A Data-science Approach to evaluating Honours candidates

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 is to our knowledge the first published application of data science to the UK Honours system. We have identified the key challenges which need to be overcome before data science can be applied to increase public confidence in the Honours System. In this paper we describe a system for scraping the internet for publicly available information about applicants to the Honours system and then attempting to form an opinion about them using machine intelligence. In order to form an opinion about applicants for the UK King's Honours, we have evaluated two existing sentiment algorithms (AFINN, VADER and then created our own novel algorithm MINOS). Various other Natural techniques)(E.g. Coreference resolution) have then been explored and their impact or improvements to the performance and efficacy of our system has been compared. The promising results in this work indicate that this system can be used to augment human evaluation to better judge the suitability of candidates for receiving honours and awards. In addition it can avoid major faux-pas by granting awards without thoroughly investigating the backgrounds of candidates. The system can be extended in future for other awards and recognition programs.

Keywords: Awards, Data Science, Artificial Intelligence, Published Information

1 Introduction

It is always an honour for a person to be nominated for an award. Since 1066, the monarch of the United Kingdom has recognised pre-eminent citizens by awarding Honours [1]. These stand out amongst national awards (see Annex 2 of [2] for a comparison) for their longevity, adaptability to keep the system relevant and accessible in the modern age, and the wide range of achievements rewarded. Since these awards are highly coveted, thorough due diligence needs to be done on each applicant especially the prospective candidates. Failing this, the general public might lose faith in the fairness of the Honours system. So far this entire process of due diligence would be done manually by officials. Furthermore, a core principle of the Honours System is that the award must be maintained via continued good conduct by the recipient [3]. Should the recipient's behaviour "bring the Honours system into disrepute", they may be stripped of their Honour to preserve public confidence in the system, a process known as "forfeiture" [3]. The obvious question, then, is how to do this fairly, robustly, and without undue public expense? With over 150 000 recipients of the Order of the British Empire alone [2], this is an arduous task and can benefit from leveraging the power of modern data science techniques. We examine the challenges in this unique problem in this paper. At the same time, this work tries to answer the questions:

How to measure public sentiment about a living individual?

How to define and identify "red flag" behaviours with open-source intelligence?

This would in turn have wide-ranging applications including award decisions for many other honours and awards, corporate Human-resource management, investigative journalism to national security vetting.

To reduce human intervention and produce impartiality, the procedure needs to be automated. This paper does not suggest *replacing* human selection. Instead our methods would augment the current human decision-making by extracing relevant information from data in an auditable and robust way. The extraction of insight would remain the human committees' responsibility.

The goals for this work are:

- **Maximise available data:** Use web-scraping to scan more of the web than a human could do, more efficiently, auditably and responsibly.
- **Minimise human sifting:** Use algorithms to identify relevant information from our large volume of data.
- **Identify key evidence:** Both the most positive and negative assertions can be flagged by looking at extremes of sentiment.
- **Compare multiple viewpoints:** Provide multiple sentiment algorithms so that assessment does not have a single point of failure.
- **Ensure attributability:** Ensure that both the original article content and the source URL are preserved in our results. This enables a human sifter to examine the source, its biases and quality, to judge any (positive or negative) evidence.

Although this appears to be a classic data science application—to extract a small amount of relevant information from a large corpus—we need careful handling to avoid misleading results. Three different sentiment analysis algorithms have been applied to

web-scraping of search results for a given subject of interest. Our subjects of interest (hereafter referred to as SOIs) are drawn from three distinct categories – Infamous individuals also referred to as test cases, successful Honours recipients, and Honours forfeiture cases – to illustrate the complexity in classifying *algorithmically* what to a human are three highly distinct behavioural groups.

Some of the key challenges addressed in this work:

- Search engines do not necessarily rank their results in order of relevance to their search query, which biases the impressions of a human assessor.
- Judging relevance at an article level leads to misleading sentiment results; only at the individual setnence level can we guarantee relevance.
- Evaluation of Co-reference resolution algorithms to extract all sentiment-bearing sentences about a given individual, rather than considering irrelevant information.
- Off-the-shelf algorithms have difficulty identifying "red flag" behaviours which would lead to forfeiture, even in the test group.
- Improving on the quality of the result — rather than increasing the complexity of the existing algorithms — therefore a simple approach was developed from first principles.

Hence, the three core components of the proposed method are web-scraping, coreference resolution and sentiment analysis.

Web-scraping has existed in various forms since the earliest days of the internet [4] including web crawling, web indexing and archiving of web content. We use "web-scraping" to refer specifically to the extraction of raw text from the (non-dark) web. Once extracted, the text is cleaned, chunked, and piped to a sentiment analyser. This process is explained in Section Section 3.1.

Coreference resolution is the task of associating different linguistic expressions which refer to the same entity [5]. It is a fundamental building block for more complex NLP tasks. We applied this to trace references to an SOI throughout an article. In this way we identified sentences which specifically involved our individuals, examining only the sentiment from those sentences to minimise noise in the results.

Many sentiment analysis approaches exist, and it is well beyond the scope of this paper to canvas them (see [6] for one such attempt). We apply the popular VADER model [7] optimised for social media sentiment analysis. We also use the simpler AFINN model, which is a composite of existing lists and some internet slang [8]. We define a custom sentiment model MINOS which we benchmark against two operating modes of VADER and the results of AFINN. We summarise all three algorithms in 3.4.

Sentiment analysis of public figures appears to be (curiosity) limited both in the scope of source data and the depth of analysis. This paper extends the *status quo* in three ways:

- it uses a broad range of sources rather than purely social media or news articles;
- a custom sentiment algorithm has been utilised alongside using two existing methods [7, 8].
- the relevance of any text input is validated using coreference resolution.

- SOIs from many different fields and with varying degrees of celebrity (or non-celebrity) status have been analysed.

The research in [9] uses the fraction of positive tweets by 400 celebrities as a proxy for the individuals' positivity or negativity. This differs from our method, where we measure their sentiment based upon open-source information *about* the public figures themselves, not necessarily (nor solely) produced *by* them. The paper [10] applies TextBlob's polarity measures to tweets returned from the query "Elon Musk", as an estimate of how Twitter users felt about Musk's proposed acquisition of the Twitter platform. The only academic paper we found similar to this work is [11], which sources large-scale European news data and estimates the sentiment of named entities (including public figures). However, this paper [11] aims to measure reporting bias amongst news sources—by assuming that their sources are biased, they preclude themselves from using news stories to obtain an unbiased measure of public sentiment. We make the opposite assumption, i.e. that our sources represent genuine differences in opinion or report different subsets of (truthful) facts.

The structure of this paper is as follows. In Section 2 we summarise the aspects of the Honours system which are relevant to our forfeiture problem and define our SOIs. Our methodology is described in Section 3, including data collection, cleaning, sentiment analysis and extraction of results. Section Section 4 shows the results thus highlighting the advantage of webscraping over manual research by examining the fraction of web-scraped text which survived our relevance filter criteria. Thereafter the results corresponding to various options and alternatives for the algorithm using coreference resolution have been illustrated. This is followed by the final results which can be used by analysts to determine if a candidate is suitable for an award or not. The finally concludes (5) with the contributions of this work and potential future work.

2 Background

In this section we select a specific use-case for analysing public opinion of well-known figures: that of detecting forfeiture in the Honours system. We briefly summarise how the Honours system works in Section 2.1 and the concept of forfeiture in Section 2.2. Section 2.3 describes how we selected the three groups of individuals used to test our algorithm.

2.1 The Honours system

The Honours System has existed for at least 850 years [12]. It is a public show of gratitude by the State to recognise exceptional service [13] for civic achievements (not gallantry) by living individuals. The system is dynamic and evolves in the machinery of how it operates, what Honours it awards, where nominations are sourced and why Honours are awarded.

The operational side of the Honours system is complex (see Annex 4 of [2] for various flow charts). For the purposes of this paper, a simplified process is [14, 15]:

1. Nominations are sourced from the Civil Service and the wider public;

2. Sifts by panels in the relevant government departments and specialist committees of the Main Honours Committee decide on a shortlist;
3. The Main Honours Committee determines the final list to be considered;
4. Merit checks ensure that the evidence for nomination can be substantiated;
5. Successful nominees undergo "proberty and propriety checks" e.g. criminal records and tax checks;
6. The Sovereign bequeaths the Honour by inducting the recipient into the appropriate Order at the appropriate rank.

Lists of Honours recipients are published bi-annually (at New Year and on the King's Birthday in June) in the London Gazette [16, 17]. Thus information on the name, Honour awarded, and a short[1] description of their achievement is publicly- available. This process takes around 12 months [15].

As one might expect of such a long-standing system, there are many different Orders that can be awarded, and at different ranks of merit. Since the Order of the British Empire covers $\sim 90\%$ of appointments, we concentrate on this Order for the purposes of this paper: Annex 1 of [2] covers other Orders in more detail. Each Order is split into a series of ranks. From most[2] to least senior, in the Order of the British Empire these are [14, 18]:

Knight / Dame Sustained commitment recognised by one's peers, with a national impact, pre-eminence in one's field.

Commander Prominence at national level, or a conspicuous leading role at regional level, for innovative and distinguished contributions to a given field.

Officer A distinguished role at regional level, or work which has gained a national profile.

Member Service which stands out as an example to others at regional or local level, with long-term impact in the community.

British Empire Medal Short-term service (a few years') which has a significant impact on the local community.

This enables the system to reflect achievement with varying impacts and over different durations.

To further simplify this paper, we consider only one of the three Lists, the Prime Minister's List (see §42-45 of [2] for details on the Lists), which covers nominations from the Home Civil Service and the general public. Approximately 1 800 nominations are considered by the sifting committees per round (as at 2011, [19]), which must be accompanied by supporting evidence [14]. Nominations are increasingly sourced from the general public (exact figures in [2, 12, 19–21]). However, the majority continue to come from within the Home Civil Service. The main factor which is important to this paper is that both the nominee (and the sifters) need publicly-available information to write (or in the case os sifting, vet) a potential candidate, without contacting the nominee directly. In the modern age the first step is usually the internet.

[1] At the highest levels, Knight and Dame, the long citation is published with more detail on their specific achievements.

[2] For simplicity we have grouped the Knight / Dame Grand Cross with the usual Knight / Dame.

Honours are awarded on the basis of several factors. The overriding criterion is one of "going above and beyond" in such a way that brings value to the UK. Nominees must show service beyond "doing the day job" (particularly relevant for senior civil servants and high-profile business figures [12]) which demonstrates that they [13]:

- "have delivered in a way that has brought distinction to British life and enhanced the UK's reputation";
- "have shown sustained achievement against the odds which has required moral courage in making tough choices and hard applications";
- are role models in their field, respected by their peers for exceptional achievement;
- are innovative and made a difference to their field of work or community

Regardless of their specific field, all nominees must have a record of charity and voluntary service to pass sifting. A nominee must still be active in the civic field in which they are nominated, including [14]:

- culture, sport, media and the arts
- education
- medicine and healthcare
- science and technology
- business and the economy
- civil or political service
- charitable, local and voluntary service

Since the nominee must still be active in their field of excellence, this precludes the award of Honours posthumously. An Honour is a living award, which ceases to be conferred upon an individual's death.

2.2 Forfeiture

An Honour is awarded to recognise an individual's contribution and commitment to the society at large. In case any awarded individual falls drastically in their standards of conduct this might bring the Honours system into disrepute. When a recipient brings the Honours system into disrepute, they may be stripped of their award through a process known as *forfeiture*. This involves the Sovereign "cancelling and annulling" a living individual's right to their Honour [22].

The Sovereign takes advice on when to cancel an Honour from an independent Forfeiture Committee. This body has neither investigatory, policing, nor legal powers. Instead it relies upon reporting from members of the public [3] and even motions in the House of Commons [22] to trigger a new case. Forfeiture may be considered appropriate in cases of any conduct which brings the Honours system into disrepute [22]. For example, forfeiture is almost certain in the following cases [3, 22, 23]:

- 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

Historically, to avoid forfeiture, Honours were conferred at the end of a career. This ensured that the exceptional service threshold for obtaining an Honour reflected the net effects of one's positive and negative contributions to the field [12]. The example given in [12] is that of Fred Goodwin, who was awarded a KBE for services to banking, which was cancelled and annulled after his leadership of RBS caused severe damage to the UK economy. Had his award been postponed until the end of his career, it would not have been awarded at all. However, since the early 2000s, timeliness of award is viewed as more important than waiting (potentially decades) to avert forfeiture risk, and so this is no longer seen as a tenable option.

