Evaluating UK Honours Candidates using a Novel Data-Analytics Pipeline

Francesca von Braun-Bates[1*], Sunreeta Sen[2], Indraayudh Talukdar[3], Anirban Lahiri[4]

[1*]Ministry of Justice, Government of United Kingdom, London, United Kingdom.
[2]Arndit Ltd., Cambridge, United Kingdom.
[3]Indian Institute of Technology Delhi, New Delhi, India.
[4]Kainos, London,United Kingdom.

Abstract

This paper explores the application of data science to the UK Honours system, focusing on how publicly available information can support due diligence and enhance public confidence. We identify key challenges that must be addressed before such methods can be applied responsibly in this context. We present a data-processing pipeline that collects open-source information via web scraping, extracts relevant text using coreference resolution, and applies sentiment analysis to identify potentially salient evidence about candidates. We evaluate two established sentiment analysis algorithms (AFINN, VADER) and introduce a novel method, MINOS, tailored to this use case.

Our results indicate that sentence-level filtering and domain-specific sentiment modelling improve the identification of relevant positive and negative signals. The proposed system is intended to augment, rather than replace, human judgement by highlighting potentially high-impact information for further review. This approach may help reduce the risk of oversight in candidate assessment and can be extended to other awards and recognition processes.

Keywords: Sentiment Analysis, Natural Language Processing, Open-Source Intelligence, Decision Support Systems

1 Introduction

This paper considers the problem of extracting sentiment signals about individuals from noisy open-source text. Given a subject of interest (SOI), we aim to identify sentiment-bearing statements that are directly relevant to that individual and distinguish them from sentiment arising from unrelated context. This problem is difficult because publicly available corpora are heterogeneous, weakly relevant, temporally mixed, and not associated with deterministic ground-truth outcomes.

The principal contribution of this work is a data-processing pipeline combining web scraping, sentence-level relevance filtering, coreference resolution, and sentiment analysis. We evaluate two established sentiment analysis methods, AFINN and VADER, and introduce a domain-specific model, MINOS, designed to prioritise the detection of negative evidence. The central result is that sentence-level filtering and cost-sensitive treatment of negative sentiment produce more interpretable and discriminative outputs than general-purpose sentiment methods when applied to heterogeneous web data.

The application context motivating this work is the analysis of publicly available information relating to the UK Honours system, particularly cases involving forfeiture of an honour. This setting is useful because it combines several difficult characteristics relevant to entity-level sentiment analysis more generally. Relevant evidence is sparse, distributed across heterogeneous sources, and accumulates over time. Furthermore, forfeiture decisions are not governed by fixed rules, making the problem unsuitable for

conventional supervised classification approaches. The methodology developed here is therefore intended as an evidence-gathering and prioritisation system to support human evaluation, rather than as an automated decision-making framework.

A key challenge addressed in this work is that sentiment expressed within a document is not necessarily attributable to the SOI. General-purpose sentiment models frequently assign sentiment based on local lexical cues without resolving whether the sentiment refers to the individual, surrounding context, or unrelated entities. As a result, strongly negative cases may appear neutral or positive due to contamination from irrelevant context. This effect persists even after document-level filtering and motivates the use of sentence-level relevance extraction and conservative detection of negative evidence.

To evaluate the behaviour of the pipeline, we construct three proxy-labelled SOI groups: awarded individuals, infamous individuals, and forfeiture cases. These groups are not used as training labels for a supervised learning task, but instead provide structured exemplars for assessing whether the extracted sentiment signals are qualitatively consistent with known characteristics of each group. Evaluation therefore focuses on the behaviour and interpretability of the extracted sentiment distributions rather than predictive accuracy.

This work builds upon prior research in sentiment analysis, entity-level text analysis, and open-source intelligence gathering [1**?** –3]. Existing sentiment analysis methods are primarily designed for short, self-contained text such as social media posts [4, 5]. In contrast, the present work considers long-form heterogeneous corpora in which relevance and attribution are themselves major challenges. More broadly, the evaluation approach adopted here is consistent with established practices in anomaly detection and exploratory data analysis, where systems are assessed using known exemplars rather than trained to optimise classification accuracy [6, 7].

The structure of the paper is as follows. Section 2 summarises the application context and defines the SOI groups used for evaluation. Section 3 describes the proposed pipeline, including relevance filtering, coreference resolution, and sentiment aggregation. Section 4 evaluates the behaviour of the system with respect to the challenges identified in the background section and compares the baseline sentiment models with MINOS. Finally, Section 5 summarises the findings, discusses limitations, and outlines directions for future work.

2 Background

This section provides minimal domain context and defines the subjects of interest (SOIs) used to evaluate our methodology.

2.1 The Honours system and forfeiture

The UK Honours system recognises individuals for exceptional contributions across a range of fields. Awards are based on evidence of sustained achievement, public impact, and service beyond an individual's primary role. Once awarded, an Honour may be revoked if subsequent conduct is deemed to bring the system into disrepute, a process known as forfeiture.

From a data analysis perspective, this creates a classification challenge: individuals may exhibit predominantly positive, predominantly negative, or mixed public sentiment over time. In particular, forfeiture cases represent individuals whose publicly observable behaviour changed significantly after recognition.

2.2 Forfeiture as a data analysis problem

Forfeiture decisions are not governed by fixed rules or deterministic thresholds. Instead, they are made on a case-by-case basis, based on qualitative judgements about whether an individual's conduct brings the Honours system into disrepute [8, 9]. For example, forfeiture is almost certain in the following cases [8, 10, 11]:

- disbarment from a professional body in the field in which the recipient was nominated
- censure by a regulator if directly relevant to the nomination
- criminal conviction resulting in a prison sentence of three months or more
- any criminal offence under the Sexual Offences Act 2003 and related legislation

However, it is not guaranteed, and past cases do not create an automatic precedent. This lack of a well-defined decision boundary makes the problem difficult to formalise as a conventional classification task.

In addition, observed forfeiture outcomes are sparse and often occur long after the underlying behaviour. Comparing the annual number of awards to forfeitures in *e.g.* Annex B of [12] show that this is a sub-percent occurrence. Table 3 shows the lag between award and forfeiture for our SOIs can be decades. Relevant evidence is expressed across heterogeneous sources and in indirect forms, making it difficult to identify using document-level or keyword-based approaches.

The problem is further complicated by the temporal evolution of an individual's public profile. Individuals subject to forfeiture frequently exhibit both positive and negative sentiment over time, leading to mixed or non-stationary signal distributions. Finally, the asymmetric cost of errors—where missing negative evidence is more consequential than identifying spurious signals—motivates a conservative approach that prioritises the detection of potentially adverse information.

2.3 Subjects of interest

To evaluate the proposed methodology, we construct three groups of SOIs:

- **Awarded**: individuals who have received an Honour and retained it, representing consistently positive public records.
- **Infamous**: individuals associated with serious misconduct or criminal activity, representing strongly negative cases.
- **Forfeited**: individuals who were awarded an Honour and later had it revoked, representing mixed or evolving public sentiment.

Each group contains twenty individuals drawn from a range of domains and levels of public prominence. These groups are intended as a structured test set rather than a representative sample of the population.

The "infamous" group provides a negative extreme, where strong negative sentiment is expected. The "awarded" group provides a positive extreme, where sentiment should be consistently favourable. The "forfeited" group represents the primary use case, in which both positive and negative signals are present. A successful methodology should distinguish between these groups and identify cases where negative evidence is sufficiently prominent to warrant further investigation.

This framing allows us to assess whether sentiment signals extracted from open-source data can provide meaningful support for identifying high-risk or anomalous cases within a broader population.

3 Methodology

This section describes a data-processing pipeline designed to extract sentiment signals about a subject of interest (SOI) from heterogeneous open-source text. The pipeline comprises: (i) data collection, (ii) relevance filtering, (iii) sentence-level extraction with coreference resolution, (iv) sentiment analysis using baseline and domain-specific models, and (v) aggregation of results. Its overall logic is shown in Figure 1.

3.1 Data collection and processing

For each SOI, we query a search engine using the individual's full name and collect the top ∼60–70 results. From each result, we extract raw text and retain the source URL to enable traceability and subsequent human verification [13, 14].

We restrict analysis to publicly accessible content on the open web and do not access paywalled or authenticated sources. This introduces sampling bias but reflects realistic constraints faced by human evaluators and common limitations of web-scraping approaches [3, 15, 16].

All text is tokenised at the sentence level using the NLTK toolkit, as sentiment is evaluated at the level of individual statements rather than entire documents [17, 18].

