# Quite Simple Approaches for Authorship Attribution, Intrinsic Plagiarism Detection and Sexual Predator Identification

## Notebook for PAN at CLEF 2012

Anna Vartapetiance, Lee Gillam

University of Surrey
{A.Vartapetiance, L.Gillam}@surrey.ac.uk

**Abstract.** Tasks such as Authorship Attribution, Intrinsic Plagiarism detection and Sexual Predator Identification are representative of attempts to deceive. In the first two, authors try to convince others that the presented work is theirs, and in the third there is an attempt to convince readers to take actions based on false beliefs or ill-perceived risks. In this paper, we discuss our approaches to these tasks in the Author Identification track at PAN2012, which represents our first proper attempt at any of them. Our initial intention was to determine whether cues of deception, documented in the literature, might be relevant to such tasks. However, it quickly became apparent that such cues would not be readily useful, and we discuss the results achieved using some simple but relatively novel approaches: for the Traditional Authorship Attribution task, we show how a mean-variance framework using just 10 stopwords detects 42.8% and could be obtain 52.12% using fewer; for Intrinsic Plagiarism Detection, frequent words achieved 91.1% overall; and for Sexual Predator Identification, we used just a few features covering requests for personal information, with mixed results.

## 1 Introduction

The PAN activity has been around since the 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, and has subsequently evolved to become Uncovering Plagiarism, Authorship, and Social Software Misuse[1]. The first competitive PAN activity in 2009 had two parts, an external task of checking document content against a collection, which largely remains as the Plagiarism Detection task though with differences in approach between 2009 and 2012, and an intrinsic component apparently looking at writing style changes within a document. This intrinsic component is now just one small part of the Authorship Identification track (Tasks E and F), which also includes Traditional Authorship Attribution (Task A, B, C, D, I, J) and Sexual Predator identification. Authorship Attribution requires identifying, based on a sample of given texts, a likely

[1] Presumably the N of PAN now comes from the conjunction.

author – in essence, identifying the closest match to other texts, as author names need not be given. Intrinsic Plagiarism detection involves the separation of text fragments from a single document where fragments are combined from two or more authors. For Sexual Predators, the result comes from a binary classification in which the data of interest relate to those conversations where some are attempting to deceive underage children (mainly) to perform actions of an immoral and potentially illegal nature.

In this paper, we outline the approach taken at the University of Surrey to these quite varied tasks for PAN2012. In section 2, we discuss why current deception detection cues seem to be unsuited for these tasks, which leads us to develop our own approaches. Sections 3, 4 and 5 focus on each of the individual tasks and the results obtained using relatively few features in most cases. Section 6 concludes the paper with considerations for future work.

## 2 Authorship Identification as Deception Detection?

In Vartapetiance and Gillam [1], we discussed why current systems and approaches for deception detection might not be effective in a variety of deceptive situations, one reason being a lack of common data sets upon which to experiment. However, as far as we can tell such approaches had not been explored systematically for PAN. We can readily consider attempts at plagiarism (or copyright infringement) to be deceptive acts, and certainly Sexual Predation would appear to be an attempt to deceive. Hence, the treatment of deception would seem to be relevant to Authorship Identification and vice versa. However, most of the cues for deception (e.g. in DePaulo et al. [2]) are based on non-verbal behaviour (visual and vocal), so are immediately not fit for such purposes. There are then different sets of verbal cues defined by various research groups, though most are covered by three major categories: (1) Overall Impression (2) Quantitative cues and (3) Qualitative cues. Overall impression covers human judgement – does somebody think it is truthful? – with all its subjective responses, restricting us to Quantitative, including word counts and average words per sentence, and Qualitative, that considers features such as self-references and occurrences of negative words. But such kinds of cues appear to be used in yet other measures, for example the Quantitative elements used in readability measures and the Quantitative elements used to determine sentiment polarity. Indeed, some researchers have used readability as an indication of deception in financial reporting, but unreadable text is not necessarily an indication of deceiving so it is important to understand what is being measured and how [3].