To mitigate the forfeiture risk while also recognising achievemnts as soon as possible after the fact, and subject to sensibly and proportionately spending public funds, we investigate augmenting the current system with a data science approach.

2.3 Selecting subjects of interest

In order to analyse and understand the applicability of existing sentiment analysis algortihms including coming up with additional algortihms, a wide and balanced distribution of people needs to be considered. In this work, 3 groups or sets of people referred to as Subjects of Interest(SOI) were considered as outlined below.

- **Awarded** subjects who have received an Honour and retained it since, suggesting continued behaviour which supports the public good;
- **Infamous**: people whose behaviour stands in direct contrast to enhancing public good therefore should never receive an award
- **Forfeited** : subjects who were initially conferred awards and then later forfeited their Honour, or who were posthumously stripped of the Honour by the Committee.

We selected twenty SOIs per group. Tables A1 to A3 summarises each group and includes relevant details of their award, forfeiture, criminal convictions *etc.* as appropriate. To make our figures easier to read, each SOI is given an index number which can be cross-referenced to the table and a prefix according to their group (*e.g.* **??** 1 is Professor Dame Winifred Mary BEARD). It is the group to which the individual belongs, rather than the individual themselves, which is of interest in examiing the performance of our method.

The "infamous" group is used as a negative extreme. These SOIs are connected to serious crime, either having been convicted (thus meeting and/or far exceeding the "criminal conviction" threshold for forfeiture) or their crimes are public knowledge without the individuals facing trial (e.g. terrorists). While such individuals would never be considered for an Honours, we need to test whether they are distinguishable from the Forfeiture group (since their crimes should far exceed the misdemeanours of the latter). Our methodology should be designed to return a high degree of negativity for

these SOIs, without becoming trapped by noise from positive text which (hopefully) does not actually relate to the SOI's actions.

Conversely, the "awarded" group represents the positive extreme. We selected a range of award levels, from the highest (Knight/Dame) to the lowest (Member) [24]. We were careful to include a broad range of fields and achievements across the ten Honours Committees, ranging from the arts and sciences to business, sport and civil service [14]. We also used a mixture of male and female recipients. Most SOIs received multiple Honours (e.g. an OBE followed by a CBE) or related commendations (e.g. election as a Fellow of their relevant Royal Society or award of the Queen's Police Medal). Since this group has retained their Honour, this implies that they continue to exhibit good behaviour, in addition to the remarkable achievements which merited the Honours in the first place. Therefore we expect a well-behaved algorithm and a well-constructed method to return a high degree of positivity for these SOIs.

Finally the "forfeiture" group tests the algorithm in our specific use-case. These SOIs all received Honours of varying degrees and via varying Committees, and over a period of decades. However, unlike the "awarded" group, all SOIs were stripped of their Honours for poor conduct. This level of poor conduct varies greatly in degree of criminality, the length and nature of the offence, and the delay between award and forfeiture. In all cases, these SOIs form a more nuanced group than the other two. We expect less negative evidence than the "infamous" group (since their actions are not as extreme). We expect a similar degree of positive evidence as the "awarded" group (since they were worthy of nomination and passed sifting). This evidence will be obfuscated in the actual data, in which we expect a mixture of positive and negative sentiment. By definition, our methodology should return a clear negative score for these SOIs—otherwise it would be of no practical use as a forfeiture early-warning system. However, we also want the results to be distinguishable from the "infamous" group, who would never have been worthy of an Honour under any circumstances.

Now we have a specific problem to solve and a set of subjects on which to test our solution. We detail our approach to solving this problem in Section 3.

3 Methodology

Section 2 established the need for a data-driven, machine-learning-based approach to vet potential and existing Honours recipients. This section proposes a model to achieve this in practice.

Figure 1 shows the high-level flow of the methodology, the key stages are illustrated in the following sub-sections as given below.

1. collection of open-source text - Section 3.1
2. extraction of relevant data - Section 3.2
3. efficacy of coreference resolution - Section 3.3
4. application of appropriate sentiment algorithms - Section 3.4
5. aggregation of results for measuring the risk of a given SOI - Section 3.5
6. high-impact data for human analysis and fact-checking - Section 3.6

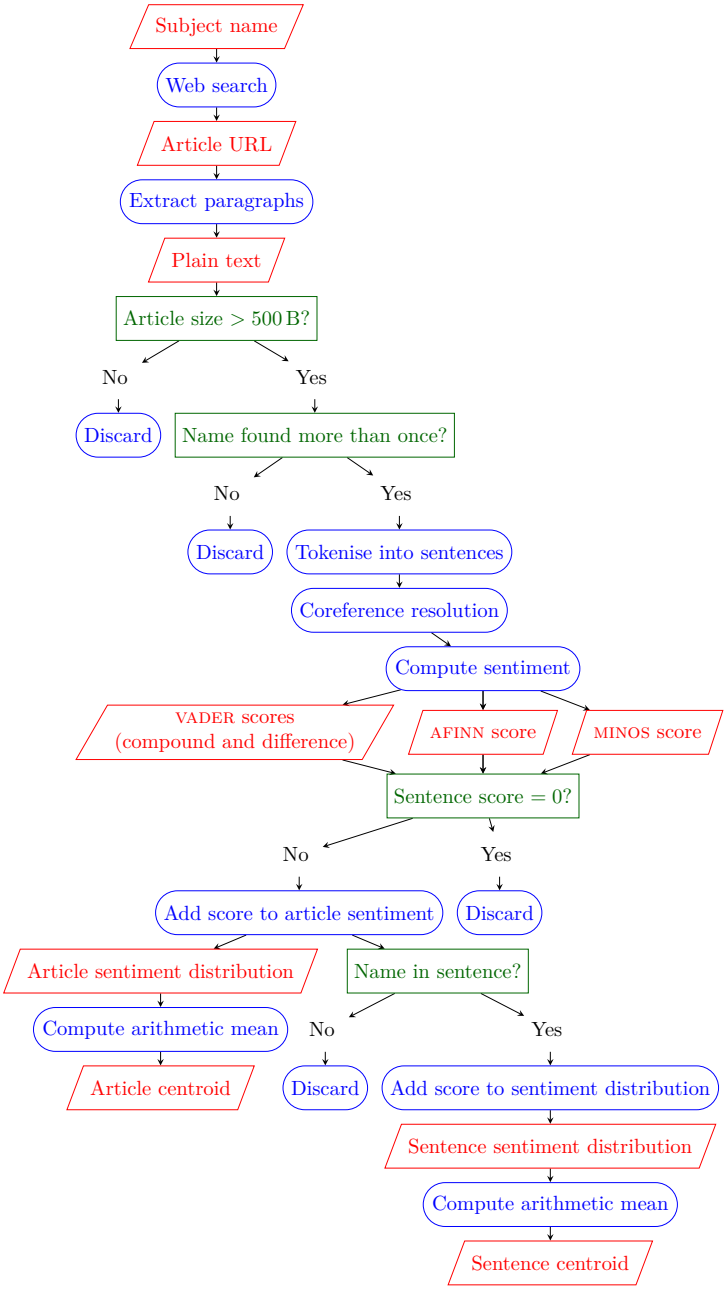


Fig. 1: Control flow of our methodology. Red trapezia represent outputs (except the starting input), green rectangles are binary decisions and blue ellipsoids are outcomes.

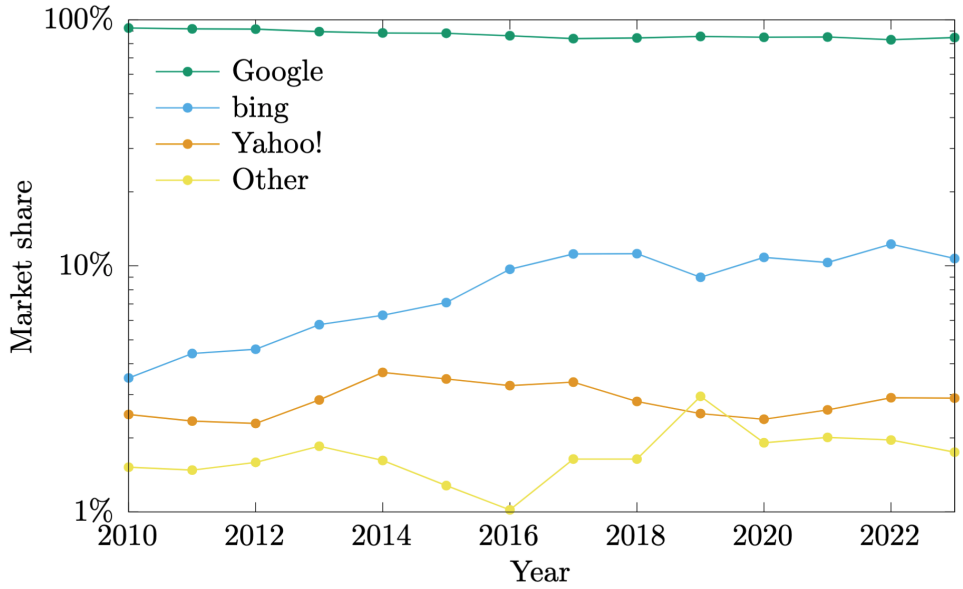


Fig. 2: Search engine market share on desktop computers in the United Kingdom [25].

3.1 Data collection and processing

Our data collection mirrored that of a human writing (or sifting) a citation as closely as possible. It differed only in our leverage of a web scraper to collect the data more efficiently than any human user.

Since the personal information submitted in an Honours citation form is often sparse or a "best guess" from a friend or colleague, we wanted to rely on the absolute minimum of detail. Similarly, notices of an award, which are published in the Gazette, consist of only a short sentence. Therefore we decided that an appropriate and realistic search query is the full name of the SOI.

To obtain public-domain data, we used a single search engine to canvas the open web. Google was an obvious choice due to its dominant market share in the UK for over a decade [25]: see Figure 2. More sophisticated crawling e.g. of social media platforms in [26] or the dark web (see [27] for a review) is a task in itself and beyond the scope of this paper. This introduces an immediate limitation, namely that Google only indexes 0.004% of the internet, or approximately 2% of the "surface" web [28].

We used the Selenium webdriver's python bindings [29] to automate the process of sending a search query to Google and visiting the URLs of each result. The number of links is generally in the range 60-70. We used the HTML and XML parser BeautifulSoup [30] to store the URL of each search result and extract raw text from each URL returned in the search results using the pre-defined `.get_text()` method[3]. This ensured that any given text could be traced to a particular source URL. Thus one can

[3]Documented at https://www.crummy.com/software/BeautifulSoup/bs4/doc/#get-text

estimate (via human verification) the quality of the source and therefore the reliability of the underlying text for each sentiment datum.

This approach limited our access to content, including:

1. being redirected to a website's `robots.txt` page
2. content missing due to Javascript not being enabled
3. we could not log in or sign up to services which required logins to access free content
4. nor did we access any paid or subscription content

While Items 1 and 2 are inevitable when using a web-scraper, Items 3 and 4 are nuisance factors which would also frustrate a human user (especially one visiting a large number of these websites, or prevented from accessing them via a corporate firewall). Thus we did not attempt to circumvent these limitations in our data collection.

For all our sentiment analysis methods, we tokenize every file by sentence. This is because the smallest sense unit in VADER is at the sentence rather than the individual word level [31]. Then each of the sentences were fed to the respective functions to calculate the sentiment value:

- The compound scheme of VADER uses the entire sentence to computer a weighted score;
- For all other algorithms, the individual words are tokenised and their individual scores summed to form the sentence score.

We used the NLTK toolkit's [32] sentence tokenizer which is trained on a large corpus of English text[4] using the unsupervised machine learning algorithm of [33]. We identified several edge cases in the corpus and tested that the tokeniser parsed these correctly or sufficiently close to avoid changing the semantics of the sentence.