3.2 Article relevance criteria

The extracted text contains a substantial proportion of irrelevant or weakly related content. We therefore apply two simple filters as proxies for relevance:

1. File size ≥ 500 characters;
2. The SOI's name appears at least once.

The file size threshold removes trivial or non-content pages (e.g. navigation or login prompts), while the name filter ensures a minimal level of topical relevance. Although naming conventions introduce edge

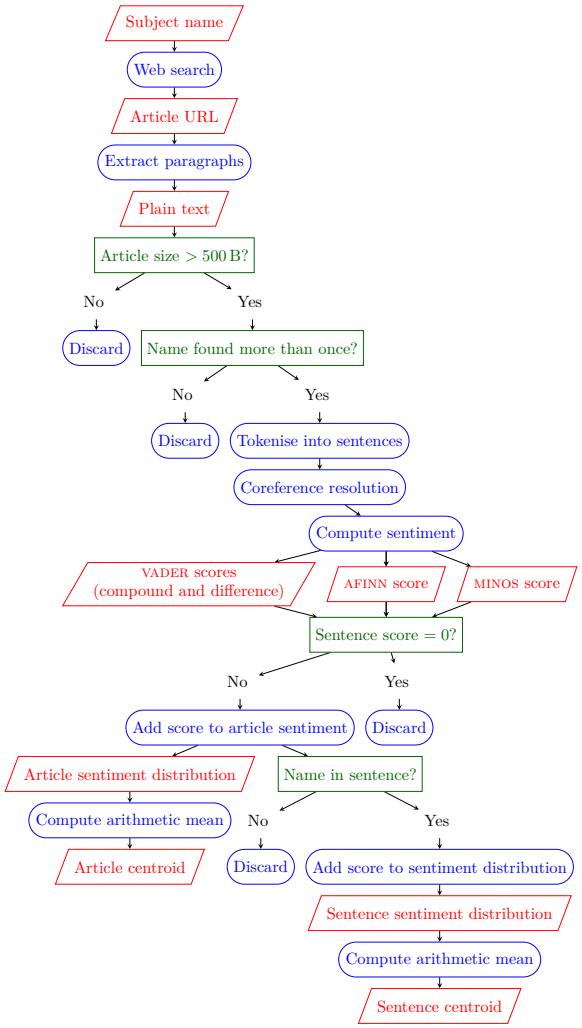


Figure 1: Summary of pipeline methodology. Trapezia represent outputs (except the starting input), rectangles are binary decisions and rounded rectangles are outcomes

cases (e.g. titles, multiple surnames, or alternative forms), a simple heuristic based on commonly used names was found to be sufficient for filtering purposes.

3.3 Coreference resolution

Document-level filtering is insufficient to guarantee that sentiment-bearing text refers to the SOI. Even within relevant documents, many sentences describe other entities or general context. To address this, we apply coreference resolution to identify expressions that refer to the SOI and restrict analysis to those sentences [19, 20].

Coreference resolution replaces pronouns and other referring expressions with their corresponding named entities, allowing indirect references to the SOI to be captured. This reduces false negatives (relevant sentences excluded) and improves the precision of sentiment extraction, though the process is inherently probabilistic and may introduce minor errors.

3.4 Sentiment analysis

We evaluate two established sentiment analysis methods alongside a domain-specific model.

AFINN is a lexicon-based approach that assigns integer sentiment scores to individual words [4, 21, 22]. VADER extends this approach with rules that account for negation, emphasis, and punctuation, producing both component and compound sentiment scores [5, 23]. Both methods are designed primarily for short, informal text and can perform poorly on heterogeneous, long-form sources.

To address this limitation, we introduce MINOS, a domain-specific sentiment model designed to prioritise the detection of negative signals. MINOS uses separate positive and negative lexicons derived

from multiple sources, including domain-relevant terminology (e.g. criminal activity and honours-related terms). Words not present in either lexicon are treated as neutral.

The key design principle of MINOS is that negative evidence dominates: if a sentence contains any negative term, the overall sentence sentiment is treated as negative. This reflects the asymmetric cost of errors in the application context, where missing negative signals is more consequential than identifying spurious positives.

3.5 Aggregation of sentiment

Each sentence is assigned a sentiment score based on the selected model. We summarise sentiment for an SOI by computing the arithmetic mean of non-zero sentence scores, excluding neutral statements that do not contribute meaningful signal.

Sentences are treated as the fundamental unit of analysis—rather than words, paragraphs or articles. For a given SOI, the aggregate sentiment is therefore obtained by integrating over the distribution of sentence-level scores across all "relevant text" (where the level of relevance is controlled by the filtering). The resulting value represents the centroid of this distribution.

This sentence-level aggregation is necessary because article-level scoring is not equivalent across models. For lexicon-based approaches such as AFINN and VADER, sentiment can be accumulated across words or sentences. However, for MINOS, which assigns negative sentiment whenever any negative term is present, aggregating at the article level would collapse mixed-sentiment articles into purely negative scores. This would obscure the presence of both positive and negative evidence within the same document.

The full distribution of sentence-level scores therefore provides additional information beyond the centroid, particularly in cases where sentiment is mixed. In the results, we analyse both the centroid and the distribution to characterise differences between SOI groups.

The sentiment models considered in this work operate on different scales and scoring schemes. As a result, their outputs are not directly comparable in absolute terms. Instead, comparisons are made within each model by examining the relative separation between SOI groups and the structure of their sentiment distributions. This allows us to assess how effectively each method distinguishes between positive, negative, and mixed sentiment profiles without requiring cross-model normalisation.

3.6 Output for human analysis

The objective of the pipeline is to support human evaluation by identifying potentially relevant positive and negative evidence within large text corpora. In addition to aggregate sentiment measures, the system retains the underlying sentences and their source URLs, allowing analysts to inspect high-impact evidence directly and assess its reliability and context.

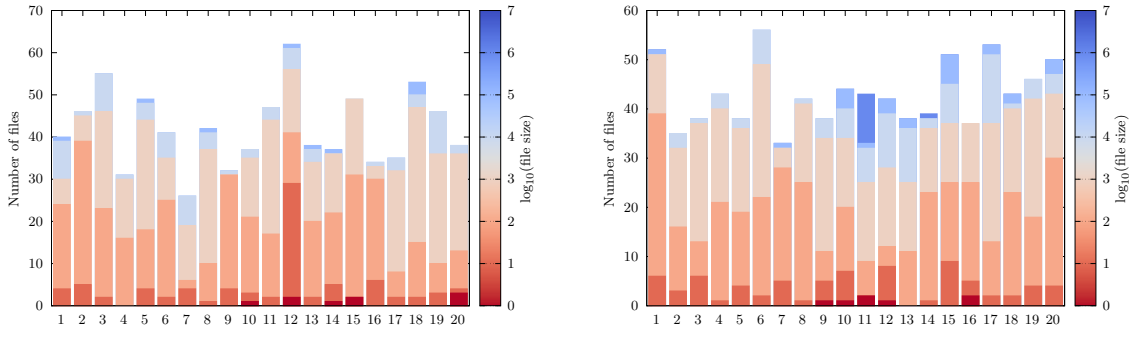
4 Results, Analysis and Discussion

This section evaluates the proposed methodology with respect to the challenges identified in Section 2. In particular, we assess the extent to which the pipeline is able to extract meaningful sentiment signals from noisy, heterogeneous open-source data and distinguish between different categories of subjects of interest.

Recall that analysis is structured around three SOI groups: awarded, infamous, and forfeited. These groups represent positive, negative, and mixed sentiment profiles respectively, providing a basis for evaluating how well the methodology captures different types of behavioural signals.

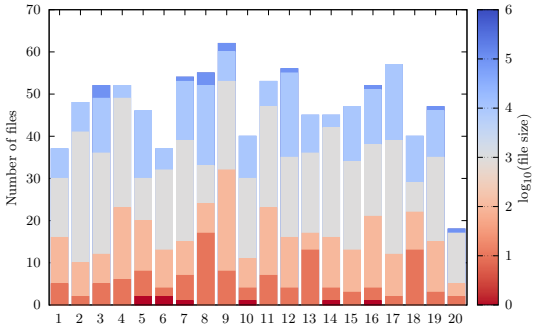
The SOI groupings provide proxy-labelled test cases for evaluation; the objective is not to train a model to classify individuals, but to assess whether the extracted sentiment signals are consistent with known characteristics of each group.

We evaluate the methodology against the principal challenges identified in Section 2. The analysis proceeds from data quality to model behaviour. Section 4.1 shows that relevance filtering substantially reduces noise in heterogeneous open-source corpora. Section 4.2 demonstrates that sentence-level filtering and coreference resolution improve the extraction of SOI-relevant text. Having established the importance of relevance extraction, we compare the behaviour of different sentiment models, first examining cases where they produce consistent qualitative conclusions in Section 4.3, then investigating situations in which their differing assumptions lead to divergent interpretations. Finally, we consider mixed sentiment profiles and discuss the appropriate interpretation of model outputs in a human-centred decision-support context.



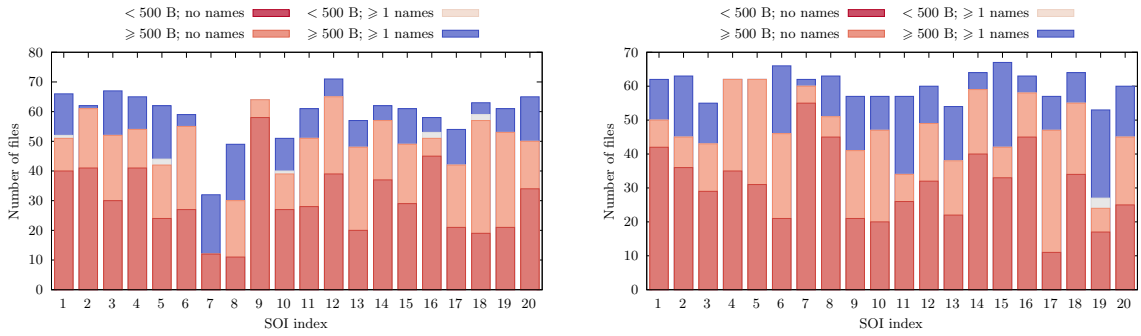
(a) Awarded

(b) Forfeited



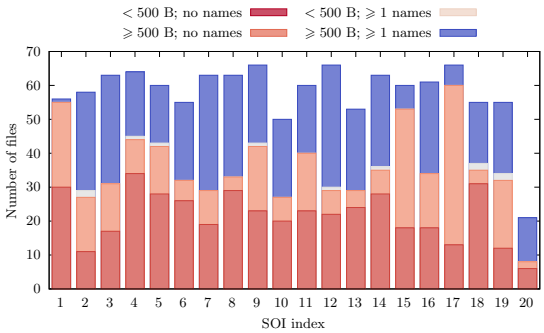
(c) Infamous

Figure 2: File size distribution per SOI group. The colour scale represents the (logarithmically-binned) file size in bytes, with the height of each stack is the number of articles per SOI.