Consider, for example, Pennebaker's work, which has been widely used (e.g. [4-6]), it is suggested that deceptive text will have (1) fewer self-references (2) more negative words (3) more exclusive words and (4) fewer motion/action verbs. Pennebaker introduces the Linguistic Inquiry and Word Count (LIWC) system which it is claimed can detect deception based on the same cues [6]. But note that the requirement for detecting deception is contrastive – more or fewer of something. LIWC can offer information about how much of what kinds of words are contained, but something needs to be available against which to make such a contrast. There is also an issue with the point of reference – are such items to be measured for each

document, for each paragraph, for each senteice, or for each sub-clause? How can we have a consistent contrast when we have technical documents, which we expect to have few self-references, reviews of bad products, which we expect to have negative words, and so on? Absent answers to such questions, we explored what might be possible with the online version of LIWC[2]:

1. **Traditional Authorship Attribution:** Measures are mainly qualitative and likely to be context (topic) specific, but there are too few full-text samples to derive useful per-author ranges and readily select an author. Using the data for training text and 12Atest01, the identifiable Author B has used 0.1% self-references, but Author B's samples have values of 9.23% and 11.11%, which would suggest that author A with 0.64% is the closest match. For the various training texts, explorations of such features did not begin to suggest a workable methodology.
2. **Intrinsic Plagiarism:** For task E, since the number of authors is unknown, not only would we need to ascertain where more or fewer was relevant, we would also have to determine how many such distinctions to make. Such an ad hoc approach is unlikely to generalise well.
3. **Sexual Predators**: The conversational nature of the task between the predator (deceiver) and the prey (children) requires an initial separation, but in a number of the conversations there is some indication of the desires of the predator but a difference in intention. Indeed, some predators are certainly not being at all deceptive about what they would like to do, and are happy to use quite a number of self references, social words, and indicate positive emotions.

Following various apparently unsuccessful attempts to make use of such cues, we considered whether such kinds of deception may not be suited to detection using these cues, and looked instead at what we might obtain first from simple features across the data.

## 3 Authorship Attribution: tasks A, B, C, D, I, J

Much literature discusses the use of numerous NLP techniques that operate over bags of words, N-grams, and parts of speech (POS), with varying degrees of success. In many cases, stopwords are either an integral part of the analysis, without consideration for how much they drive the analysis, or are dropped from processing. Prior research in this task does not appear to have addressed whether authors' writing styles and preferred topics lead to distinctive positional preferences for stopwords.

Church & Hanks [7] describe a mean-variance framework for detecting strong associations between co-occurring words and being able to distinguish amongst patterns using this. Their examples are of fixed phrases such as "bread and butter", which demonstrate a clear preference over "butter and bread". As an indicator of style, we explored grammatical preference using a mean-variance framework with just 10 stopwords.

[2] Available at: http://www.liwc.net/tryonline.php

### 4.1 AA, Closed dataset

The approach taken for the *closed* dataset was:

**Step 1** Select the 10 most frequent words from the Oxford English Dictionary:
the, be, to, of, and, a, in, that, have, I

**Step 2** Generate regular expressions for all pairs of these words, e.g. the+have, have+the, and use a specific size of window N (here, 5).

**Step 3** Extract concordances containing the regular expressions for all author texts

**Step 4** Calculate per-author frequency, mean and variance information for the pairs.

**Step 5** Calculate the frequency, mean and variance for the test data (per document) in the same way.

**Step 6** Select the author with closest match values

An example of the values derived for three authors is shown below in Table 1, against text 12Atest01. The selected author is, in essence, decided on by the number of votes cast by matches to frequency, mean and variance as shown in Table 2.

**Table 1.** Example comparison of a document to a set of 3 authors, closest matches displayed in bold font.

| | | A*I | A*And | A*Have | A*In | A*the | … |
|---|---|---|---|---|---|---|---|
| Frequency | A | 5.5 | 20 | --- | 9.5 | **28.5** | … |
| | B | 19.5 | **49** | 1.5 | **16.5** | 25.5 | … |
| | C | 1.5 | 21.5 | --- | 5 | 18.5 | … |
| | 12Atest01 | --- | **49** | --- | **14** | **50** | … |
| | Closest match | --- | B | --- | B | A | … |
| Mean | A | 2.75 | **3.08** | --- | 2.71 | 3.35 | … |
| | B | 2.79 | 2.88 | 3 | 3 | 3.4 | … |
| | C | 3 | 2.69 | --- | **3.33** | **3.7** | … |
| | 12Atest01 | --- | **3.5** | --- | **3.5** | **3.57** | … |
| | Closest match | --- | A | --- | C | C | … |
| Variance | A | 0.19 | 0.69 | --- | 0.49 | **0.46** | … |
| | B | 0.6 | 0.63 | 0 | **0.73** | 0.24 | … |
| | C | 0 | **0.59** | --- | 0.89 | 0.21 | … |
| | 12Atest01 | --- | **0.25** | --- | **0.75** | **0.39** | … |
| | Closest match | --- | C | --- | B | A | … |