NLTK correctly parsed full stops within a sentence [34]:

```
1 There have been a few improvements, including a reduction of
    brain swelling, according to Dr. Robert S. Kurtz,
    assistant chief of surgery and director of the surgical
    intensive care unit, and Dr. Kent Duffy, chief of
    neurosurgery.
```

including the use of ellipses [35]:

```
1 Murder... murder... Can you prove it was murder?
```

and unusually-punctuated sentences [36]:

```
1 Look at the MESS!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

Citation marks presented some minor problems, which did not affect the meaning of the text, e.g. [37] :

[4]Documented at: https://www.nltk.org/api/nltk.tokenize.punkt.html

```
1 Section 35 of the South African Bill of Rights provides that
    "Every accused person has a right to a fair trial, which
    includes the right... to be tried in a language that the
    accused person understands... ".[42] At the start of the
    trial, Masipa told the court that the proceedings would be
     held in English with the assistance of interpreters, and
    confirmed that Pistorius spoke English.[43]
```

was parsed as four separate sentences:

```
1  Section 35 of the South African Bill of Rights provides that
     "Every accused person has a right to a fair trial, which
     includes the right... to be tried in a language that the
     accused person understands...
2  ".
3  [42] At the start of the trial, Masipa told the court that
     the proceedings would be held in English with the
     assistance of interpreters, and confirmed that Pistorius
     spoke English.
4  [43]
```

More complex sentences created minor problems, e.g. this sentence which nests two questions and ends with the inverted commas (there is no trailing full stop) [38]:

```
1 In a telephone conversation with an old friend that summer,
    Ahmed was peppered with the standard questions  "What are
    you doing now? What are you up to?"
```

In this last example, NLTK split the text as:

```
1 In a telephone conversation with an old friend that summer,
    Ahmed was peppered with the standard questions  "What are
    you doing now?
2 What are you up to?"
```

whereas ideally we would prefer to have the entire sentence including both questions marked as a single unit of teat.

These pitfalls were sufficiently minor that we judged further training of the model on a more "chat-speak"-oriented training set (to better reflect, e.g. forum threads, blog comments and email digests in]our sample) to be disproportionate to the possible improvement in tokenisation. A key factor is this decision was that the word is the main unit of meaning in three of our four algorithms. If a sentence were incorrectly tokenised, NLTK would be more likely to split sentences too soon that to amalgamate sentences into larger portions of text. Although this would affect different sentiment algorithms differently, outliers caused by very long sentences should be easy to spot by their unusually large score (especially in AFINN) and examined individually.

3.2 Article relevance criteria

The plain text having been extracted from each search result, we now need to assess its relevance. We used two measures as a proxy for relevance:

1. File size ≥ 500 B (i.e. 500 characters)
2. Name of SOI mentioned at least once

Item 1 was chosen as a sensible lower bound since this is about the length of a short paragraph. We found a number of edge cases in Item 2 in the names of many Honours recipients, including:[5]

- surname more than a single word is not *a priori* distinguishable from several forenames (e.g. "Stephen William Hawking" has one surname);
- surname is also a common noun (e.g. "Beard"; "Cook" etc.)
- better-known by their middle names (e.g. Dame Winifred Mary Beard publishes academically and is known in layman's circles as "Mary Beard" not "Winifred Beard");
- individual referred to by their title (e.g. Baron or Lord Bhattacharyya rather than "Sushanta Kumar Bhattacharyya")
- individual referred to by their peerage (e.g. "the Duke of Wellington" rather than "Arthur Wellesley")

We defined the exact criterion for Item 2 as at least one occurrence of the forename by which an individual is best known, followed by a single space, followed by all parts of their surname, and required the last element of the surname to be capitalised. This balanced addressing the edge cases above with a desire to avoid hard-coding specific exemptions (particularly those for peerages). Given the informal manner of much online text, we felt that byname followed by surname was likely to avoid filtering out articles which were actually relevant; whereas requiring someone's formal name as specified in the London Gazette would exclude too many articles.

We applied these filters separately to examine their impact, which we cover in **??**. To be a "relevant" file, we demanded that both filter criteria were met.

3.3 Coreference resolution

A key limiting factor in our analysis is the resolution at which we can link a piece of text to an individual. Our filters in Section 3.2 exclude or include entire articles. To filter at the paragraph or sentence level, we must determine whether that excerpt refers to our given SOI.

Consider the following excerpt from [39]:

> Dame Winifred Mary Beard, DBE, FSA, FBA, FRSL (born 1 January 1955) is an English scholar of Ancient Rome. She is a trustee of the British

[5] Inclusion as an example in this list does not indicate whether or not an individual was included in our SOI list.

> Museum and formerly held a personal professorship of Classics at the University of Cambridge. She is a fellow of Newnham College, Cambridge, and Royal Academy of Arts Professor of Ancient Literature.
>
> Beard is the classics editor of The Times Literary Supplement, where she also writes a regular blog, "A Don's Life". Her frequent media appearances and sometimes controversial public statements have led to her being described as "Britain's best-known classicist". The New Yorker characterises her as "learned but accessible".

The only obvious *deterministic* criterion for relevance is whether the SOI's name is included or not. This leads to a large number of false negatives whereby our criterion would exclude setnences which are actually relevant. In our example above, only one sentence is guaranteed to be about Mary Beard (since the "Beard" in the second paragraph may not be her). However, it is intuitively obvious to a human reader that all of the sentences are about Dame Mary, thanks to our understanding that:

1. Pronouns "she" and "her" all mean "Dame Winifred Mary Beard";
2. The anaphora "Beard" is the same Beard as Dame Mary;
3. The co-ordinating conjunctions (both "and") transfer the meaning of "she" from the first half of the sentence to also apply to the second half (*i.e.* Beard is a Fellow and the same Beard is also a Professor);
4. The preposition "as" link the descriptions to Beard (as the object of them) and not to the New Yorker (as the subject making the declaration).

This shows that our problem is not limited to the identification of pronouns. Rather, it is a more complex problem of grammar and semantics (discussed in detail in [5]).

Coreference resolution is the task of identifying expressions in text which all refer to the same entity. Usually the references are anaphorae, *i.e.* a short-hand (*e.g.* Items 1 and 2) or even an elision (*e.g.* Items 3 and 4) of an entity, whose presence is inferred thanks to grammatical rules. Although several open-source coreference resolution methods are available, we chose [40] on the basis of it being a reputable, long-term, Python implementation. The results are assigned based on supervised learning from a training set of English text, mapping anaphorae to their most likely root entities. The text is then output with each anaphora replaced by their named entities. The main risk is that any resolution is *probabalistic*, whereas a human does this so intuitivly that it appears to be deterministic. Thus we must "trust but verify" by quality assuring a sample.

To quality assure the results, we compared a sample of articles in both the resolved and raw states. We chose articles which passed both our relevance filters. From this shortlist we examined one reputable article (e.g. the encyclopaedia entry cited above), on the basis that the articles which are most reliable are those which should be given the most weight when considering an individual's reputation. In these articles we looked for particularly complex sentences which had the highest likelihood of confusing the resolution algorithm. Due to the volume of text involved, we were unable to examine a text from every SOI in detail.

3.4 Choice of sentiment algorithms

Sentiment analysis is the automated evaluation of emotive language from text. For the (relatively simple) purposes of this paper, we concentrate on "valence" (strength of feeling, [8]) and "polarity" (positivity or negativity [41]) of sentiment. There are two main approaches: lexicon-based, or machine-learning-based [41, 42].

We applied two "off-the-shelf" approaches: one lexical (AFINN) and the other machine-learning (VADER) [31]. Neither of these produced satisfactory results [cref] because they were tailored to meet different contexts (principally analysis of social media) which did not align with our source material. Therefore, we created a bespoke method MINOS, which uses a rule-based algorithm applied to a specialist lexicon. We benchmarked this by comparing its performance to VADER and AFINN. This subsection outlines the three different algorithms.

3.4.1 AFINN

AFINN is the simplest algorithm in terms of logic [8]. It measures strength and polarity of feeling in 1 468 unique words. A word list is maintained (see [43] for the various versions of the list), derived from a variety of prior sources. Each word is assigned a value based largely on experiential (human) classification. The values are integers in the range $[-5, 5]$, where the sign denotes polarity and the magnitude denotes valence [44]. For more detail on the construction of the list consult [8].

AFINN's default unit of meaning is the individual word. Since we tokenised by sentence, not by word, we need to construct a sentence score. We lemmatise the words in the sentence and search for them in the AFINN list. The total sum is the final sentiment of the sentence. Since we do not use any contextual information which is conveyed in the larger unit of sense, this has some obvious drawbacks:

- We cannot identify negations (*e.g.* "bad" and "not bad" both only count "bad");
- Adverbs and adjectives do not compound ("very bad","bad","not too bad" all score equally);
- Punctuation and case are ignored ("BAD","BAD!!!" and "bad?" all score equally; "bad-ass" becomes "bad ass" which scores as "bad" plus "ass");
- Contextual meaning is ignored (is "ass" in the example above referring to a donkey or a pejorative expression for a person, or a bottom?).

To mitigate these limitations, we also applied a more complex methodology, VADER.

3.4.2 VADER

VADER is a rule based sentiment intensity analyser. A dictionary containing negative words is maintained along with another one on booster words. These words enhance or decrease the sentiment values of other words. Idioms are also considered here. Initial cleaning includes removal of punctuations, converting words with punctuations to only words. A check is made to determine whether preceding words to a particular word increases or decreases its valency. Words containing all uppercase letters are given separate sentiment values. VADER algorithm provides the positive,negative and

neutral scores for a particular piece of text. Also, it contains a score normalised from the three aforementioned scores called a compound score. This compound score gives the general sense of the sentence. If the sentence is positive or negative or neutral is a different concept from whether the individual words are summed.

The version of VADER that we used had 7 525 words and acronyms in its lexicon (see [31] for the most up-to-date corpus).

3.4.3 MINOS

Our own algorithm, MINOS[6] (Minimal-complexity INference for OBE Sentiment), is a polarity lexicon combined with simple algorithmic logic. Similarly to AFINN, we use a lexical approach drawing upon different wordlists. However, unlike AFINN's valence approach, which scores on a scale of $[-5, 5]$, we take only the polarity of the valence, scoring as plus (minus) one in the positive (negative) list. This was a practical decision since we needed to integrate different wordlists, which had different rating scales (e.g. $\mathbb{R} \in [-1, 1]$ for VADER and $\mathcal{Z} \in [-5, 5]$ for AFINN) without manually re-computing the valence.

We maintain one list for positive words and another for negative words. The positive list contains 2 024 words and the negative 4 657. The latter is more detailed because our primary objective is to identify forfeiture-likely activities, which are inherently negative. In particular, we include 145 synonyms of criminal activity, covering a range of words from informal ("filching" i.e. theft, "rubout" i.e. contract killing) to formal ("embezzlement") registers and including legal terms ("larceny","manslaughter"). While this targets one key criterion of forfeiture (a criminal conviction resulting in a sentence of 3 years or more), it also illustrates the difficulty in transforming a simple legal judgement into the wide variety of ways this might be expressed. The positive list includes specialist decorations and orders, both as acronyms (e.g. "BEM", "OBE" etc.) and phrases (e.g. "Order of St Michael and St George") because we expect these to occur in our specific context. Our lists also contain some deliberate mis-spellings (e.g. both "accessable" and "accessible") as done by [7]. Words not appearing in either list are implicitly neutral and scored as zero.

We search for each entry (word or phrase) in a given sentence. Presence of a positive (negative) word will increment the sentence score by plus (minus) one. However, if a sentence has at least one negative word, positive words will not be checked any further. This approach mirrors the requirement for good conduct to maintain an Honour: if (sufficiently) negative evidence is unearthed, no amount of positive evidence will compensate.

3.5 Integration over the parameter space

Since a distribution is an inconvenient measurement to give to stakeholders, we devised a method to marginalise over the distribution, summarising our sentiment as a single value.

[6] In Roman mythology, King Minos weighs the hearts of the deceased against a feather: if the deceased committed any wrong-doing, the scales became unbalanced and the unfortunate was barred from the Elysian Fields.

First we define this more rigorously. We assume that there exists some true sentiment value θ (which we could obtain if we had all the information about an SOI, if our sntiment algorithm were perfectly capable of reflecting all numances within this text, and if we had enough computing power to undertake such a task). Since none of those pre-requisites is true, instead we approximate θ by a combination of our prior information $\pi(\theta)$ and the data $\mathbf{x}$ collected in our distributions $f(\mathbf{x})$. Via Bayes' Theorem, we find that:

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})} \tag{1}$$

gives the probability that θ is the "true" value of our sentiment given the data $\mathbf{x}$. Since some of our sentiment algorithms are continuous while others are discrete in distribution, we drop the normalisation factor $f(\mathbf{x})$.

We decide on our prior using two straightforwad assumptions:

1. We are equally ignorant of all sentiment values
2. Adding neutral information should not change the sentiment

Item 1 suggests a uniform prior $\mathcal{U}[a,b]$ where $-\infty < a < b < \infty$ are the bounds set by our sentiment algorithms:

$$\pi(\theta) = \begin{cases} 0 & \text{where } \theta < a \\ \frac{1}{b-a} & \text{where } a \leq \theta \leq b \\ 0 & \text{where } \theta > b \end{cases} \tag{2}$$

$$= \frac{1}{b-a}\left[\mathcal{H}(\theta - a) - H(\theta - b)\right] \text{ where } \lim_{x\to a^+} = \lim_{x\to b^-} = 1 \tag{3}$$

where $\mathcal{H}(\theta)$ is the Heaviside step function and our conventions are set by the bounds of the parameter space (i.e. $x \geq a$ must be right-continuous since x only approaches a from the right, whereas $x \leq b$ must be left-continuous since x only approaches b from the left).

However, Item 2 forces us to exclude zero from this distribution. This can be done by invoking the Dirac delta (provided that $a \neq 0$ and $b \neq 0$). Combining these we find the prior:

$$\pi(\theta) = \frac{1}{b-a}\left[1 - \delta(\theta)\right]\ \left[\mathcal{H}(\theta - a) - \mathcal{H}(\theta - b)\right] \tag{4}$$

The likelihood $f(\mathbf{x}|\theta)$ is straightforwardly computed from our data. Each sentence produces one datum within the range $[a,b]$ and we weight all sentences equally.

Thus our posterior is the arithmetic mean of all nonzero sentiment scores. We can integrate over all results or compute a result from each article. Since each datum is equally weighted there is no difference between the arithmetic mean of all the results on a per-article basis and the centroid marginalising over all articles at once.

We are left to find $[a,b]$ which is the range of our parameter space for θ. For VADER the limits are $[-1,1]$ regardless of the unit of tokenisation [7]. AFINN has a(n integer) valence per word in $[-5,5]$ which maps to $[-5N,5N]$ for a sentence with N words [8]. Similarly MINOS only takes the values $\{-1,0,1\}$ per word, but at the sentence level

this extends to $[-N, N]$. Therefore we must normalise the centroid before comparing them.

3.6 High-impact data for human analysis and fact-checking

All the previous subsections explained the steps in the methodology which have been optimised to produce a high-quality clean data for rapid analysis and fact-checking by a human analyst. As mentioned previously, the objective of this work is not to eliminate the human from the loop rather to aid and assist human beings in making high-quality decisions based on clear indicative data and radically reduce the probability of errors and omissions.

4 Results, Analysis and Discussion

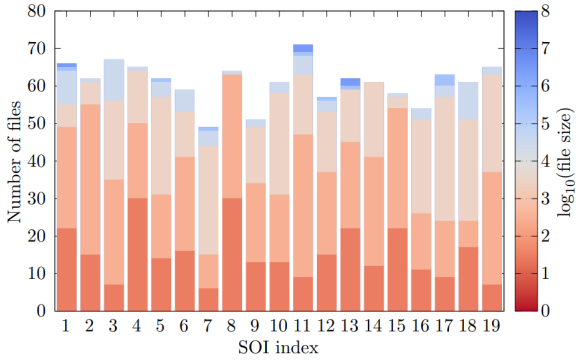
This section presents the obtained results and discusses the key observations. The section is organised in the sequence of the steps performed in the methodology as explained in section (Section 3). The analysis shows how each step transforms the data in order to extract key valuable information as well as discard the noise and impurities from the data. The following sub-section outlines the results by comparing the input and output data from each key step of the methodology.