(a) Awarded

(b) Forfeited



(c) Infamous

Figure 3: File size distribution per SOI group. The colour scale represents the (logarithmically-binned) file size in bytes, with the height of each stack is the number of articles per SOI.

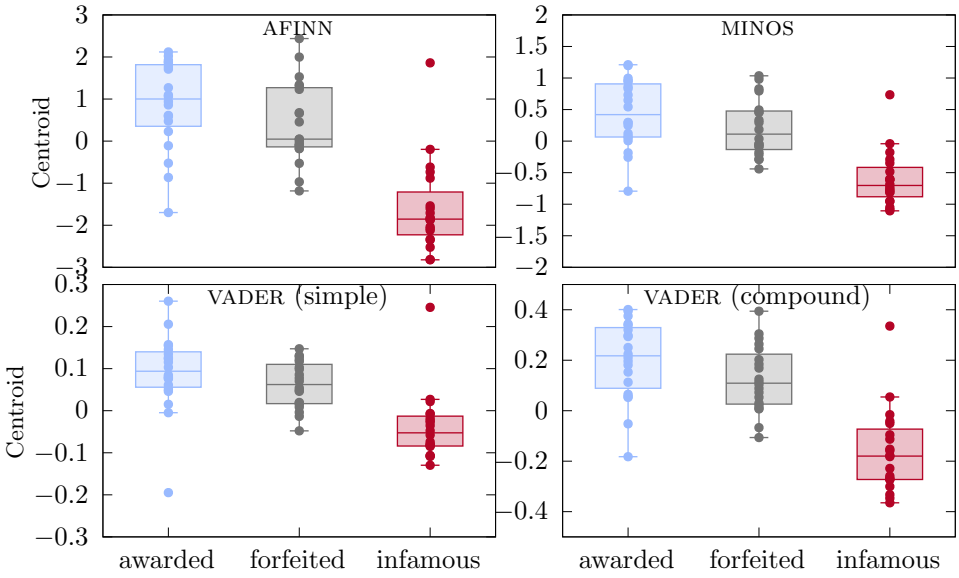


Figure 4: Box plot of centroid distributions from all raw data.

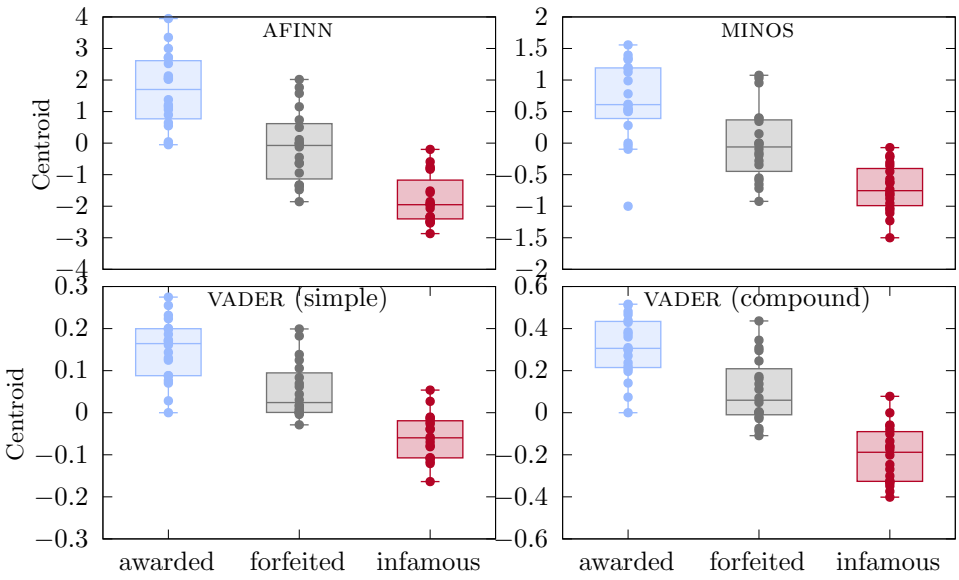


Figure 5: Box plot of centroid distributions from raw data, including only articles which contain the SOI name.

4.1 Amplifying weak signals via article-level criteria

A primary challenge identified in Section 2 is the heterogeneous and weakly relevant nature of open-source text. Web search results contain large volumes of content that are either unrelated to the SOI or contain sentiment that does not refer to the individual. This subsection evaluates the effectiveness of the proposed filtering steps in reducing this noise.

Figure 2 shows the distribution of file sizes for retrieved articles across all SOIs. The results indicate substantial variation in article length, spanning several orders of magnitude. A significant proportion of files are either too short to contain meaningful information (e.g. navigation pages or login prompts) or sufficiently large that they are unlikely to be focused on a single individual.

Applying a file size threshold removes a large number of trivial or spurious results. However, file size alone is insufficient to ensure relevance. Figure 3 shows the proportion of articles that pass the file size and name-based filters. In all SOI groups, a substantial fraction of retrieved content fails at least one of these criteria, demonstrating the extent of noise in the raw data.

Name-based filtering further reduces irrelevant content by requiring explicit mention of the SOI. However, even after these filters are applied, many retained articles contain sentences that are not directly

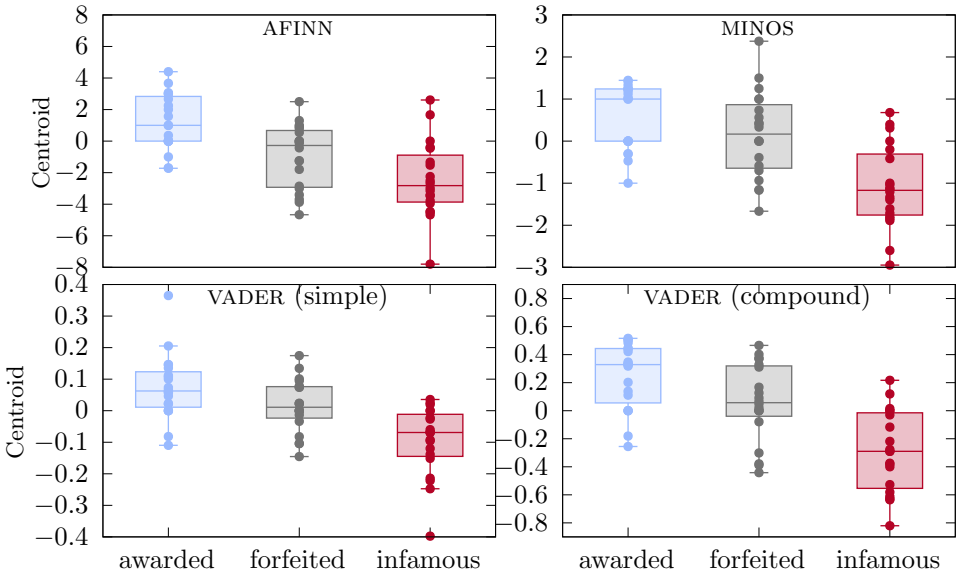


Figure 6: Box plot of centroid distributions from raw data, including only sentences which contain the SOI name.

