastic
Sigrid Kaag meldt zich voor lijsttrekkerschap D66	sarcastic	non-sarcastic	non-sarcastic
Rutte: Ik besefte te laat dat Griekenlandvakantie van koning niet kon	sarcastic	non-sarcastic	non-sarcastic
Trump in videoboodschap positief over gezondheid*ingewijdenbezorgd*	sarcastic	non-sarcastic	non-sarcastic
Bedreigde docent van Rotterdamse school staat nog niet voor de klas	sarcastic	non-sarcastic	non-sarcastic
Rechtszaak Baudet en televisieprogramma Buitenhof op 11 maart	sarcastic	non-sarcastic	non-sarcastic
Signalen over mogelijke wraakacties na fatale steekpartij Scheveningen	sarcastic	non-sarcastic	non-sarcastic
Verkiezingsupdate: Tussen hoop en vrees kibbelen Democraten over knuffel	sarcastic	non-sarcastic	non-sarcastic
Biden reikt Trump-stemmer de hand in overwinningsspeech	sarcastic	non-sarcastic	non-sarcastic
Wit-Russische agenten laten schilden zakken en worden geknuffeld	sarcastic	non-sarcastic	non-sarcastic
Adopteren van huisdieren in coronatijd blijft onverminderd populair	sarcastic	non-sarcastic	non-sarcastic
Nederlandse handelsmissie naar Saoedi-Arabië definitief afgeblazen	sarcastic	non-sarcastic	non-sarcastic
Nieuwe school kan voortaan zonder religieuze of levensbeschouwelijke richting	sarcastic	non-sarcastic	non-sarcastic
Welke sneltests zijn er en wanneer kunnen we ze gebruiken?	sarcastic	non-sarcastic	non-sarcastic

Continued on next page

Table 5 – continued from previous page

Headline	SVM Prediction	BERTje Prediction	True Label
Presidentsverkiezingen Malawi moeten over vanwege gesjoemel met tipp-ex	sarcastic	non-sarcastic	non-sarcastic
Zo voorkom je dat je bril beslaat als je een mondkapje draagt	sarcastic	non-sarcastic	non-sarcastic
Trump wil nog steeds geen verlies nemen$^{n}oemtongefundeerde'fraude'weer$	sarcastic	non-sarcastic	non-sarcastic
De Jonge wil toch meer tijd voor invoering quarantaineplicht	sarcastic	non-sarcastic	non-sarcastic
Kwart van mensen die coronatest willen doen moet langer dan 48 uur wachten	sarcastic	non-sarcastic	non-sarcastic
Boeren ontvangen proces-verbaal na ode aan Bevrijdingsdag met tractoren	sarcastic	non-sarcastic	non-sarcastic
Kijk mee naar het boerenprotest bij het RIVM	sarcastic	non-sarcastic	non-sarcastic