4.1 Results for Relevance Filtering

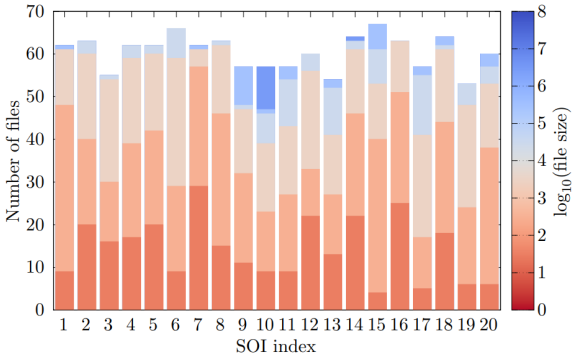
As outlined in section Section 3, the articles or web-pages obtained for an individual needs to be filtered based on relevant to find the relevant articles while removing the irrelevant or spurious articles. This prevents the following following steps in the process flow from generating poor quality of results.

In this section, we analyse the impact of the steps followed in the methodology as described in Section 3 and rationale behind using them.

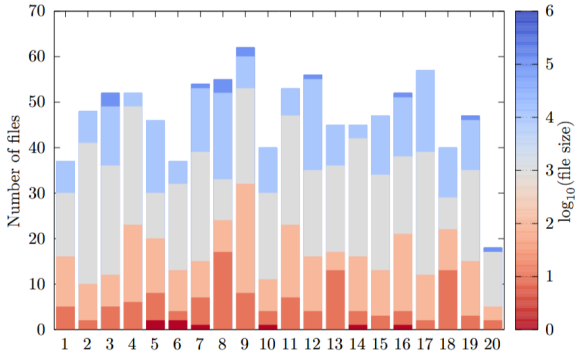
4.1.1 Filtering Relevance with File size and Name Counts



(a) Awarded



(b) Forfeited



(c) Infamous

Fig. 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 3 shows the distribution in file sizes for each article by group and individual. Each individual in the group has a stacked bar whose height represents the number of articles. Most SOIs have an internet footprint of at least 50 files. The exceptions are individuals A.7 and A.10 (refer to Table A1) in the awarded group with around 40 articles and individual T.20 in the test group (refer Table A2) with fewer than 20 articles. This is more material than a human (either sifting or nominating) would likely read about an SOI, which immediately shows the need for machine assistance to curate what is published about an individual.

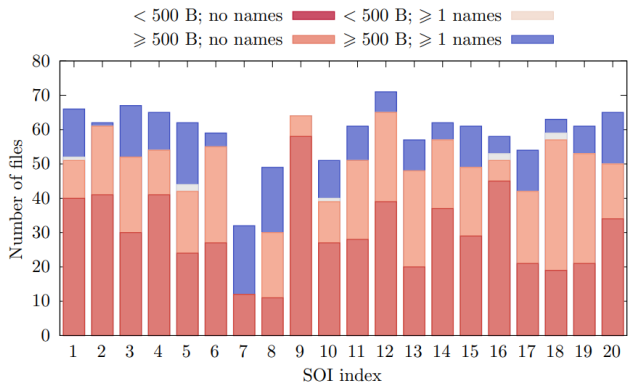
On closer examination, the article lengths span seven orders of magnitude. Within the stack, the files are binned into log-spaced groups whose colour represents the file size. Red files are smaller (approximately less than 100 B), blue files are larger (than 1 MB) and neutrally-coloured files are between these sizes. We can draw inferences about the relevance of these articles by examining both the total number of files and the distribution in size.

No SOIs have files of fewer than 10 bytes (deep red). The bright red files are $10-100$ bytes, which is too short to contain relevant information about an individual. All SOIs have some files of this brevity, which are likely to be due to paywalls and login prompts for some websites. Individuals A.4 and A.8 (refer Table A1) and SOI F.7 (refer Table A3) stand out for about thirty files of this size, which means that most of their results are spurious. In Figure 3 pale red indicates files that are $100 - 1\,000\ B$, fewer than a thousand characters, or about the length of a paragraph. Files need to contain at least a paragraph of text for them to be relevant. Some SOIs, e.g. A.8 in the Awarded group, have a corpus entirely of short files, so their internet footprint is unlikely to be suitable for sentiment analysis.

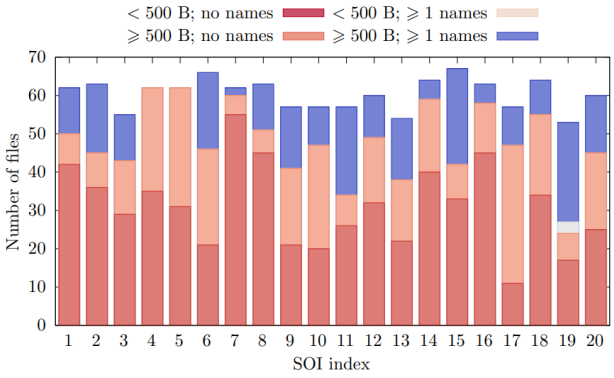
Large files at the blue end of the colourmap will dominate the sentiment results of an SOI. The deep blue files—just visible at the top of A.11 and A.13 (refer Table A1) from the Awarded group are $10 - 100$ MB in size, which is an enormous amount of text ($\gtrsim 10^6$ words). To a lesser extent, the files indicated by the bright blue boxes in Figure 3 which are of sizes $1 - 10$ MB are suspect for the same reason. Very few people, e.g. A.1, A.11, A.13, A.17 in the Awarded group (refer to Table A1) and F.10, F.14 in the Forfeited group (refer Table A3) have such large articles or web pages in their search results. The individual F.10 (refer Table A3) is especially noteworthy since there are ten files of 1 MB or larger. No SOIs in the test group have files $\gtrsim 1\ MB$. Such large files are highly unlikely to be focused on one individual. Although the blue files will contain some relevant information, this needs to be extracted appropriately from the rest of the text.

This leaves the neutrally-coloured files of $1\ kB - 1\ MB$, in pale red, grey and pale blue respectively. These are sufficiently large to contain non-negligible text, yet sufficiently small to plausibly focus on one individual. However, not all individuals have a large number of files in this category. Thus, a non-negligible fraction of the internet footprint of our SOIs will either need to be discarded or carefully filtered to obtain data which are specifically relevant to that individual.

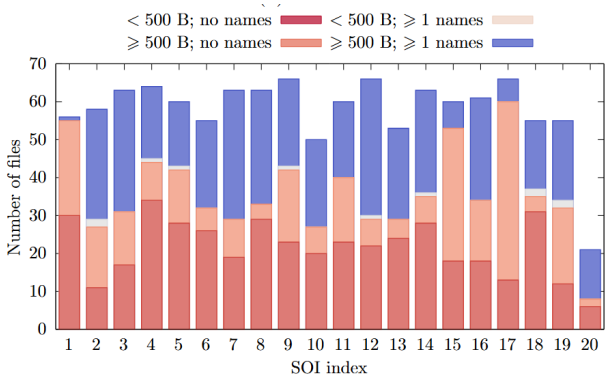
4.1.2 Filtering relevance with Name count



(a) Awarded



(b) Forfeited



(c) Infamous

Fig. 4: Files which survive each filter per SOI group.

Figure 4 shows the fraction of articles which survived neither, one or both filters. Each bar chart has one bar per SOI with one vertical unit per article, so the total height is the number of articles retrieved. Red articles failed both filters, i.e. were less than $500B$ and had no instances of the SOI's name. Orange articles were larger than $500B$ and had no instances of the SOI's name. Grey articles did contain the soi's name, but at fewer than $500B$, these articles are unlikely to contain any useful information. The blue articles are those which survived both filters and were used for sentiment analysis.

Some individuals stand out due to the lack of relevant information retrieved in particular two individuals A.8 (refer to Table A1) and F.7 (refer Table A3). In all three groups (awarded, test and forfeited), the number of articles failing the first filter (based on file size) is much larger than that failing the second filter (based on the number of times their name is included in the article).

4.2 Comparison of Centroids Sentiment Analysis Algorithms

This subsection compares the centroids obtained for individuals from each of the groups - awarded, infamous/test and forfeited using various sentiments analysis algorithms. Figure 6 shows the centroids corresponding to the three algorithms outlined in Section 3.4 using Raw-data. The term Raw-Data indicates that the data before any of the filters in the previous sections have been applied.

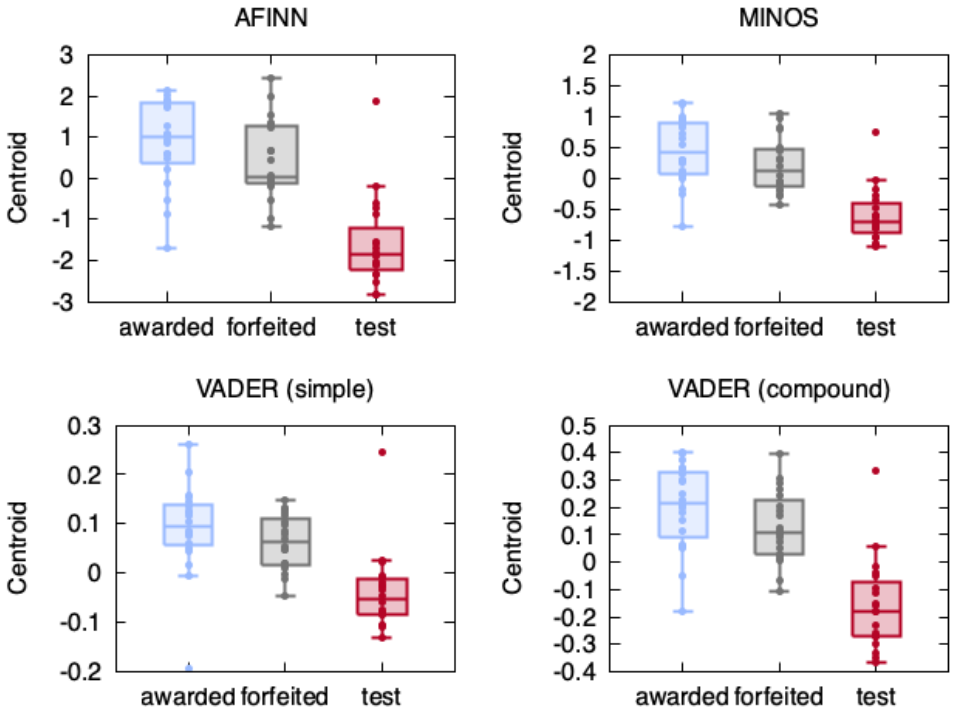


Fig. 5: Box plot of centroid distributions from raw data without any filtering. Each datum is the centroids for one SOI in that group. The boxes show the inter-quartile range of the SOI group, with the mean shown as a horizontal line. Red, gray and blue show test/infamous (Table A2), forfeited (Table A3) and awarded (Table A1) groups respectively.

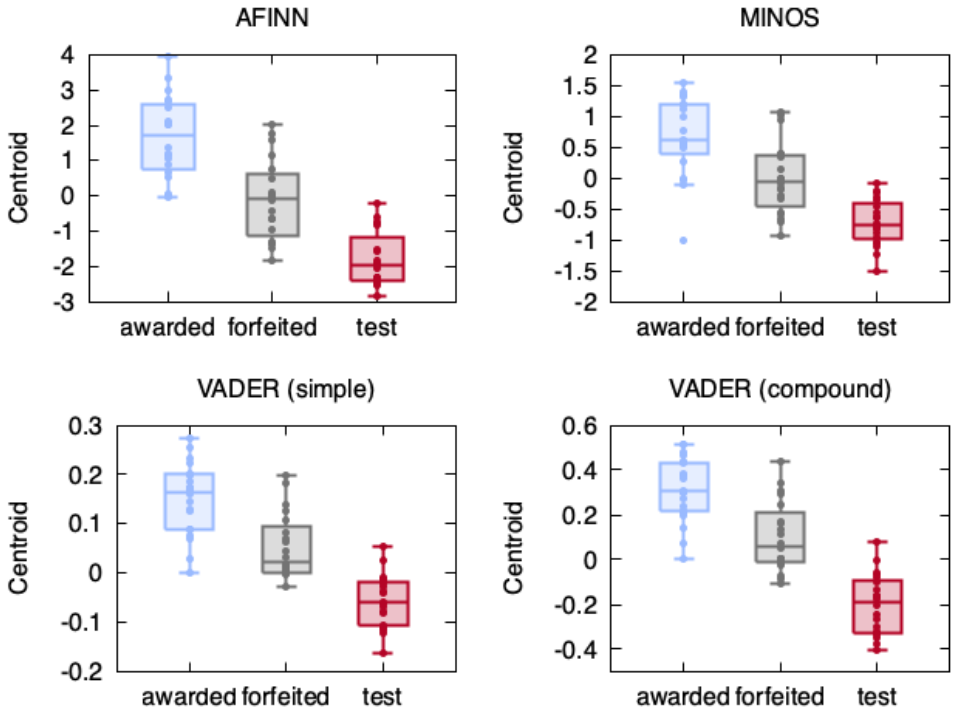


Fig. 6: Box plot of centroid distributions from raw data, including articles which only pass the filters in Section 4.1.2. Each datum is the centroids for one SOI in that group. The boxes show the inter-quartile range of the SOI group, with the mean shown as a horizontal line. Red, grey and blue show test (Table A2), forfeited (Table A3) and awarded (Table A1) groups respectively.

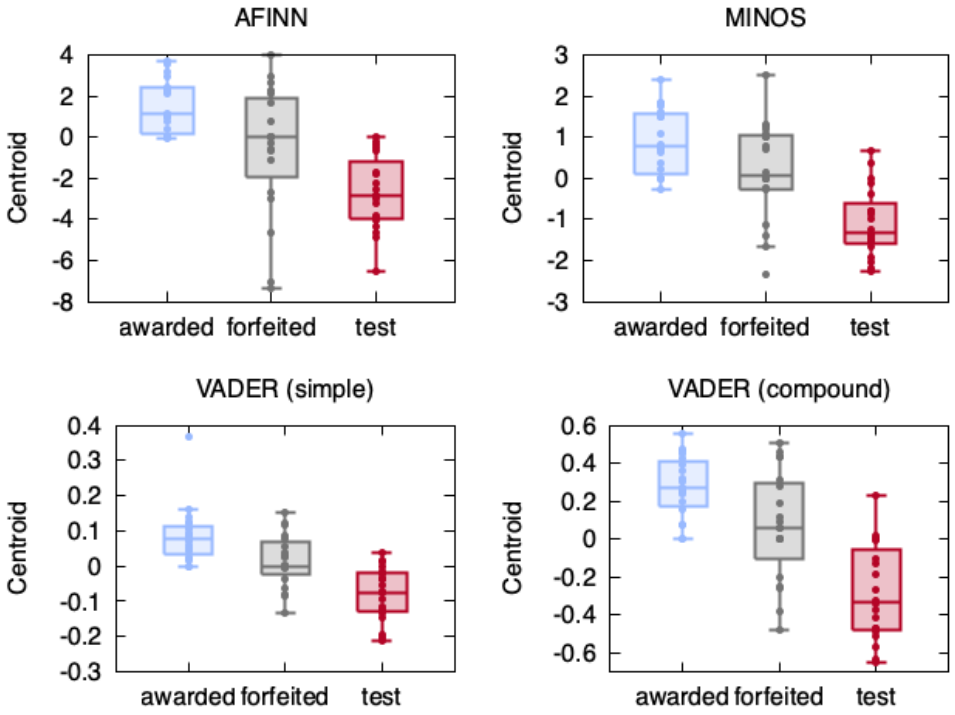


Fig. 7: Box plot of centroid distributions from raw data, including only sentences which contain the SOI name. Each datum is the centroids for one SOI in that group. The boxes show the inter-quartile range of the SOI group, with the mean shown as a horizontal line. Red, grey and blue show test (Table A2), forfeited (Table A3) and awarded (Table A1) groups respectively.

4.3 Analysis of an Individual's Sentiment Score Distribution

This subsection presents the impact of the various options, steps and alternatives in the proposed algorithm through the analysis of sentiment score distribution. The results have been shown for example individual Howard Marks for demonstrating the impact of the various key steps of the algorithm. These are results are shown in Figure 8 to Figure 11, each of which correspond to a sentiment analysis algorithm. Each of these plots are in turn composed of subplots (a) to (f), which correspond to various possible options evaluated for the algorithm as follows. In addition the salient observations from these pots have been explained below.

(a) **Article 0 Threshold:** Corresponds to running the sentiment analysis on the article after filtering based on file size, before any name reference based filtering. Therefore it shows a high number of sentences which have sentiment scores compared to the other subplots.

(b) **Coref Article 0 Threshold:** Results after running the Coreference resolution algorithm. The reduction in the number of sentences is a side-effect of the coreference resolution package used as it sometimes combines together a few sentences to make the name reference clear.