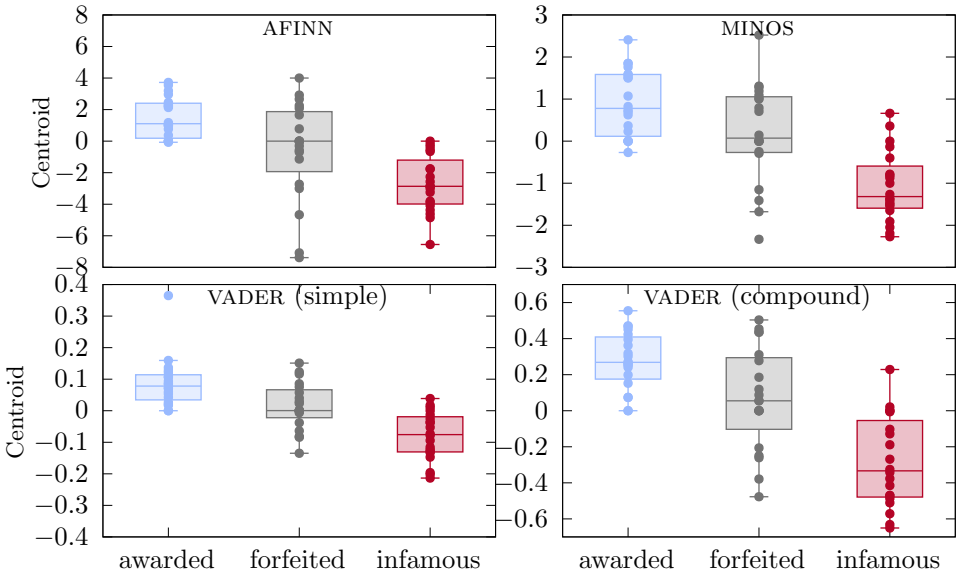


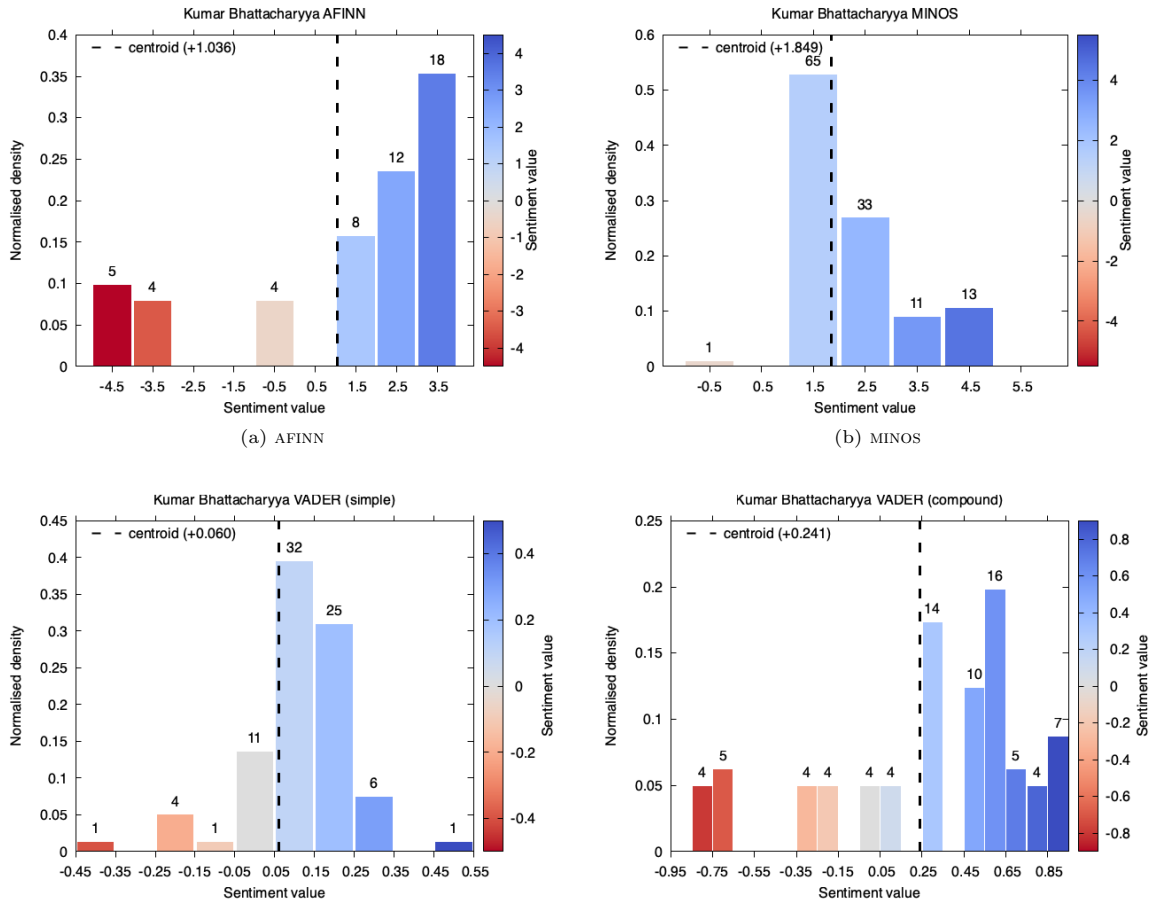
Figure 7: Box plot of centroid distributions from coreference-resolved data, including only sentences which contain the SOI name.

related to the individual. This motivates the use of sentence-level filtering and coreference resolution, which are evaluated in subsequent sections.

Overall, these results demonstrate that relevance filtering is a necessary preprocessing step for extracting meaningful sentiment signals. Without such filtering, sentiment analysis would be dominated by noise arising from unrelated or weakly relevant text.

4.2 Maximising signal relevance at the sentence-level with coreference resolution

The article-level filtering described in Section 3.2 removes a substantial fraction of irrelevant content. However, article-level relevance does not guarantee that sentiment-bearing text within an article refers to the SOI. Many retained articles contain descriptions of events, organisations, locations, or other individuals whose sentiment may contaminate the overall signal.



(a) AFINN (b) MINOS

(c) VADER (simple) (d) VADER (compound)

Figure 8: Sentence-level sentiment distributions for Lord Bhattacharyya.

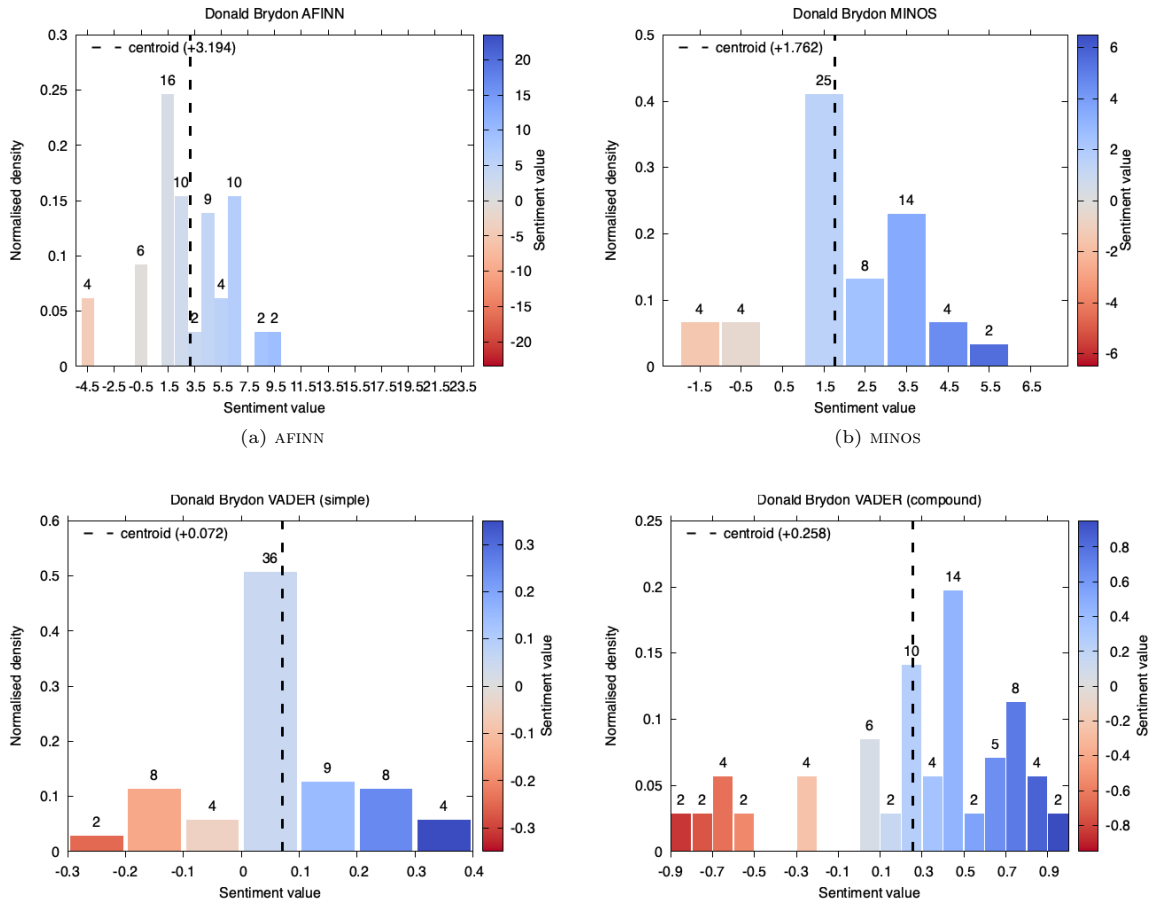
To assess the impact of increasingly selective relevance extraction, we compare centroid distributions obtained from four stages of the pipeline: raw data, article-level filtering, sentence-level filtering, and sentence-level filtering with coreference resolution.

Figure 4 shows the centroid distributions obtained from all retrieved text without any filtering. The three SOI groups exhibit substantial overlap. In particular, the infamous group, which should be strongly negative, frequently appears close to neutral. This indicates that sentiment from unrelated or weakly relevant text dominates the aggregate signal.

Applying article-level filtering improves the separation between groups. Figure 5 shows the corresponding distributions after filtering by file size and name occurrence. As suggested by Figure 3, much of the noise has been removed safely, by excluding articles which do no mentioned the SOI substantively. Although the overlap is reduced, the centroid distributions remain broad and poorly separated. This demonstrates that simply identifying articles which mention the SOI is insufficient to isolate sentiment directly relevant to that individual.

A larger improvement is obtained by restricting the analysis to sentences that explicitly contain the SOI name. Figure 6 shows that sentence-level filtering substantially improves separation between the awarded, forfeited, and infamous groups. This confirms that much of the noise in the corpus arises from sentiment-bearing text that is not directly associated with the SOI. However, unlike the article-level filtering, this also throws away false negatives because sentences which are about the SOI frequently employ anaphora such as pronouns in place of the SOI's name. This is clearly an unsuitable filter because it removes signal as well as noise.