**Table 2.** Example counts, three potential authors with author B selected on totals

| | A | B | C |
|---|---|---|---|
| Frequency | 19 | 54 | 10 |
| Mean | 22 | 41 | 20 |
| Variance | 20 | 63 | 63 |
| Sum | 61 | ***158*** | 93 |

### 4.2 AA, Open dataset

For the *open* dataset, to account for data not belonging to any of the authors in the training set, we used a simple confidence measure: if the count difference between the 1st and 2nd highest values is less than 5, it is reported that there is no author. Table 3 shows an example where matches have been made to different authors but with insufficient confidence (difference = 3).

**Table 3.** Example count, with no author (NA) selected

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Average Frequency | 5 | 7 | 15 | 17 | 20 | 13 | 5 | 6 |
| Mean | 13 | 6 | 15 | 14 | 13 | 10 | 8 | 9 |
| Variance | 20 | 8 | 10 | 13 | 8 | 13 | 6 | 10 |
| Sum | 38 | 21 | 40 | 44 | 41 | 36 | 19 | 25 |

### 4.3 Results

Results from PAN2012 show that this method achieves 41%, flagging 28 out of 71 documents correctly. In post-competition analysis, we investigated effects of changing the gaps size (5, 10 and 25), changing the confidence (2, 3, 5, 10) and looking at subsets of the 10*10 stopword combinations (four directional subsets of 5*5, denoted as pairs of S1, S2). Table 4 shows results those comparable to or better than our competition result. Of particular interest is that:

- Judicious use of the 5*5 performs better, with the same threshold of 5 (case 3 and 4).
- Patterns starting with S2 did not help detection
- For open datasets, lower threshold seems to work better
- Best results would have been achieved with S1*S1 for closed and S1*S2 for open data sets; improving the results by almost 10%

**Table 4.** Post-competition investigations into the approach/parameters.

| | | A | B | C | D | I | J | A | B | C | D | I | J | Overall | Corr. | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Correct | 6 | 10 | 8 | 17 | 14 | 16 | % | % | % | % | % | % | % | % | 71 |
| **1** | **AF-3-S1*S1/S1*S2** | **5** | **6** | **4** | **10** | **5** | **4** | **83** | **60** | **50** | **59** | **36** | **25** | **52.15** | **47.89** | **34** |
| 2 | AF-5-S1*S1/S1*S2 | 5 | 6 | 4 | 11 | 5 | 2 | 83 | 60 | 50 | 65 | 36 | 13 | 51.04 | 46.48 | 33 |
| 3 | AF-5-S1*S2 | 5 | 3 | 4 | 8 | 5 | 4 | 83 | 30 | 50 | 47 | 36 | 25 | 45.18 | 40.85 | 29 |
| 4 | AF-5- S1*S1 | 4 | 6 | 1 | 11 | 6 | 2 | 67 | 60 | 13 | 65 | 43 | 13 | 43.2 | 42.25 | 30 |
| 5 | Surrey | 4 | 6 | 1 | 3 | 7 | 8 | 67 | 60 | 13 | 18 | 50 | 50 | 42.8 | 40.85 | 29 |

(1) using S1*S1 for closed dataset and S1*S2 with threshold of 3 or more for open dataset
(2) using S1*S1 for closed dataset and S1*S2 with threshold of 5 or more for open dataset
(3) using S1*S2 for all dataset with threshold of 5 or more for open dataset
(4) using S1*S1 for all dataset with threshold of 5 or more for open dataset

## 5 Intrinsic Plagiarism Detection: task E and F

Depending on the approach, Intrinsic Plagiarism Detection might be categorized as Authorship Attribution - it can be related to identifying parts of a text least likely to have been produced by the current author. However, there are various differences: for Authorship Attribution, (1) texts are usually longer, (2) there are training samples (3) two long texts are usually compared (4) the boundaries for comparison are known (5) decisions are usually for an individual. But this is not necessarily true for Intrinsic Plagiarism which is often (1) in short sections (2) only internally comparable (3) with unknown boundaries (4) with unknown number of plagiarised sections (5) with many possible decisions. For these reasons, the approach outlined above would not usefully flag the plagiarized content – here, paragraphs. Instead, we looked to a new approach, starting with Task F as it was mentioned that task E might have more than 2 authors.