(c) **Article 1 Threshold:** This subplot refers to the sentiment score for sentences where the name of the SOI occurs at least once in the article. Without this filter it might possible that the article is not about the SOI at all.

(d) **Coref Article 1 Threshold:** In this case the coreference algorithm is applied after the name filter which ensures that the SOI's name occurs at least once in the article.

(e) **Sentence:** This plot corresponds to distribution of sentiment scores based on sentence level, where the sentence contains the name of the SOI at least once. Therefore, the numbers on the y-axis are much lower corresponding to the other subplots (a) to (d).

(f) **Coref Sentence:** In this plot the coreference algorithm is applied before the sentiment analysis algorithms. The distribution corresponds the sentiment scores for each sentence which contains the SOI's name at least once. Therefore y-axis values are much lower than subplots (a) to (d) but higher than that of (e). The reason being, after coreference the pronouns like 'he' or 'she' is replaced by the name of the SOI hence there are greater numnber of sentences with the SOI's name.

Comparing the plots in Figure 8 to Figure 11 shows that the proposed MINOS sentiment algorithm greatly reduces the spurious information and noise from the input data. It provides a clear view of whether an individual is worthy of an award. Minos greatly reduces number of sentences which are not directly related to the individual

being assessed. This is explained below using an example related to SOI Ajmal Kasab (refer to T.1 in Table A2).

The example sentences below from an article on the Mumbai terrorist attack, 2008 which happened in India is as below:

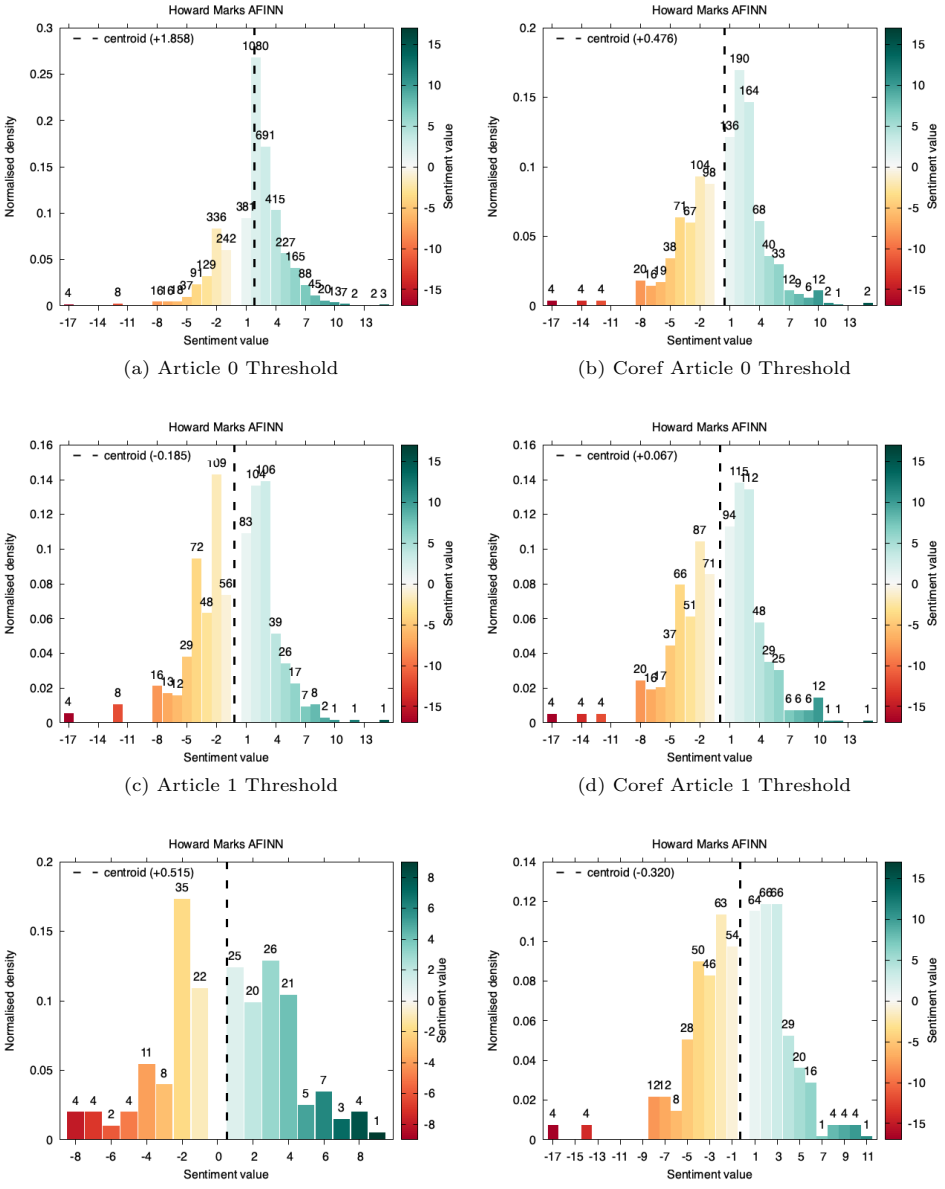
Sentence A : The Taj Mahal Palace Hotel is an iconic luxury hotel in Mumbai built in 1903.

Sentence B: When terrorists attacked the Indian city of Mumbai in 2008, employees of the Taj Mumbai hotel displayed uncommon valor.

Sentence C: Assistant Sub-Inspector Tukaram Omble courageously held onto Kasab's weapon, enabling Omble's colleagues to capture Kasab alive.

From the above 3 sentences, AFINN and Vader algorithms indicate a total positive sentiment for sentences A and B. The first sentence A contains positive words like 'iconic' and 'luxury' therefore it is positive in nature. Sentence B contains the positive word 'valor', hence produces a positive sentiment. However sentences A and B do not not talk about Kasab. Hence these should not contribute a positive score to Kasab. Sentence C mentions Kasab name, however the positive sentiment from the word 'courageously' is ascribed to the police officer Tukaram Omble who helped capture Kasab. However, AFINN and Vader both incorrectly indicate a net positive sentiment for the sentence.

These fallacies has been addressed by the MINOS in two ways. Firstly, the sentiment score of a sentence is only taken into account if the SOI's name was mentioned directly or indirectly (determined through co-reference resolution) within the sentence. Secondly, if there is a single negative word in the sentence then the entire sentence is assigned a net negative score, since it often implies that it might have a negative connotation for the SOI. Therefore, the algorithm errs on the side of caution. This is a necessary safeguard which will prompt the human analyst using the system to investigate the details the negative sentences which are output at the end of algorithm's execution. It is envisaged that this additional caution will avoid having to forfeit awards later.



(a) Article 0 Threshold

(b) Coref Article 0 Threshold

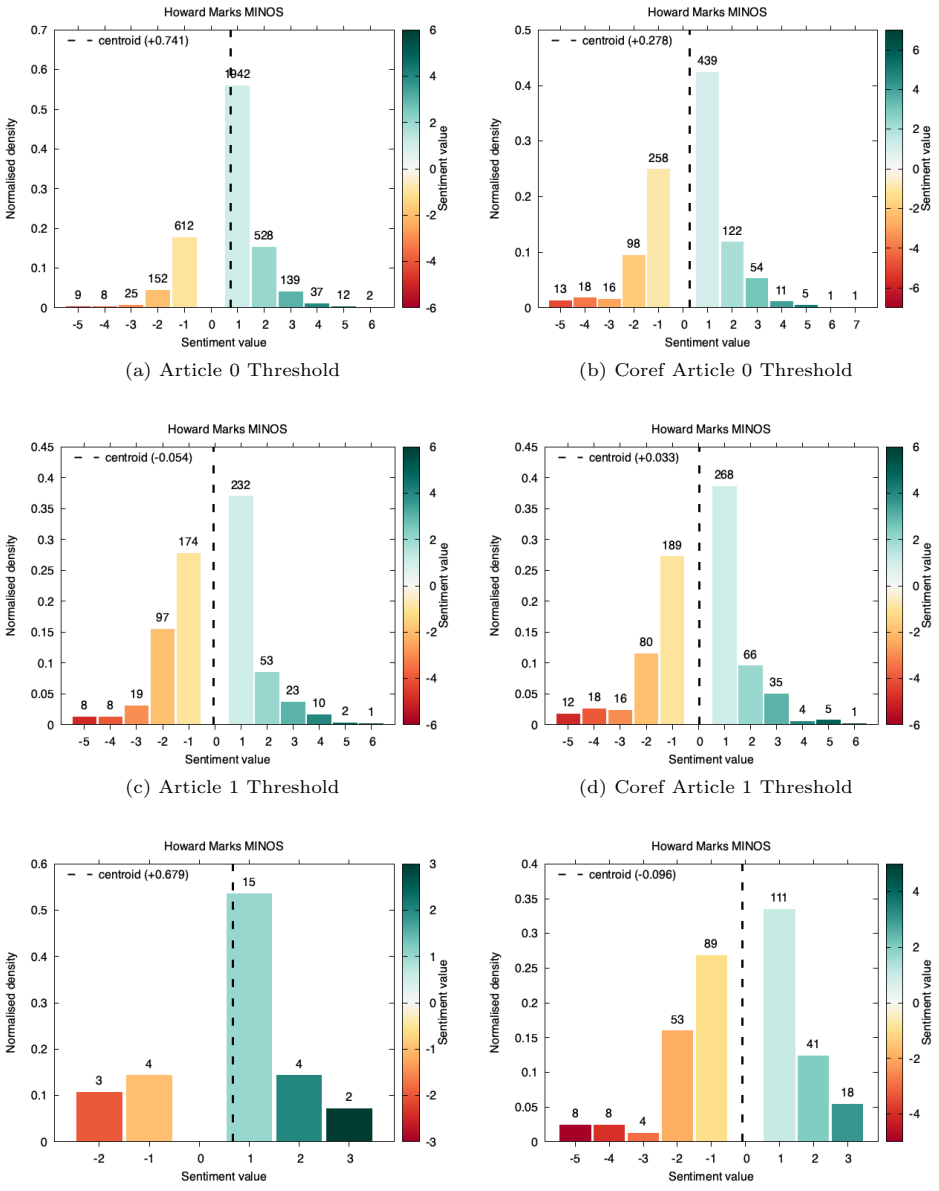
(c) Article 1 Threshold

(d) Coref Article 1 Threshold

(e) Sentence

(f) Coref Sentence

Fig. 8: Distribution of sentiment scores using AFINN algorithm corresponding to Howard Marks



(a) Article 0 Threshold

(b) Coref Article 0 Threshold

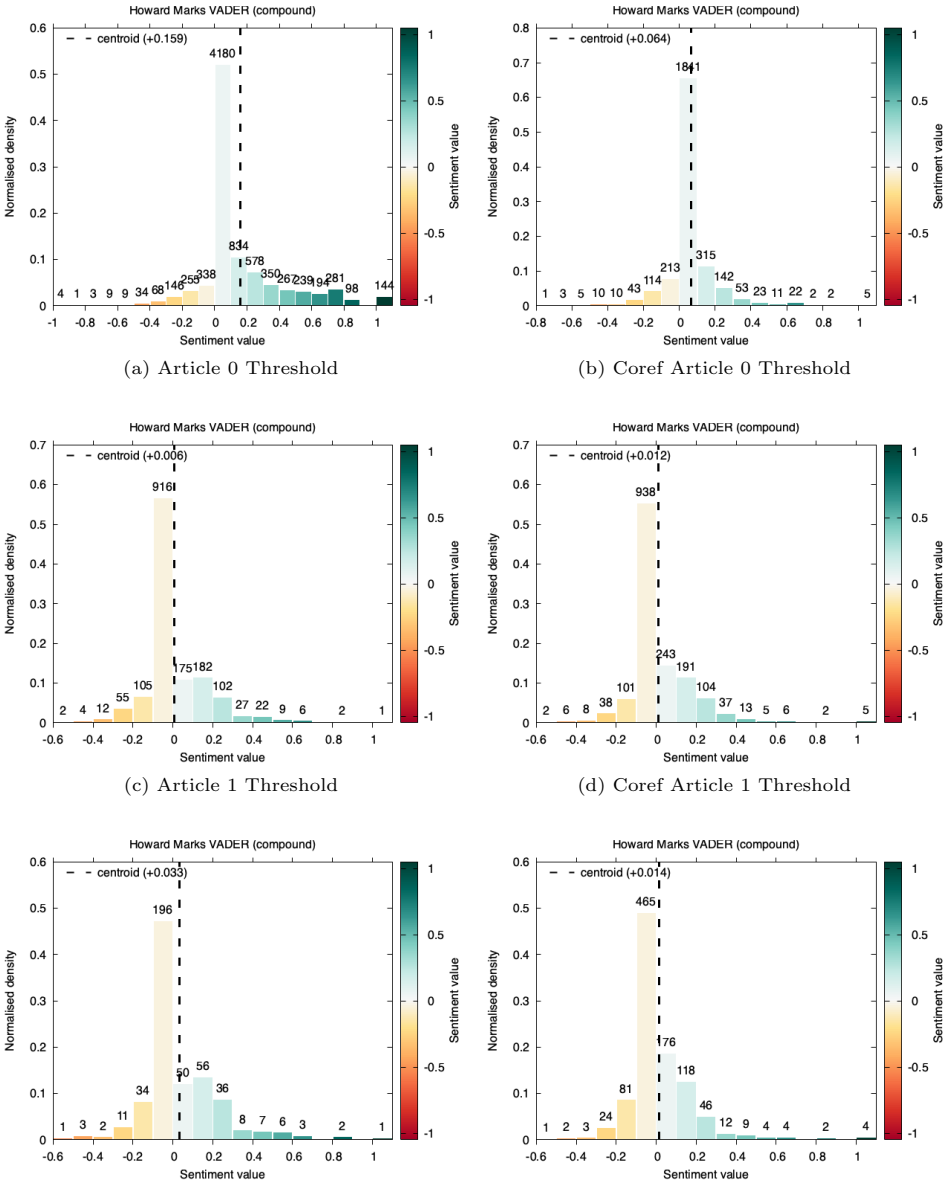
(c) Article 1 Threshold

(d) Coref Article 1 Threshold

(e) Sentence

(f) coref Sentence

Fig. 9: Distribution of sentiment scores using MINOS algorithm corresponding to Howard Marks



(a) Article 0 Threshold

(b) Coref Article 0 Threshold

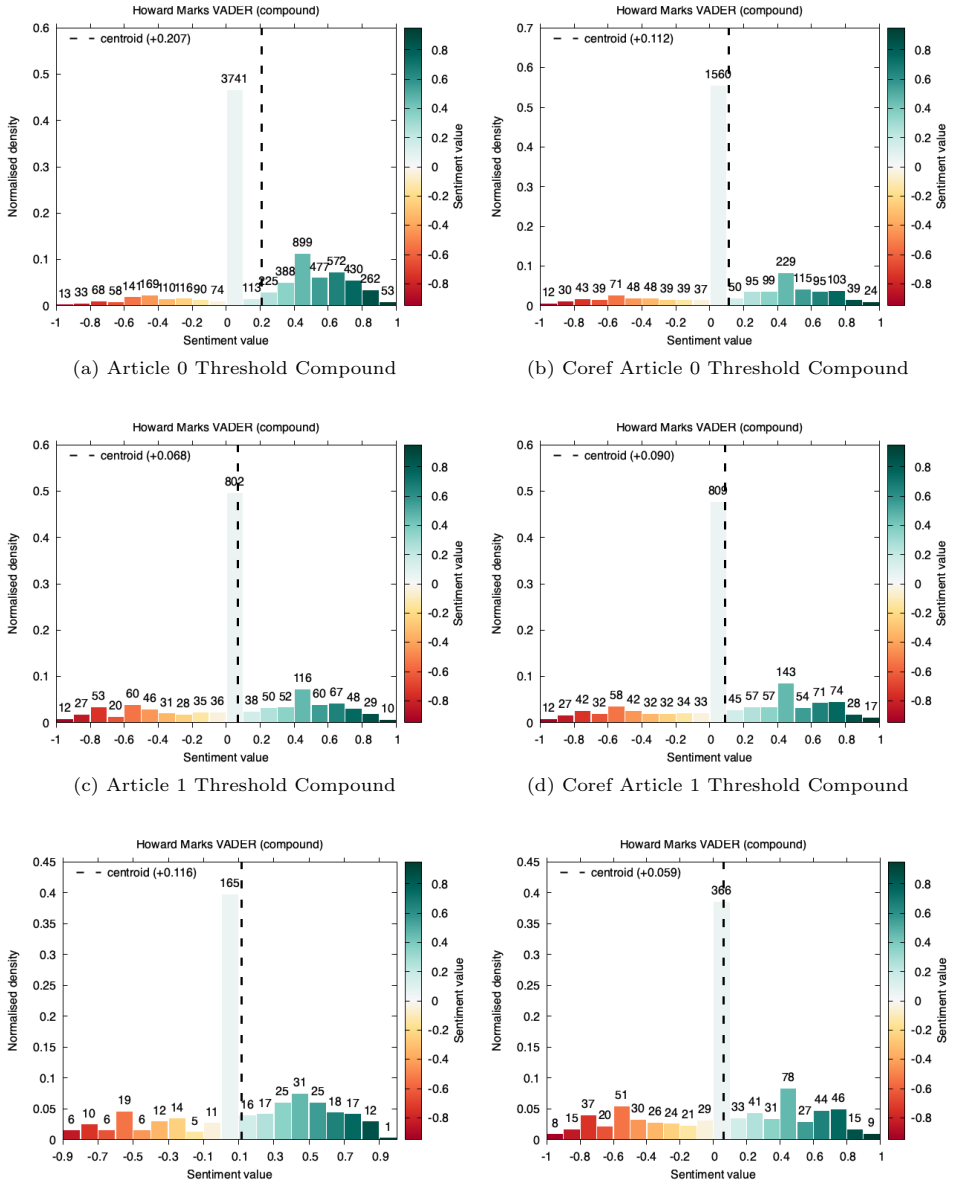
(c) Article 1 Threshold

(d) Coref Article 1 Threshold

(e) Sentence

(f) Coref Sentence

Fig. 10: Distribution of sentiment scores using VADER Simple algorithm corresponding to Howard Marks



(a) Article 0 Threshold Compound

(b) Coref Article 0 Threshold Compound

(c) Article 1 Threshold Compound

(d) Coref Article 1 Threshold Compound

(e) Sentence Compound

(f) Coref Sentence Compound

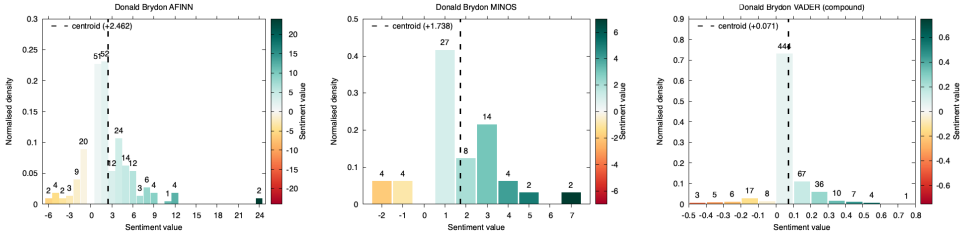
Fig. 11: Distribution of sentiment scores using VADER Compound algorithm corresponding to Howard Marks

4.4 Final Output

This section illustrates the final output of the evaluation system described in this work. As mentioned previously the objective of the system is to make it as easy for an analyst to understand whether an individual is of good character and therefore worthy of an award or not. The results have been divided into groups for awarded, infamous and forfeited individuals as follows. Note, the results shown in this section look a lot more neat and concise for all 3 sentiment algorithms compared to the results before the various filters have been applied (refer Section 4.3). This indicates the efficacy of the filters in weeding out all the spurious data-points.

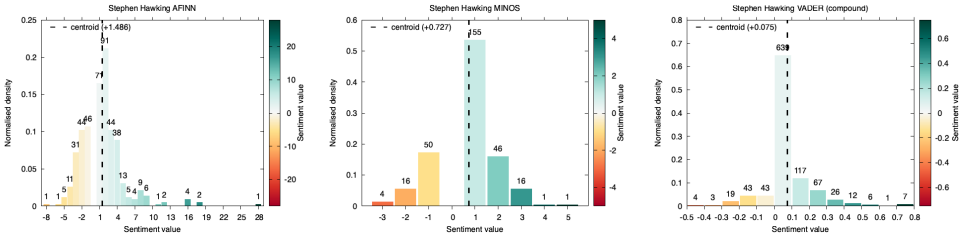
4.4.1 Awarded Individuals

This subsection outlines the results for the awarded individuals. Figure 12 to Figure 13 shows the distribution of the scores for sentences evaluated using the 3 sentiment analysis algorithms respectively. The plots indicate that Minos algorithm presents a much more summarized and clear indication indication regarding the character of the individual rather than a wide spread distribution which can be difficult to assess. In addition, the centroid gives whether person has a positive or negative personality rather being close to zero as in the case of the Vader algorithm or sometimes the Afinn Algorithm. This becomes further evident when compared to the distributions of the infamous individual shown in Figure 18 to Figure 21.



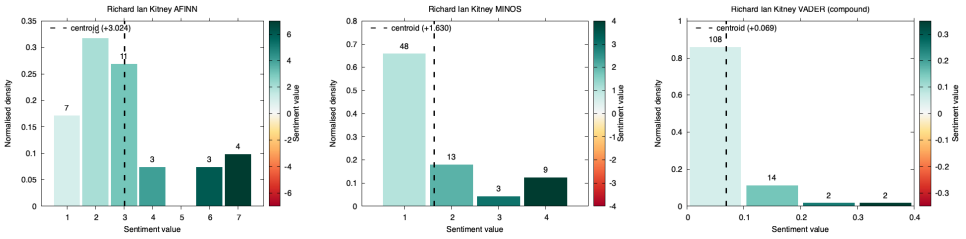
(a) AFINN (b) MINOS (c) VADER

Fig. 12: Donald Brydon



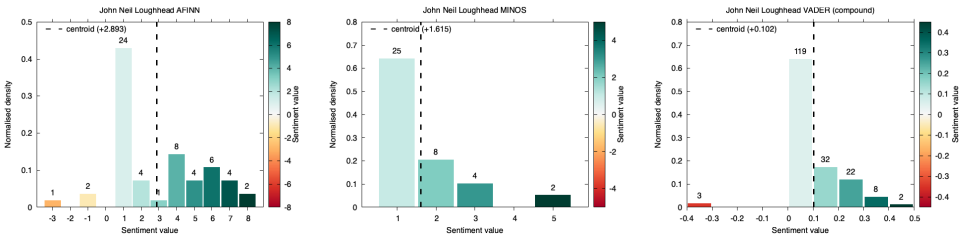
(a) AFINN (b) MINOS (c) VADER

Fig. 13: Stephen Hawking



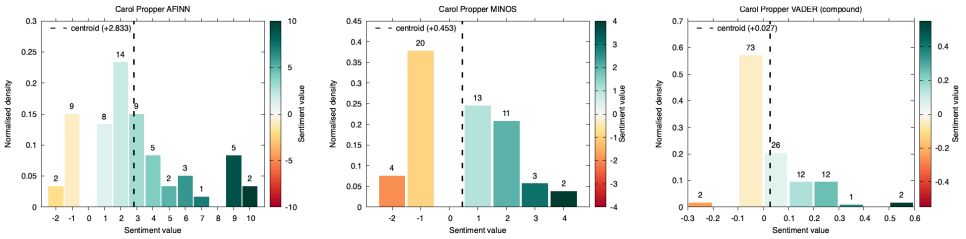
(a) AFINN (b) MINOS (c) VADER

Fig. 14: Richard Ian Kitney



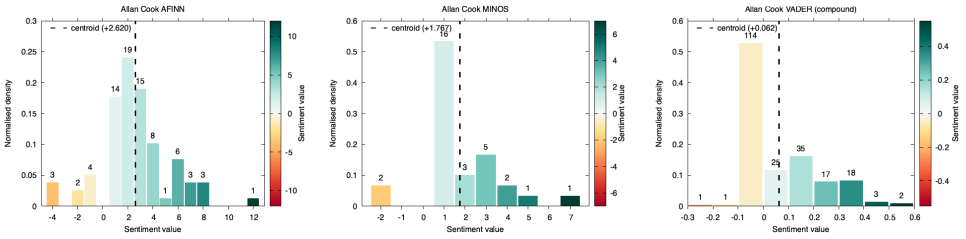
(a) AFINN (b) MINOS (c) VADER

Fig. 15: John Neil



(a) AFINN (b) MINOS (c) VADER

Fig. 16: Carol Propper

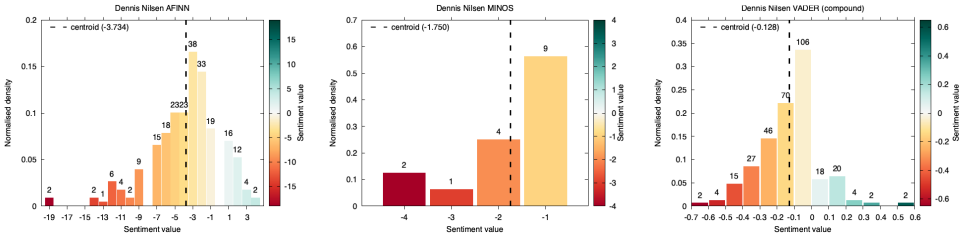


(a) AFINN (b) MINOS (c) VADER

Fig. 17: Allan Cook

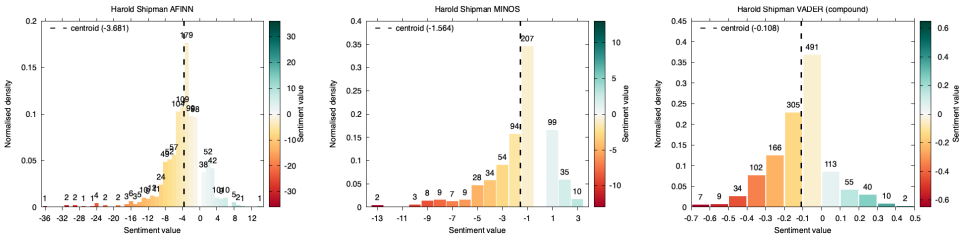
4.4.2 Infamous Individuals

This subsection shows the score distribution plots for infamous individuals who should not receive any award at all. Figure 18 to Figure 21 shows the plots for these individuals. Again Minos is effective in trimming out spurious sentence sentiments which are wrongly scored as positive due to the artifact explained in a previous section. The centroids for these individuals shown by the Minos algorithm is overwhelmingly negative whereas those shown by Afinn or Vader are close to zero.



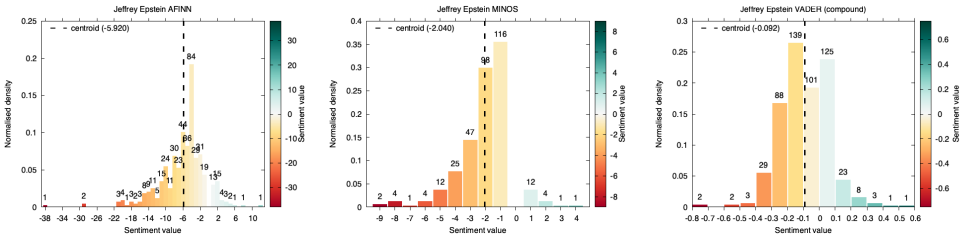
(a) AFINN (b) MINOS (c) VADER

Fig. 18: Dennis Nilsen



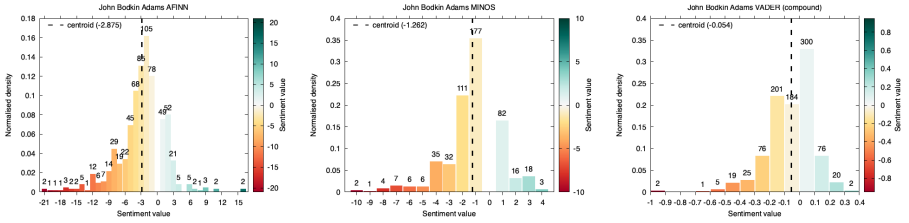
(a) AFINN (b) MINOS (c) VADER

Fig. 19: Harold Shipman



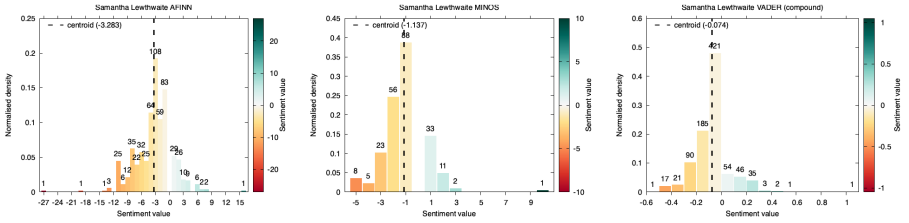
(a) AFINN (b) MINOS (c) VADER

Fig. 20: Jeffrey Epstein



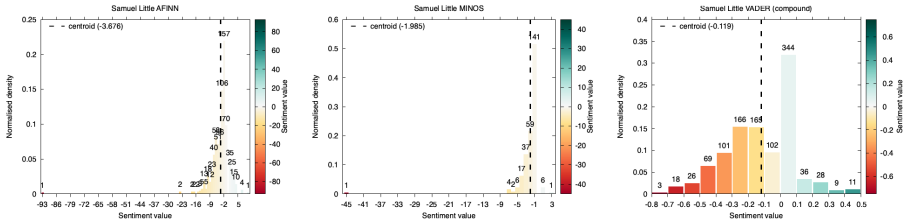
(a) AFINN (b) MINOS (c) VADER

Fig. 21: John Bodkin Adams



(a) AFINN (b) MINOS (c) VADER

Fig. 22: Samantha Lewthwaite

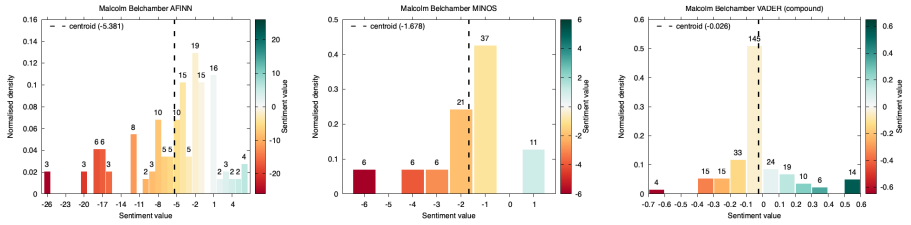


(a) AFINN (b) MINOS (c) VADER

Fig. 23: Samuel Little

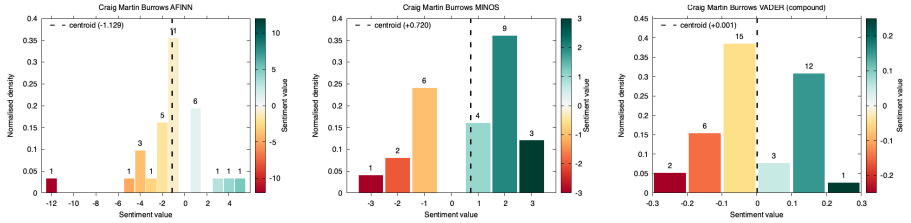
4.4.3 Forfeited Individuals

The sentiment score distribution shown by the MINOS algorithm for forfeited individuals are usual bi-partite. This is due to fact that at some point in time they received a lot of positive press coverage and then at a later date received a lot of negative press. Such distribution can be observed in Figure 24 to Figure 29. The trend is not always as clear for AFINN and Vader algorithms as in Figure 29. It is expected that if any negative press existed before an individual is awarded an honour then the current system would be able to detect it and thereby avoid a future forfeiture.