By applying coreference resolution before sentence-level filtering, the risk of discarding false negatives is greatly reduced. Figure 7 shows the results of coreference-resolved, sentence-level filtering including text which would have been discarded in Figure 6. Relative to simple sentence filtering, coreference resolution further improves separation between the groups and increases the amount of relevant text available for analysis. This occurs because references expressed through pronouns and other indirect forms are recovered, reducing the number of relevant sentences excluded from the analysis.



(a) AFINN (b) MINOS

(c) VADER (simple) (d) VADER (compound)

Figure 9: Sentence-level sentiment distributions for Sir Donald Brydon.

Overall, these results demonstrate that a substantial proportion of the improvement in sentiment discrimination arises from increasingly precise extraction of SOI-relevant text. Before considering the choice of sentiment model, relevance extraction alone produces progressively clearer separation between the three groups.

4.3 Robust sentiment signals: when models agree

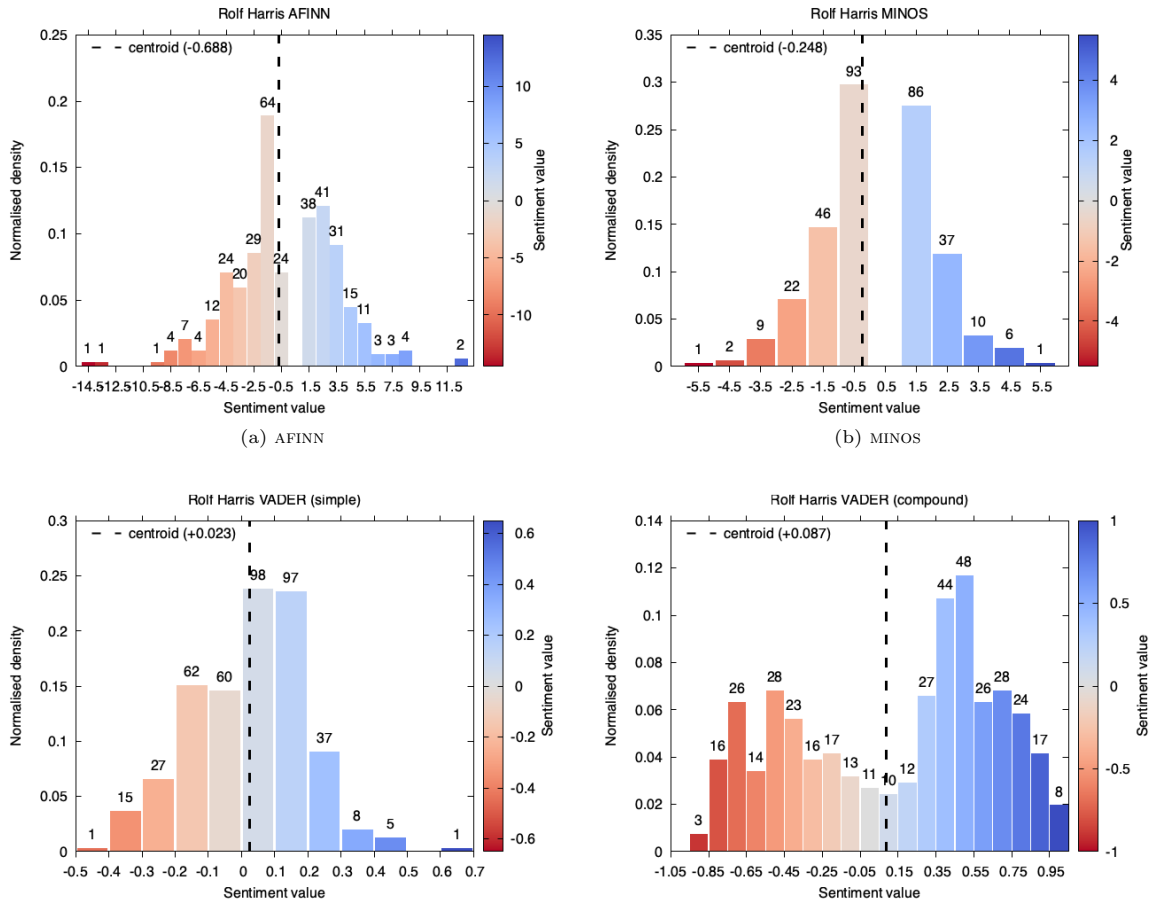
The preceding sections show that relevance extraction substantially improves the quality of the text passed to the sentiment models. We now examine the behaviour of the sentiment models themselves. In particular, we first consider cases where AFINN, VADER, and MINOS agree in the sign of the centroid, despite producing different distributional shapes.

This comparison is useful because the three models encode different assumptions. They differ in their wordlists, the scores assigned to sentiment-bearing terms, and the way sentence-level scores are aggregated. Agreement between models should therefore not be interpreted as proof that the inferred sentiment is "true". However, it does provide evidence that the qualitative interpretation is robust to the modelling assumptions embedded in a particular sentiment algorithm.

We begin with two examples from the awarded group. Figure 8 shows the sentence-level sentiment distributions for Lord Bhattacharyya, while Figure 9 shows the corresponding distributions for Sir Donald Brydon. In both cases, all three models return positive centroid values.

PLACEHOLDER: Insert detailed interpretation of the Bhattacharyya and Brydon figures. Note the differences in distribution shape between afinn, vader, and minos, and explain why the shared positive centroid supports a robust positive interpretation.

We then consider two examples from the infamous group. Figure 12 shows the sentiment distributions for Charles Sobhraj, while Figure 11 shows the distributions for Oscar Pistorius. These cases differ in their public profiles and the likely mixture of reporting within their corpora, but all three models return negative centroid values.



(a) AFINN (b) MINOS

(c) VADER (simple) (d) VADER (compound)

Figure 10: Sentence-level sentiment distributions for Rolf Harris.

PLACEHOLDER: Insert detailed interpretation of the Sobhraj and Pistorius figures. Highlight whether the distributions differ in spread, skew, or concentration around zero, while noting that the centroid sign remains negative across all models.

Finally, we include one forfeiture example. Figure 10 shows the corresponding sentiment distributions for Rolf Harris. This case is important because forfeiture examples are expected to contain both positive and negative evidence: positive reporting associated with the original award and negative reporting associated with later misconduct. Where the models agree despite this mixed evidence, the result provides a useful bridge between clear positive or negative cases and the more complex mixed profiles analysed later.

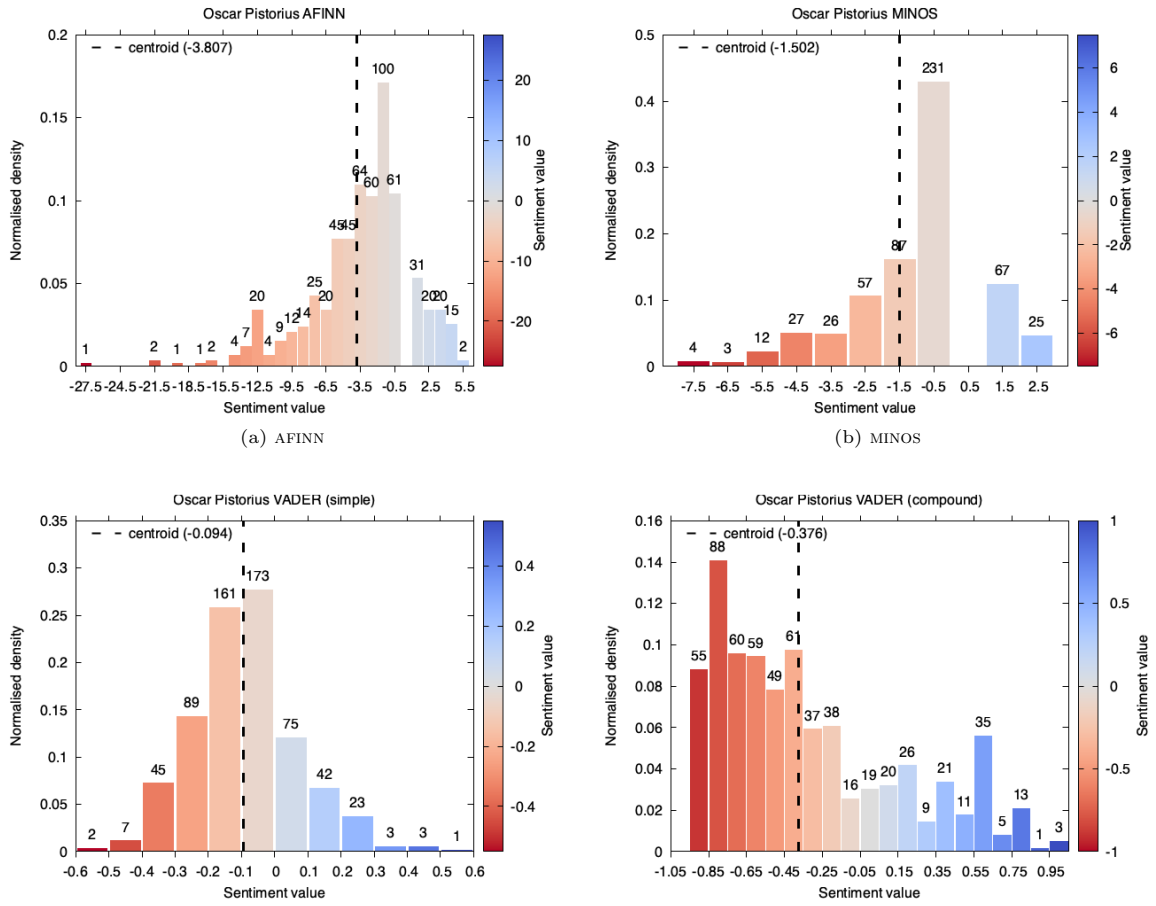
PLACEHOLDER: Insert detailed interpretation of the Harris figure. Explain whether agreement is weaker or stronger than in the awarded and infamous examples, and whether the distribution already shows the mixed-sentiment behaviour discussed in ??.