### 5.2 Intrinsic Plagiarism, task F

The approach taken for Task F was:

**Step 1** Select the 50 most frequent words from the file, after removing stopwords.

**Step 2** Determine frequency by paragraph for these 50 words

**Step 3** Select (sequences of) paragraphs with fewer similarities (e.g. < 10)

If there is more than one sequence:

**Step 3a** Select the longest sequences of paragraphs which do not share the most frequent word, and have the lowest average frequency for top 5 of these 50 words

Table 5 shows paragraphs flagged by total frequency below 10. These sequences' average frequency for the top 5 shows that the [P4, P5, P6] sequence is least relevant to the file.

For 12Ftest02, steps 1-3 identify P01, P04 and [P06, P07, P08]. Calculating with the 5 most frequent words suggested [P06, P07, P08] to be the sequence. However, all similarly shared the most frequent words suggesting that they are all related to the topic. We allocated "no author" to that file, even though it was not suggested in the competition that the dataset could have open answers.

**Table 5.** Example of task F detection process

| p | ALL | P01 | P02 | P03 | P04 | P05 | P06 | P07 | … | P11 | P12 | P13 | P14 | P15 | P16 | P17 | P18 | P19 | P20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| id | 18 | --- | 3 | 2 | --- | --- | --- | 3 | … | --- | --- | 1 | --- | 2 | 1 | --- | 2 | --- | 1 |
| time | 13 | --- | --- | --- | --- | 3 | 3 | --- | … | 2 | 2 | --- | --- | --- | --- | 1 | --- | --- | 1 |
| back | 12 | --- | 2 | --- | 1 | --- | --- | --- | … | --- | --- | --- | 1 | 1 | 2 | --- | 2 | 1 | 1 |
| made | 11 | 1 | --- | --- | 1 | 1 | --- | --- | … | --- | --- | --- | 2 | 1 | 1 | --- | 2 | --- | --- |
| bowker | 11 | --- | --- | --- | --- | --- | --- | 1 | … | --- | 1 | --- | 1 | 2 | --- | --- | --- | --- | --- |
| … | … | … | … | … | … | … | … | … | ... | … | … | … | … | … | … | … | … | … | … |
| Total frequency | 50 | 10 | 13 | 10 | 7 | 7 | 9 | 14 | … | 12 | 8 | 8 | 8 | 10 | 7 | 8 | 9 | 7 | 8 |
| Frequency of 5 | | | | | 2 | 1 | 1 | | | | 2 | 1 | 3 | | 3 | 1 | 3 | 1 | 3 |
| Average | | | | | 1.67 | | | | | | 2 | | | | 2.2 | | | | |

**Table 6.** Example of task E detection process

| | Fqall | P01 | P09 | P11 | P14 | P16 | P18 | P22 | P25 | P26 | P28 | P30 | | P04 | P07 | P08 | P12 | P13 | P15 | P17 | P21 | P29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A1 | | | | | | | | | | | | A2 | | | | | | | | |
| john | 13 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| johnson | 11 | NA | NA | NA | NA | NA | NA | NA | NA | 1 | NA | NA | | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| simon | 7 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| rizzo | 6 | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | 1 | | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| jan | 6 | 1 | 1 | 2 | NA | NA | NA | NA | NA | NA | 2 | NA | | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| correct | | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A1 | A2 | A1 | A3 | | A2 | A2 | A2 | A2 | A2 | A2 | A2 | A2 | A2 |
| | | | | | | | | | | | | | | | | | | | | | | |
| | Fqall | P03 | P05 | P06 | P19 | P23 | P24 | P27 | | P02 | P10 | P20 | | | | | | | | | | |
| | | | | | | | | A3 | | Unknown | | | | | | | | | | | | |
| john | 13 | NA | NA | NA | NA | NA | NA | NA | | NA | NA | NA | | | | | | | | | | |
| johnson | 11 | NA | NA | NA | NA | NA | NA | NA | | NA | NA | NA | | | | | | | | | | |
| simon | 7 | 1 | 2 | 2 | NA | NA | 2 | NA | | NA | NA | NA | | | | | | | | | | |
| rizzo | 6 | NA | NA | 1 | 1 | 1 | 1 | 1 | | NA | NA | NA | | | | | | | | | | |
| jan | 6 | NA | NA | NA | NA | NA | NA | NA | | NA | NA | NA | | | | | | | | | | |
| correct | | A3 | A3 | A3 | A3 | A3 | A3 | A3 | | A2 | A3 | A2 | | | | | | | | | | |