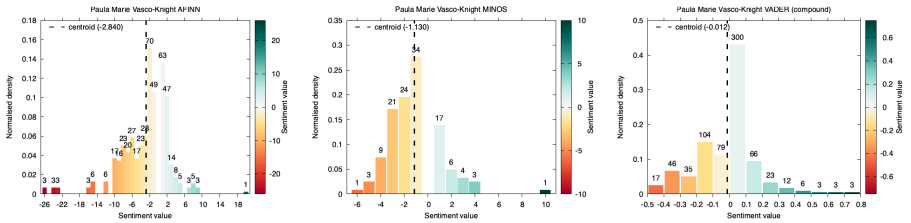
(a) AFINN (b) MINOS (c) VADER

Fig. 24: Malcolm Belchamber



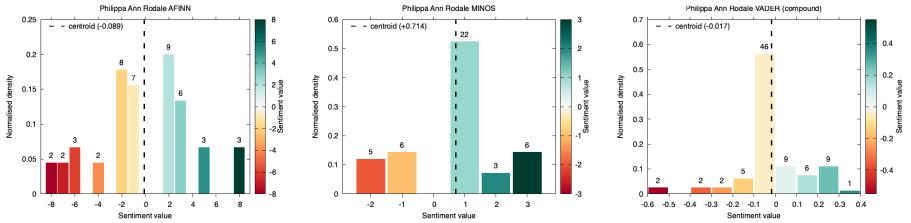
(a) AFINN (b) MINOS (c) VADER

Fig. 25: Craig Martin Burrows



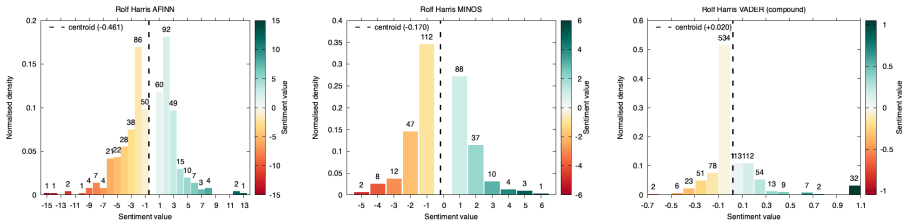
(a) AFINN (b) MINOS (c) VADER

Fig. 26: Paula Marie Vasco-Knight



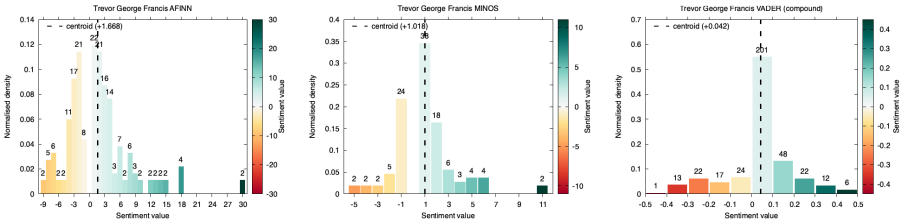
(a) AFINN (b) MINOS (c) VADER

Fig. 27: Philippa Ann Rodale



(a) AFINN (b) MINOS (c) VADER

Fig. 28: Rolf Harris



(a) AFINN (b) MINOS (c) VADER

Fig. 29: Trevor George Francis

5 Conclusions and Future Work

This paper has outlined a data-science approach and system for evaluating the Queen's Award candidates by analyzing publicly available information about them on the Internet. The system scrapes the Internet for publicly available articles on individuals and then analysing these articles to determine their suitability for the award. The data-pipeline created for the algorithm is highly efficient and about 60 candidates can be evaluated in less than an half an hour. A human sifting through the background of a single individual can potentially take much longer than that. Therefore the system is orders of magnitude more efficient than doing the process manually. However, the developed system assumes that a human analyst or award panel would be responsible for going through key information flagged by the system to come up with the final decision.

A novel sentiment analysis algorithm called MINOS has been proposed by this research which is tailored for the current purpose as compared to previous sentiment analysis algorithms like AFINN and Vader. The detailed results shown in this article justify the need for the various steps in the algorithm and their corresponding benefits. The system created by this work finds application in evaluating the backgrounds of candidates not just for the Queen's awards but many such honours and awards which expect high standards of conducts from the awardees. A possible future extension to this work would be the periodic evaluation of upcoming candidates as well as previous awardees such that any risks for forfeiture can be flagged early.

Appendix A Individuals or Subjects of Interest

Table A1: 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

Table A1: (continued)

SOI	Name	Award	Citation	Gender	Year
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 A2: The "test" group of individuals who committed serious crime.

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

[7]The gender proportion in this table is approximately aligned with UK Criminal Justice System staistics on violent and serious crime, e.g. §8 of [45]
[8]Widespread posthumous evidence of fraud.
[9]Warrant issued for charges of possession of explosives and conspiracy to commit a felony.
[10]Deprived of UK citizenship due to links to terrorism and extremism.

Table A3: The "forfeited" group of recipients who were stripped of the Honour.

SOI	Name	Award	Citation	Gender	Year awarded	Forfeiture reason	Year 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

Table A3: (continued)

SOI	Name	Award	Citation	Gender	Year awarded	Forfeiture reason	Year forfeited
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] Honours and Appointments Secretariat: History, (2022). https://honours.cabinetoffice.gov.uk/about/history/

[2] Phillips, S.H.: Review of the Honours System. Crown. https://www.gov.uk/government/publications/review-of-the-honours-system-2004

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

[4] Brief History of Web Scraping. https://webscraper.io/blog/brief-history-of-web-scraping

[5] 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)

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

[7] Hutto, C.J., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. (2015)

[8] Nielsen, F.: A new anew: Evaluation of a word list for sentiment analysis in microblogs (2011)

[9] Storrs, E.: Celebrities on twitter — tweet sentiment analysis with ulmfit (2018)

[10] Silaparasetty, N.: Twitter sentiment analysis for data science using python in 2022

[11] 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)

[12] Public Administration Select Committee: The Honours System: Public Administration Select Committee. https://www.gov.uk/government/publications/pasc-report-on-the-honours-system-2012

[13] Phillips, S.H.: Review of the Honours System 2004. https://www.gov.uk/government/publications/review-of-the-honours-system-2004

[14] The Gazette: Everything you need to know about nominating someone for a UK honour (2022). https://www.thegazette.co.uk/awards-and-accreditation/content/103437

[15] Cabinet Office: How the Honours System Works. https://www.gov.uk/government/publications/how-the-honours-system-works

[16] London Gazette: Birthday and New Year honours lists (1860 to 1939) (2022). https://www.thegazette.co.uk/all-notices/content/100862

[17] London Gazette: Birthday and New Year honours lists (1940 to 2021) (2022). https://www.thegazette.co.uk/honours-lists

[18] Cabinet Office: The First Report on the Operation of the Honours System. https://www.gov.uk/government/publications/operation-of-the-honours-system-2008

[19] Cabinet Office: Second Report on Operation of the Reformed Honours System. https://www.gov.uk/government/publications/second-report-on-operation-of-the-reformed-honours-system

[20] Cabinet Office: Operation of the Honours System 2014. https://www.gov.uk/government/publications/operation-of-the-honours-system-2014

[21] Cabinet Office: Operation of the Honours System 2019. https://www.gov.uk/government/publications/operation-of-the-honours-system-2019

[22] Armstrong, H.: Honours: History and Reviews

[23] OBE, P.A.J., OBE, P.S.M.E.K., Frank, I., CBE, D.S.: I.2 Operation of the Honours System, (2020)

[24] Crown: Types of honours and awards, (2022). https://www.gov.uk/honours/types-of-honours-and-awards

[25] Desktop Search Engine Market Share United Kingdom 2010 - 2023 (2023). https://gs.statcounter.com/search-engine-market-share/desktop/united-kingdom/#yearly-2010-2023

[26] Agarwal, S., Sureka, A., Goyal, V.: Open source social media analytics for intelligence and security informatics applications. In: Kumar, N., Bhatnagar, V. (eds.) Big Data Analytics, pp. 21–37. Springer, Cham (2015)

[27] 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)

[28] Hewitt, J.: What Is "Open Source Intelligence" and How Can Our Police Forces and LEAs Use It to Better Effect? https://synalogik.com/what-is-open-source-intelligence-and-how-can-our-police-forces-and-leas-use-it-to-better-effect/

[29] The Selenium Browser Automation Project (2004). https://www.selenium.dev/

documentation/

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

[31] Hutto, C.J.: cjhutto / vaderSentiment. https://github.com/cjhutto/vaderSentiment

[32] 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) arXiv:cs/0205028

[33] 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 https://direct.mit.edu/coli/article-pdf/32/4/485/1798345/coli.2006.32.4.485.pdf

[34] Pitt, D.E.: Gang attack: Unusual for its viciousness

[35] Wikipedia, t.f.e.: John Bodkin Adams. https://en.wikipedia.org/wiki/John_Bodkin_Adams

[36] Ernst, E.: Profits before ethics: pharmacist continue to recommend and sell bogus treatments. https://edzardernst.com/2017/02/profits-before-ethics-pharmacist-continue-to-recommend-and-sell-bogus-treatments/

[37] Wikipedia, t.f.e.: Trial of Oscar Pistorius. https://en.wikipedia.org/wiki/Trial_of_Oscar_Pistorius

[38] Bowden, M.: The death of osama bin laden: how the us finally got its man

[39] Wikipedia, t.f.e.: Mary Beard (classicist) - Wikipedia. https://en.wikipedia.org/wiki/Mary_Beard_(classicist)

[40] 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)

[41] Gonçalves, P., Araújo, M., Benevenuto, F., Cha, M.: Comparing and combining sentiment analysis methods. In: Proceedings of the First ACM Conference on Online Social Networks. COSN '13, pp. 27–38. Association for Computing Machinery, New York, NY, USA (2013)

[42] Ozdemir, C., Bergler, S.: A comparative study of different sentiment lexica for sentiment analysis of tweets. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 488–496. INCOMA Ltd. Shoumen,

BULGARIA, Hissar, Bulgaria (2015)

[43] Nielsen, F.: fnielsen / afinn. https://github.com/fnielsen/afinn/tree/master

[44] Nielsen, F.: 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/

[45] 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