Taken together, these examples show that model agreement can occur across awarded, infamous, and forfeited SOIs. This is important for two reasons. First, it shows that MINOS is not simply forcing all cases into the expected category: in clear cases, its qualitative output agrees with the baseline models. Second, it shows that differences in distributional shape do not necessarily invalidate centroid-based summaries when the dominant signal is strong enough to survive aggregation.

The situations of greatest practical interest are therefore not those in which all models agree, but those in which the choice of sentiment model materially affects interpretation. These cases reveal the trade-offs between general-purpose sentiment estimation and conservative detection of negative evidence. We examine this distinction in the following subsection.

4.4 Limitations of off-the-shelf sentiment models and negative recall risk

The previous subsection demonstrated that sentence-level filtering and coreference resolution substantially improve the relevance of the extracted text. However, even when analysis is restricted to text that refers



(a) AFINN (b) MINOS

(c) VADER (simple) (d) VADER (compound)

Figure 11: Sentence-level sentiment distributions for Oscar Pistorius.

to the SOI, standard sentiment analysis models still face an important limitation: they do not determine whether sentiment-bearing language is actually attributable to the individual.

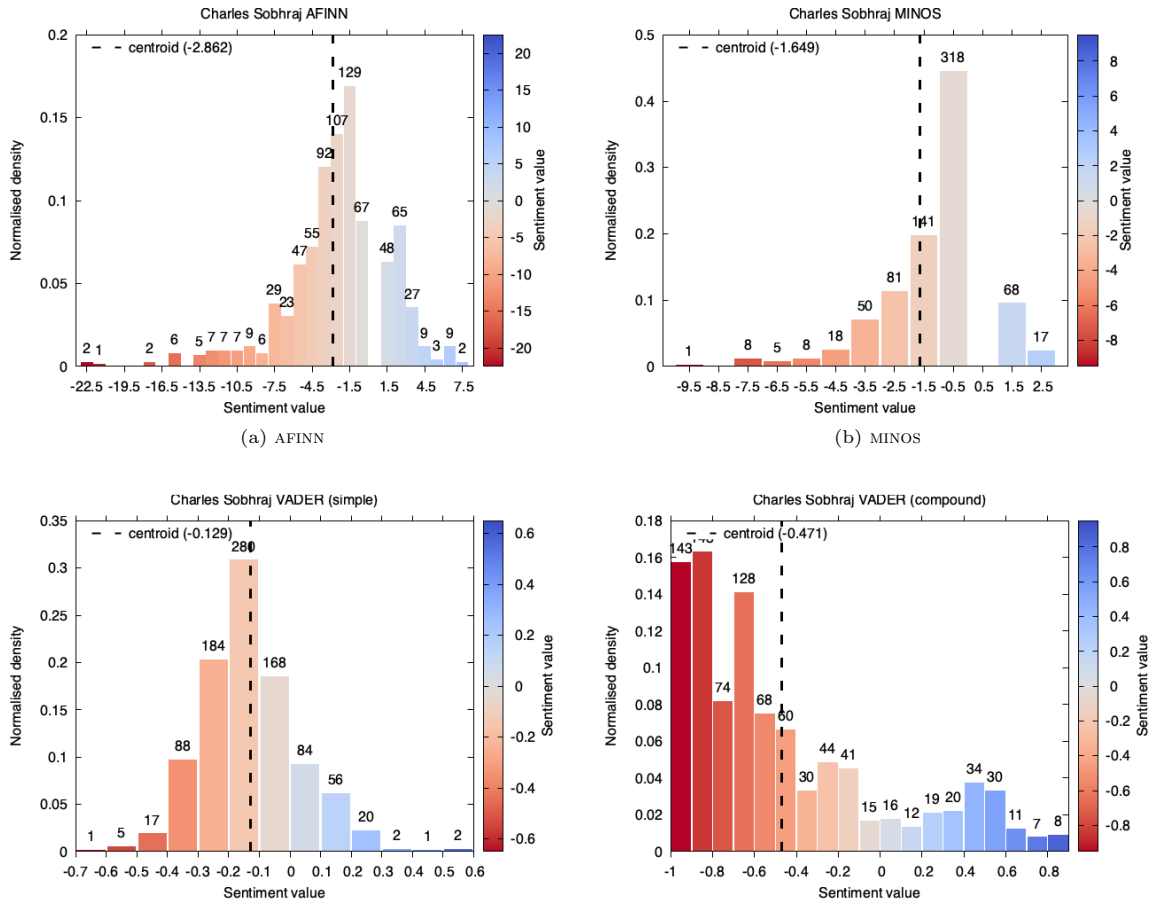
This effect arises because baseline sentiment models operate on local lexical cues without resolving which entity a sentiment-bearing expression refers to. While coreference resolution allows us to identify sentences that mention the SOI, it does not determine whether the sentiment expressed in those sentences is attributable to the individual.

As a result, sentiment associated with surrounding context—such as events, topics, or other entities—can be incorrectly attributed to the SOI. For example, Dame Mary Beard's sentiment scores are negatively affected by "death" appearing in the title of her work *Pompeii: Life and Death in a Roman Town*, even though it does not describe her character.

This limitation leads to systematic misattribution of sentiment, particularly in heterogeneous text where descriptive or contextual language is common. Although filtering and coreference resolution substantially reduce noise, they do not eliminate the influence of irrelevant sentiment. This motivates the need for a domain-specific sentiment model that explicitly accounts for the asymmetric importance of negative signals, which we address in the following subsection.

In our particular application to the UK Honours System, risks are asymmetric: failing to identify relevant negative evidence (false negatives) is more consequential than incorrectly flagging spurious signals (false positives). The objective is therefore not to produce an unbiased estimate of sentiment, but to maximise the likelihood of identifying potentially adverse evidence for subsequent human review.

This objective manifests differently across SOI groups. For individuals in the "awarded" group, the goal is to maximise sensitivity to potential forfeiture signals. Even a small number of negative statements may be important, and should be surfaced for triage by a human analyst. Missing such signals represents a failure of the system. Conversely, for individuals in the "infamous" group, we expect overwhelmingly negative sentiment. In this case, the presence of positive or neutral sentiment in the output indicates contamination from irrelevant context, which obscures the true signal.



(a) AFINN (b) MINOS

(c) VADER (simple) (d) VADER (compound)

Figure 12: Sentence-level sentiment distributions for Oscar Pistorius.

We first consider a representative case from the "awarded" group. **fig:marks-afinn (placeholder)**, **fig:marks-vader (placeholder)**, and **fig:marks-minos (placeholder)** show the distribution of sentence-level sentiment scores for the same SOI under AFINN, VADER, and MINOS respectively. The baseline models produce a broad distribution with both positive and negative values. While this reflects the presence of diverse language in the corpus, it also indicates that negative signals are diluted by unrelated positive context. As a result, potentially relevant negative evidence is not clearly distinguished from background noise.

We now consider a corresponding example from the "infamous" group. **fig:kasab-afinn (placeholder)**, **fig:kasab-vader (placeholder)**, and **fig:kasab-minos (placeholder)** show the equivalent distributions for this SOI. Despite the strongly negative nature of the underlying behaviour, both AFINN and VADER assign a substantial number of positive or near-neutral scores. This uncovers an axiomatic principle of both AFINN and VADER: that they were designed to estimate overall sentiment without privileging positive or negative evidence. As a result, positive language in surrounding context—such as descriptions of institutions or actions of other individuals—can offset genuinely negative signals.

The contrast between the two cases highlights a fundamental limitation of baseline sentiment models: their design objective is to provide an unbiased estimate of sentiment, rather than to prioritise the detection of specific types of evidence. In this application, such neutrality is undesirable, as it increases the likelihood of false negatives by allowing negative signals to be masked by unrelated positive context.

The MINOS model addresses this limitation by explicitly prioritising negative evidence. For both SOIs, MINOS assigns negative sentiment whenever a sentence contains a negative term, regardless of the presence of positive language. This increases the likelihood that potentially relevant negative statements are surfaced. In the "awarded" example, this results in a small number of clearly identifiable negative signals, suitable for human review. In the "infamous" example, it produces a distribution that is overwhelmingly negative, reducing the influence of irrelevant positive context.

Taken together, these results demonstrate that prioritising the detection of negative evidence provides a more appropriate representation of risk in this setting. While this approach may increase the number

of sentences flagged for review, it reduces the likelihood that important signals are missed, aligning the behaviour of the model with the requirements of the application.