### 5.2 Intrinsic Plagiarism, task E

The approach to Task F would not distinguish as readily. We adapted this for task E as follows:

**Step 1** As task F

**Step 2** As task F

**Step 3** Extract the nouns from the 50 most frequent words (excluding stopwords)

**Step 4** For the highest frequency noun, create a cluster and remove from consideration all other nouns enclosed by this – i.e. occurring in the same paragraphs. Repeat this step to produce new clusters from the remaining nouns.

Where paragraphs are not allocated to a cluster:

- If the number of consecutive unallocated paragraphs is greater than 5, these form a new cluster.
- For others: (a) paragraphs between two in the same cluster are allocated to that cluster; (b) paragraphs between different clusters are allocated to the subsequent cluster.

A sample of the process and the results from one of the test sets in presented in Table 6, with errors highlighted with gray.

To validate our reasoning behind not using the Task F approach, we used the 5 most frequent words (from task F) and used Step 4 of task E to cluster them into groups. Interestingly, for 12Etest01 this only mis-classified P19 as Author 2, but we have mis-classified 2 for our final submission. However, such a gain would have come at the cost of losses elsewhere.

### 4.3 Results

Competition results show these approaches gave 100% (Task F) and 82.2% (Task E) accuracy, a simple average of which would make us 2nd in just this task (91.1% against 94.2%). We are now looking further at what might have improved methods to improved performance in Task E without introducing complexity.

**Table 6.** Result for PAN2012 for Intrinsic Plagiarism detection only

| TEAM | E % | F % | Overall | Docs corr. |
|---|---|---|---|---|
| EVL Lab | 92.22222 | 96.25 | 94.23611 | 94.11765 |
| ***Surrey*** | ***82.22222*** | ***100*** | ***91.11111*** | ***90.58824*** |
| CLLE-ERSS 1 | 73.33333 | 93.75 | 83.54167 | 82.94118 |

## 6 Sexual Predator Detection

Just like other sub-tasks in this paper, Sexual Predator actions carry a level of deception which may or may not lie in the text but with the actions followed. Since we have never attempted the analysis of such a corpus or topic previously, we have taken relatively straightforward approach, and with reference to the training corpus this appears to offer good performance (up to f1=0.66), but for which we would have concern over the rate of false negatives as we discuss later.

### 6.1 Process of Sexual Predators Identification

As it was the first time we have attempted such a task, we randomly took 10 predators IDs from the training set to set about manually discovering patterns. We found obvious similarities, and classified these as described below (examples are shown below in Table 7):

**Address**: asking for the address of the house or somewhere close to drive to in other to meet up. Mostly, it's the predator who asks the question about the child's address and it was rare for children to ask whether the predator would need/like their address. This alone detects 58 out of 142 in the training data, appearing more than once, and 28 times it appears twice or more with very high precision (85%).

**Parents**: another strong feature. Questions about parents are usually because of:

- **Secrecy**
  - Making sure children are alone while chatting
  - Making sure the chat history will be deleted later
  - Saying nothing to their parents
- **Seclusion**
  - To determine whether parents are around
  - To ascertain how long they would be gone for

This feature detects 84 out of 142, when appearing once or more, and 49 when appearing twice or more. Combining "address" and "parents" would detect 105 and 74 respectively.

**Age**: Some predators might lie about their age but most seem quite open about their age. They would usually highlight the fact that they are older, wishing the child were older, and this is mainly to retain the secrecy.

**Intention:** Interestingly, not many of these chats have direct references to their sexual intentions. They usually focus on the concept of meeting up and having fun time, watching TV, listen to music and have some alcohol. In some cases there may be mentions of what they would like to do – these are mostly limited to *cuddles and kisses* and so sexual activities are more apparent.