4.5 Mixed sentiment profiles and temporal behaviour

A further challenge identified in **sec:forfeiture-challenges (placeholder)** is that an individual's public profile may evolve over time, resulting in mixed or non-stationary sentiment signals. This is particularly relevant for forfeiture cases, where individuals may have a substantial body of positive reporting prior to the events that led to the withdrawal of an honour. In such cases, sentiment cannot be characterised by a single polarity, and aggregate measures must be interpreted with care.

We therefore examine the distribution of sentence-level sentiment scores for individuals in the "forfeited" group. **fig:vasco-knight (placeholder)** and **fig:rodale (placeholder)** show representative examples under AFINN, VADER, and MINOS. In both cases, the distributions exhibit a mixture of positive and negative scores, reflecting the coexistence of positive reporting associated with earlier achievements and negative reporting associated with subsequent events.

This behaviour contrasts with the more homogeneous distributions observed in the "awarded" and "infamous" groups. For example, **fig:hawking (placeholder)** shows a predominantly positive distribution, while **fig:epstein (placeholder)** shows a predominantly negative distribution. These cases illustrate that, in the absence of temporal shifts in behaviour, sentiment signals tend to be concentrated around a single polarity.

The limitations of centroid-based summaries become particularly apparent when comparing individuals with similar aggregate values. **fig:forfeited-centroid-comparison (placeholder)** shows examples of SOIs with comparable centroid values but markedly different underlying distributions. In such cases, the centroid obscures the presence of both strongly positive and strongly negative evidence, and therefore fails to capture the qualitative differences between individuals.

Examining the full distribution of sentence-level scores provides a more informative representation. In particular, the presence of distinct positive and negative components allows forfeited individuals to be distinguished from both extremes. This supports the use of distributional characteristics, such as spread or multi-modality, in addition to aggregate measures when identifying potentially anomalous cases.

An additional consideration is the cumulative nature of the underlying data. The corpus for each SOI is constructed from publicly available text collected at a single point in time, but reflects information that has accumulated over the individual's public history. New information is therefore added to an existing body of evidence rather than replacing it. In principle, repeated application of the pipeline as new data becomes available would lead to convergence of both the centroid and the distribution, unless the new information represents a significant deviation from prior behaviour.

In this sense, the methodology captures behaviour integrated over time, partially mitigating the effect of delays in the emergence of negative evidence. While forfeiture decisions may occur after a lag, the underlying signals can be incorporated into the analysis as soon as they appear in publicly available sources. This highlights a potential advantage of automated approaches over manual review, which may be limited in its ability to aggregate historical evidence at scale.

Overall, these results demonstrate that forfeiture cases are characterised by mixed sentiment profiles that cannot be adequately captured by aggregate measures alone. A distributional approach is therefore necessary to identify individuals whose public record contains both positive and negative signals, and to distinguish these cases from those with consistently positive or consistently negative sentiment.

4.6 Interpreting outputs under limited and non-deterministic ground truth

A central challenge identified in **sec:forfeiture-challenges (placeholder)** is that forfeiture outcomes are not governed by a fixed or observable decision rule. Instead, they arise from case-by-case judgements based on qualitative assessment of an individual's conduct. In addition, the number of publicly documented forfeiture cases is limited relative to the total number of honours awarded. Together, these factors mean that there is no well-defined ground-truth label that can be used to train or evaluate a conventional supervised classification model.

The SOI groupings used in this work therefore serve as proxy-labelled test cases for evaluation rather than as training data. The objective is not to assign individuals to these categories with high accuracy, but to assess whether the sentiment signals extracted by the pipeline are consistent with known characteristics of each group. Evaluation is therefore comparative and behavioural, focusing on the separation of group-level patterns and the interpretability of sentence-level evidence. This evaluation approach is consistent with established practices in anomaly detection and exploratory data analysis, where systems are assessed using known exemplars rather than trained to optimise classification accuracy [6, 7]. Similar

patterns are observed in intelligence analysis, where accumulated evidence and representative cases guide interpretation in the absence of deterministic decision rules [24].

The outputs of the pipeline should be interpreted as descriptive rather than predictive. Aggregate measures, such as the centroid of sentence-level sentiment scores, provide a summary of the balance of positive and negative signals, while the full distribution captures the presence of mixed or extreme values. As discussed in **sec:aggregation (placeholder)** and **sec:mixed-sentiment (placeholder)**, these representations highlight different aspects of the underlying data and should be considered jointly.

Across the SOI groups, the results exhibit consistent qualitative patterns. Individuals in the "awarded" group tend to show predominantly positive sentiment, while those in the "infamous" group show predominantly negative sentiment. The "forfeited" group occupies an intermediate position, characterised by mixed sentiment distributions and overlapping centroid values. This overlap reflects the heterogeneous nature of the underlying evidence rather than an absence of signal, and reinforces the limitations of interpreting aggregate measures in isolation.

In this context, the methodology is best understood as a tool for ranking or flagging cases based on the presence of potentially relevant evidence. In particular, the identification of negative sentence-level signals provides a mechanism for surfacing information that may warrant further investigation. As discussed in **sec:minos (placeholder)**, the approach explicitly incorporates asymmetric error costs, prioritising the detection of negative evidence over the avoidance of false positives.

Human evaluation is therefore not an optional safeguard but a fundamental requirement. The system is designed as an evidence-gathering and prioritisation pipeline, in which automated methods structure large volumes of unstructured text and highlight potentially important signals, while the interpretation of those signals—including their relevance, reliability, and context—remains the responsibility of the human analyst.

This design is consistent with established principles for responsible AI systems [25], which emphasise interpretability, reliability, and the retention of human oversight in settings where decisions are context-dependent and cannot be reduced to deterministic rules. In this sense, the methodology does not attempt to replace human judgement, but to augment it by enabling more efficient and systematic analysis of publicly available information at scale.

Overall, the results demonstrate that meaningful sentiment signals can be extracted from noisy open-source data, but that these signals must be interpreted with care. The combination of sentence-level analysis, distributional representation, and conservative detection of negative evidence provides a practical framework for supporting decision-making in settings characterised by limited ground truth and non-deterministic outcomes.

5 Conclusions and Future Work

This paper investigated the problem of extracting sentiment signals about individuals from heterogeneous open-source text. Motivated by the challenge of identifying potentially relevant evidence within large volumes of publicly available information, we developed a data-processing pipeline combining web scraping, relevance filtering, coreference resolution, and sentiment analysis.

The work addressed several key difficulties inherent to this setting. First, open-source corpora contain large amounts of irrelevant or weakly related text, requiring filtering at both document and sentence level. Second, sentiment-bearing expressions are not necessarily attributable to the subject of interest, limiting the effectiveness of standard sentiment analysis methods. Third, forfeiture-related evidence is temporally mixed and accumulates over time, producing sentiment distributions that cannot be adequately represented by a single aggregate measure. Finally, the application context introduces asymmetric error costs, in which failing to identify potentially relevant negative evidence is more consequential than flagging spurious signals for review.

The results demonstrated that sentence-level filtering and coreference resolution substantially reduce noise in the extracted sentiment signal. Comparison of AFINN and VADER with the proposed MINOS model showed that general-purpose sentiment methods frequently assign positive or neutral sentiment due to contextual language unrelated to the behaviour of the SOI. In contrast, MINOS prioritises the detection of negative evidence, producing clearer separation between strongly negative and non-negative cases.

The methodology was evaluated using proxy-labelled groups of awarded, infamous, and forfeited individuals. These groups were not used as training data for a supervised classification task, but instead provided structured exemplars for evaluating whether the extracted sentiment signals were qualitatively consistent with known characteristics of each group. The results suggest that distributional representations of sentence-level sentiment contain more informative structure than centroid values alone, particularly for forfeiture cases exhibiting mixed sentiment profiles.

A key feature of the proposed approach is that it is designed as an evidence-gathering and prioritisation pipeline rather than a decision-making system. Human evaluation remains an essential component of the process, with the methodology intended to support analysts by surfacing potentially relevant evidence from large unstructured corpora. In this sense, the work aligns with broader principles for responsible AI systems, emphasising interpretability, traceability, and the retention of human oversight in contexts where outcomes are not governed by deterministic rules [25].

Several directions for future work follow naturally from these results. The current methodology evaluates accumulated evidence at a single point in time; extending the pipeline to operate continuously on streaming data would enable temporal analysis of how sentiment distributions evolve as new information emerges. In addition, distributional features such as variance, skewness, or multi-modality may provide more informative repre analysissentations of behavioural change than aggregate sentiment measures alone.

More sophisticated approaches to entity-level sentiment attribution could also reduce the misassignment of sentiment arising from contextual language. While coreference resolution allows us to identify sentences that mention the SOI, it does not determine whether the sentiment expressed in those sentences is attributable to the individual. As a result, sentiment associated with surrounding context—such as events, topics, or other entities—can be incorrectly attributed to the SOI. For example, Dame Mary Beard's sentiment scores are negatively affected by "death" appearing in the title of her work *Pompeii: Life and Death in a Roman Town*, even though it does not describe her character. Overcoming this limitation, particularly in heterogeneous text where descriptive or contextual language is common, is a key refinement of the current approach.

Finally, although this work was motivated by a specific application domain, the broader methodology is applicable to other settings involving entity-centric analysis of noisy open-source data, particularly where the objective is to support human interpretation under limited or non-deterministic ground truth.

A Individuals or Subjects of Interest

Table 1: The "awarded" group who received an Honours and maintained their award.

SOI	Name	Award	Citation	Gender	Year
A.1	Professor Dame Winifred Mary Beard OBE	DBE	Study of Classical Civilisation	F	2018
A.2	Professor Sir Sushantha Bhattacharyya CBE	KBE	Higher Education and Industry	M	2003
A.3	Sir Donald Brydon CBE	KBE	Business and charity	M	2018
A.4	Allan Cook	CBE	Defence and Aerospace Industries	M	2018
A.5	Hugh David Facey MBE	OBE	Manufacturing, Innovation, Exports and Employee Ownership	M	2018
A.6	Rear Admiral Philip Duncan Greenish	CBE	Military Division	M	2002
A.7	Professor Carole Hillenbrand OBE	CBE	Understanding of Islamic History	F	2018
A.8	Dr Mohammed Kamal Hossain	OBE	Industry	M	2009
A.9	Professor Sir James Hough OBE FRS FRSE	KBE	Detection of Gravitational Waves	M	2018
A.10	Sandra Kerr OBE	CBE	Equality and to Diversity	F	2019
A.11	John Nigel Kirkland	OBE	Derbyshire	M	1999
A.12	Professor Richard Ian Kitney	OBE	IT in Health Care	M	2001
A.13	Ursula Frances Rosamond Lidbetter	MBE	Business in Lincolnshire	F	2011
A.14	Professor John Neil Loughhead OBE	CB	Research and Development in the Energy Sector	M	2018
A.15	Miss Maria McCaffery MBE	OBE	Renewable Energy Sector	F	2017
A.16	Professor Carol Propper CBE FBA	DBE	Economic Policy and Public Health	F	2020
A.17	Dr. Frances Carolyn Saunders CB	DBE	Science and Engineering	F	2018
A.18	Ms Jennifer Margaret Saunders OBE	CBE	Tackling Fuel Poverty	F	2018
A.19	Jack Crossley Tordoff MBE	OBE	Business and West Yorkshire	M	2018

Table 2: The "infamous" group of individuals who committed serious crime.

SOI	Name	Criminal activity	Gender[1]
I.1	Ajmal Kasav	Terrorism and extremism	M
I.2	Bruce Reynolds	Theft, assualt, drug dealing	M
I.3	Charles Sobhraj	Murder, attempted murder	M
I.4	Dale Cregan	Murder	M
I.5	Dennis Nilsen	Murder, attempted murder	M
I.6	Ghislaine Maxwell	Sex trafficking	F
I.7	Harold Shipman	Murder	M
I.8	Harvey Weinstein	Sexual offences	M
I.9	Howard Marks	Drug dealing	M
I.10	Jeffrey Epstein	Sexual offences	M
I.11	John Bodkin Adams	Fraud, perverting the course of justice	M
I.12	Osama Bin Laden	Terrorism and extremism	M
I.13	Oscar Pistorius	Murder	M
I.14	Peter Sutcliffe	Murder, attempted murder	M
I.15	Reginald Kray	Murder, accessory to murder	M
I.16	Robert Maxwell	Not convicted[2]	M
I.17	Ronald Kray	Murder	M
I.18	Samantha Lewthwaite	Not convicted[3]	F
I.19	Samuel Little	Murder, attempted murder	M
I.20	Shamima Begum	Not convicted[4]	F

[1] The gender proportion in this table is approximately aligned with UK Criminal Justice System staistics on violent and serious crime, e.g. §8 of [26]
[2] Widespread posthumous evidence of fraud.
[3] Warrant issued for charges of possession of explosives and conspiracy to commit a felony.
[4] Deprived of UK citizenship due to links to terrorism and extremism.

Table 3: The "forfeited" group of recipients who were stripped of the Honour.

SOI	Name	Award	Citation	Gender	Awarded	Forfeiture reason	Forfeited
F.1	Anne Ganley	MBE	Employment	F	2012	Perverting the course of justice	2017
F.2	Ashuk Ahmed	MBE	Young people	M	2009	No criminal convictions found	2019
F.3	Craig Martin Burrows	MBE	Charitable and voluntary work	M	2004	Sexual offences	2017
F.4	David John Kemp	MBE	Education	M	2013	Sexual offences	2017
F.5	Derek Charles Eaglestone	MBE	Charitable and voluntary work	M	1994	Sexual offences	2017
F.6	Ian Richard Swingland	OBE	Conservation	M	2006	Fraud	2017
F.7	Ian Strong	MBE	Rural Community in Yorkshire	M	1997	No criminal convictions found	2019
F.8	Jawaid Mohammed Ishaq	MBE	community relations in South Humberside and North Lincolnshire	M	2000	Fraud	2016
F.9	John Anthony Coatman	MBE	young people	M	2011	Sexual offences	2019
F.10	Malcolm Belchamber	MBE	Littlehampton community	M	2004	Fraud; Forgery and Counterfeiting Act 1981	2017
F.11	Michael Nathan Cohen	MBE	Chorlton Probation Hostel	M	1998	Sexual offences	2018
F.12	Jo Shuter	CBE	Education	F	2010	Professional misconduct	2015
F.13	Patrick Robert John Rock	OBE	Political service	M	1992	Sexual offences	2017
F.14	Paul Symonds	OBE	Community Relations in Northern Ireland	M	2007	No criminal convictions found	2017
F.15	Paula Marie Vasco-Knight	CBE	Health services	F	2013	Fraud	2017
F.16	Philip Anthony Knight	OBE	British Honorary Consul-General, Antwerp	M	2001	No criminal convictions found	2017
F.17	Philippa Ann Rodale	MBE	Animal Welfare and to the community in Dorset	F	2007	Professional misconduct; animal welfare charges	2017
F.18	Robert Stanley Poots	MBE	Education	M	2010	Fraud	2017
F.19	Rolf Harris	CBE	Entertainment and the arts	M	2006	Sexual offences	2015
F.20	Trevor George Francis	MBE	Fife community	M	2012	Sexual offences	2017

References

[1] Birjali, M., Kasri, M., Beni-Hssane, A.: A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems **226**, 107134 (2021)

[2] Steinberger, R., Hegele, S., Tanev, H., Della Rocca, L.: Large-scale news entity sentiment analysis. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 707–715. INCOMA Ltd., Varna, Bulgaria (2017)

[3] Bergman, J., Popov, O.B.: Exploring dark web crawlers: A systematic literature review of dark web crawlers and their implementation. IEEE Access **11**, 35914–35933 (2023)

[4] Nielsen, F.A.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs (2011). Creative Commons Attribution 3.0 Unported

[5] Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014 (2014)

[6] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41**(3), 15 (2009)

[7] Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, ??? (1977)

[8] Cabinet Office: Having Honours Taken Away (Forfeiture). https://www.gov.uk/guidance/having-honours-taken-away-forfeiture (2021)

[9] Phillips, H.: Review of the Honours system 2004. Corporate report, Cabinet Office (July 2004)

[10] Jay, A., Evans, M., Frank, I., Sharpling, D.: I.2 operation of the Honours system. In: Allegations of Child Sexual Abuse Linked to Westminster Investigation Report, (2020)

[11] Armstrong, H.: Honours: History and reviews. Briefing Paper 02832, House of Commons Library (February 2017)

[12] Cabinet Office: Operation of the Honours system 2019. Corporate report, Cabinet Office (July 2019)

[13] Software Freedom Conservancy: The Selenium Browser Automation Project. https://www.selenium.dev/documentation/ (2004)

[14] Richardson, L.: Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/ (2004)

[15] Web Graph SIA: Brief History of Web Scraping. https://webscraper.io/blog/brief-history-of-web-scraping (2021)

[16] Margoni, T., Kretschmer, M.: A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. Zenodo (2021). https://doi.org/10.5281/zenodo.5082012

[17] Loper, E., Bird, S.: NLTK: The natural language toolkit. Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics (2002) cs/0205028

[18] Kiss, T., Strunk, J.: Unsupervised multilingual sentence boundary detection. Computational Linguistics **32**(4), 485–525 (2006) https://doi.org/10.1162/coli.2006.32.4.485

[19] Hirst, G.J.: Anaphora in natural language understanding: A survey. Master's thesis, Department of Engineering Physics, Research School of Physical Sciences, The Australian National University (1979)

[20] Qi, P., Dozat, T., Zhang, Y., Manning, C.D.: Universal dependency parsing from scratch. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,

pp. 160–170. Association for Computational Linguistics, Brussels, Belgium (2018)

[21] Nielsen, F.A.: AFINN: A New Word List for Sentiment Analysis on Twitter. https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/ (2011)

[22] Nielsen, F.A.: fnielsen/afinn. GitHub repository, https://github.com/fnielsen/afinn/tree/master (2022)

[23] Hutto, C.J.: cjhutto/vaderSentiment. GitHub repository, https://github.com/cjhutto/vaderSentiment (2022)

[24] Heuer, R.J.: Psychology of Intelligence Analysis. CIA Center for the Study of Intelligence, ??? (1999)

[25] Government Digital Service: Artificial Intelligence Playbook for the UK Government. https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government. Accessed 2026 (2025)

[26] Ministry of Justice: Women and the Criminal Justice System 2021. https://www.gov.uk/government/statistics/women-and-the-criminal-justice-system-2021/women-and-the-criminal-justice-system-2021#offence-analysis (2022)