**Table 7.** Bases for accept and reject files

| | | | |
|---|---|---|---|
| Address | Accept | 13 | Different spelling combination of following words: "your addres", "ur addres", "the addres" |
| | Reject | 78 | IT and social networking related topics such as URL, Gmail Facebook, email, e-mail, IP, Browser, … |
| Parents | Accept | 11 | Different spelling combination of following words: "your mom", "your dad", "your Parent" |
| | Reject | 26 | Reference to parents' objects or characteristics such as "Ur dads car", "Your mom's face", "Your mom is nice, young, etc".<br>IT related topics such as "Parent Class" |
| Age | Accept | 11 | Different spelling combination of following words: "you are young", "get in trouble", "underage", "to jail", "wish you were" |
| | Reject | 33 | Self-reference such as "I'm underage"<br>Reference to the others such as sister, brother, friend<br>Excluding, "wish you were here /with me" |
| Intentions | Accept | 6 | Different spelling combination of following words: "go down on you", "make you come" |

We tested all of the above mentioned categories individually, varying the number of occurrences, and in various combinations, with the PAN2012 training data. Our analysis of these results is shown in Table 8.

**Table 8.** Comparing the results from combining different elements for detection.

| # of Occurrence | Flagged | Unique | Correct | FP | FN | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|
| **Address Cues Category** | | | | | | | | |
| Once or more | 159 | 117 | 58 | 59 | 84 | 0.5 | 0.41 | 0.45 |
| Twice or more | 74 | 33 | 28 | 5 | 114 | 0.85 | 0.20 | 0.32 |
| Three times or more | 18 | 9 | 8 | 1 | 134 | 0.89 | 0.06 | 0.11 |
| **Parents Cues Category** | | | | | | | | |
| Once or more | 440 | 255 | 84 | 172 | 58 | 0.33 | 0.59 | 0.42 |
| Twice or more | 257 | 72 | 49 | 24 | 93 | 0.68 | 0.35 | 0.46 |
| Three times or more | 151 | 38 | 32 | 6 | 110 | 0.84 | 0.23 | 0.36 |
| **Age Cues Category** | | | | | | | | |
| Once or more | 124 | 88 | 33 | 55 | 109 | 0.38 | 0.23 | 0.29 |
| Twice or more | 62 | 25 | 17 | 8 | 125 | 0.68 | 0.12 | 0.20 |
| Three times or more | 21 | 10 | 9 | 1 | 133 | 0.90 | 0.06 | 0.12 |
| **Intentions Cues Category** | | | | | | | | |
| Once or more | 39 | 35 | 14 | 21 | 128 | 0.40 | 0.10 | 0.16 |
| Twice or more | 8 | 5 | 4 | 1 | 138 | 0.80 | 0.03 | 0.05 |
| Three times or more | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Combining two Cue Categories of Address and Parents** | | | | | | | | |
| Once or more | 598 | 333 | 105 | 228 | 37 | 0.32 | 0.74 | 0.44 |
| Twice or more | 366 | 101 | 74 | 27 | 68 | 0.73 | 0.52 | 0.61 |
| Three times or more | 217 | 53 | 46 | 7 | 96 | 0.87 | 0.32 | 0.47 |
| **Combining three Cue Categories of Address, Parents and Age** | | | | | | | | |
| Once or more | 722 | 388 | 112 | 276 | 37 | 0.29 | 0.79 | 0.42 |
| Twice or more | 458 | 124 | 85 | 39 | 57 | 0.69 | 0.60 | 0.64 |
| Three times or more | 280 | 69 | 58 | 11 | 84 | 0.84 | 0.41 | 0.55 |
| **Combining all four Categories together** | | | | | | | | |
| Once or more | 761 | 410 | 113 | 297 | 29 | 0.28 | 0.80 | 0.41 |
| Twice or more | 478 | 126 | 88 | 38 | 54 | 0.70 | 0.62 | 0.66 |
| Three times or more | 298 | 72 | 62 | 10 | 80 | 0.86 | 0.44 | 0.58 |
| **Main Test Data** | | | | | | | | |
| Twice or more | 630 | 159 | 97 | | | 0.61 | 0.38 | 0.48 |

### 6.2 Results

For the competition, we used the combination of all four categories that occurred twice or more, as this offered the optimal f1 score on training data (precision=0.7, recall=0.62 and F1=0.66).

However, it can easily be argued that for real detection the false negatives would be of particular concern. Let us consider an application that might filter out possible predatory conversations. What characteristics would a parent rather have: stringent filtering of suspicious behaviour but with a high false positive rate so that some genuine conversations are dropped, or conversations with real predators that may remain and appear acceptable? The competition webpage suggests that the decision

on efforts put on investigation are the dominion of the system “to optimize the time of a police agent towards the "right" suspect rather than "all" the possible suspects”. This presents a disturbing view of such a system, and we would strongly contend that the system should produce results ranked according to confidence, but resourcing judgements should be left to the conscience of human beings who are fully aware of the consequences of such missed results. Recall, then, should be a higher priority – F2, not F0.5 – even if this would reduce our placing in this task.

To see if we could have improved our results, post submission but before results, we also tested the combination of best f1 scores of all categories on the training dataset. The result for “one occurrence” across all 4 classes increased from 0.42 to 0.58 because Parent was already based on two occurrences, but the two and three occurrence scores decrease respectively to 0.6 and 0.48. Currently we are looking at other ways to increase the detection rate while keeping the simplicity of the method. We have already found a few words that can improve the detection, especially in case of sexual comments related to “intentions”.

Checking the ground truth of second section raised some questions for us based on the number of lines selected and the content. For example, lines such as:

| | | |
|---|---|---|
| 0fe0367fc3735101fbf7aa3df1cb9f4e | 37 | what grade u in |
| 6bf9b33a9f4ae1df54cb89831eac1be2 | 5 | :) |
| 94c71d9e905c390d310f3f315f9c7b19 | 41 | i promise |
| 94c71d9e905c390d310f3f315f9c7b19 | 45 | age |

## 7 Conclusion

We attempted, for the first time, the Author Identification track at PAN2012, participating in all three tasks of Authorship Attribution, Intrinsic Plagiarism detection and Sexual Predator Identification. We attempted to use fairly simple approaches in each case to determine the extent to which these might be effective, and believe there is some degree of novelty presented in each approach. It was surprising, for example, that just 5 stopwords with a mean-variance framework might be able to produce reasonable performance in Authorship Attribution, given that it is essentially a stylometric approach.

Our initial intention was to determine how the supposed cues of deception might be useful against benchmark data collections, but we have seen little indication of relevance to these kinds of deception. What we have learnt from participating in these tasks can now be applied back to reported deception experiments elsewhere to see whether deception is indicated by features which are not usually in the set selected by the researchers – positive evidence being so much easier to discern.

Our best results appear to have been obtained against Intrinsic Plagiarism, which was very much an 11$^{th}$ hour effort. However, plenty of room for improvement remains against the other tasks, and the generalisability of our approaches can now be evaluated across previous PAN datasets also.

## References


1. Vartapetiance, A., Gillam, L.: “I don't know where he's not”: Does Deception Research yet offer a basis for Deception Detectives?: Proceedings of the Workshop on Computational Approaches to Deception Detection, pp. 3-14, Avignon, France (2012)
2. DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K., Cooper, H.: Cues to Deception: Psychological Bulletin, vol. 129(1), pp. 74-118 (2003)
3. Moffitt, K., Burns, M.B.: What Does That Mean? Investigating Obfuscation and Readability Cues as Indicators of Deception in Fraudulent Financial Reports: Fifteenth Americas Conference on Information Systems, San Francisco (2009)
4. Little, A., Skillicorn, B.: Detecting Deception in Testimony: Proceeding of IEEE International Conference of Intelligence and Security Informatics (ISI 2008), pp. 13-18, Taipei, Taiwan (2008)
5. Gupta, S., Skillicorn, D.: Improving a Textual Deception Detection Model: Proceedings of the 2006 Conference of the Center for Advanced Studies on Collaborative Research, pp. 1-4, Toronto, Canada (2006)
6. Newman, M.L., Pennebaker, J.W., Berry, D.S, Richards, J.M.: Lying Words: Predicting Deception from Linguistic Styles: Personality and Social Psychology Bulletin, vol. 29(5), pp. 665-675 (2003)
7. Church, K., Hanks, P.: Word Association Norms, Mutual Information and Lexicography: Computational Linguistics, vol. 16(1), pp. 22-29 (1991)