



Academic Dishonesty or Academic Integrity? Using Natural Language Processing (NLP) Techniques to Investigate Positive Integrity in Academic Integrity Research

Thomas Lancaster¹

Accepted: 24 May 2021 / Published online: 5 June 2021
© The Author(s) 2021

Abstract

Is academic integrity research presented from a positive integrity standpoint? This paper uses Natural Language Processing (NLP) techniques to explore a data set of 8,507 academic integrity papers published between 1904 and 2019.

Two main techniques are used to linguistically examine paper titles: (1) bigram (word pair) analysis and (2) sentiment analysis. The analysis sees the three main bigrams used in paper titles as being “academic integrity” (2.38%), “academic dishonesty” (2.06%) and “plagiarism detection” (1.05%). When only highly cited papers are considered, negative integrity bigrams dominate positive integrity bigrams. For example, the 100 most cited academic integrity papers of all time are three times more likely to have “academic dishonesty” included in their titles than “academic integrity”. Similarly, sentiment analysis sees negative sentiment outperforming positive sentiment in the most cited papers.

The history of academic integrity research is seen to place the field at a disadvantage due to negative portrayals of integrity. Despite this, analysis shows that change towards positive integrity is possible. The titles of papers by the ten most prolific academic integrity researchers are found to use positive terminology in more cases than not. This suggests an approach for emerging academic integrity researchers to model themselves after.

Keywords Academic integrity · Academic misconduct · Positive integrity · Academic dishonesty · Academic honesty · Student cheating

Background

Academic integrity research has been published dating back to at least to the 1900s. Academic integrity publications span disciplines and research is published under a variety of different themes. Barnes (1904) published the paper “*Student honor: A study in cheating*”, providing an early example of using surveys to conduct qualitative and quantitative academic integrity research. Although Barnes’ research may have not been presented with the

✉ Thomas Lancaster
thomas@thomaslancaster.co.uk

¹ Department of Computing, Imperial College London, London, UK

rigour of some more recent academic integrity investigations, it still identified themes that are commonly discussed in the field today.

In Barnes (1904) students were asked how they would respond to an academic integrity scenario that had occurred in real life and for which a variety of academic penalties were available. In the scenario, examination questions were said to have been stolen, giving the students who had to access to them an unfair advantage over their peers. The responses received included disparate opinions from students regarding whether it was their business to get involved further. Thirty percent of male students and 35% of female students felt that reporting was necessary so that they would not be unfairly judged against other students who they now expected would benefit from better results. Barnes also noted differences in responses between genders, a theme which is still regularly investigated as part of academic integrity research to this day.

Barnes' choice of paper title does rather seem to present their own position on academic integrity issues. This is perhaps most clearly summed up in this quote from the paper:

"The reasons are mainly selfish; the university's interests are far less important than self-protection; while general social responsibility is comparatively little felt."

Despite being only six words long, the paper title "*Student honor: A study in cheating*" brings together two words at different ends of the academic integrity spectrum, *honor* and *cheating*. The word *honor* carries with it an expectation that students will act with positive integrity. The word *cheating* has negative connotations, with the suggestion of students getting an unfair advantage. A focus on transgressive behaviour is not necessarily wrong, but also leads to missed opportunities for people working in the academic integrity field. The conflict between whether academic integrity should be framed in a positive or negative manner still exists in paper titles today and is the focal point of the research investigation presented in this paper.

This paper uses Natural Language Processing (NLP) techniques to provide a data-driven investigation into how academic integrity paper titles have been constructed between 1904 and 2019. The research presented examines the titles of 8,507 papers published in the wider academic integrity field and is used to see how far such titles are presented to readers using a positive or negative approach. The results are intended to help the academic integrity research field to determine if it wishes to present itself from a more positive direction.

Academic Integrity

The literature on academic integrity often considers this field through both positive and negative viewpoints, with integrity itself considered as a positive term. A look at the different opinions and presentations of this research is useful to help define how the field is changing, as well as to allow positive integrity and negative integrity ideas to be demonstrated through representative examples.

The popularisation of the term academic integrity is commonly attributed to McCabe. Despite this, in the single most highly cited paper in the field "*Academic dishonesty: honor codes and other contextual influences*", McCabe and Trevino (1993) do not discuss academic integrity, but instead academic dishonesty. In the paper, McCabe and Trevino collected data using a survey methodology and discussed how this could be used to predict academic dishonesty. Despite its high citation level, the focus of both the paper title and content brings connotations of a negative presentation of integrity.

Similar observations to those made about McCabe and Trevino (1993) also appear in a literature review by Macfarlane et al. (2014). They examined 115 articles in the field across both Western and Chinese literature. Their review concluded that academic integrity is commonly defined by reference to misconduct, fraud and corruption. This paper will consider research with a focus on areas such as these as being representative of negative integrity.

An alternative group of approaches are possible. This paper will consider such approaches as representative of positive integrity, often represented by the pure term academic integrity. Macfarlane et al. (2014) define academic integrity as “*the values, behaviour and conduct of academics in all aspects of their practice*”. An alternative definition, given by East and Donnelly (2012) based on the values of the institution they work for is “*academic integrity means being honest in academic work and taking responsibility*”. That interpretation is close to how sector organisation the International Centre for Academic Integrity (ICAI) present this. ICAI take a positive integrity view and define this concept in terms of core values by asking members to commit to “*honesty, trust, fairness, respect and responsibility*” (Fishman, 2014).


Fishman (2016) discusses the variety of frameworks which academic integrity is presented under in the United States. These include moral and ethical frameworks, pedagogical frameworks, legalistic frameworks, comparing academic integrity with criminal behaviour and even considering this as a form of disease. Although these frameworks provide some opportunity for a positive discussion, the most immediate interpretation is that academic integrity should be viewed through a negative lens.

There have been opportunities for the negative viewpoint to change. McCabe and Pavela (2004) discuss principles they believe will help build a culture of academic integrity, such as making this an institutional value with consistent standards, clarify expectations with students, enabling students to take responsibility and ensuring fair assessment. How academic integrity principles are taught to students and how far teaching can take a positive approach continues to be an important part of the modern discussion (Ransome & Newton, 2018; Sefcik et al., 2020).

One underlying principle regarding making academic integrity work at an institutional level is that it should apply to the whole academic community, not just to students and not just to academics. The student voice is being increasingly considered as an essential and important part of this discussion (Pitt et al., 2020).

The fields of research studied within academic integrity have widened in recent years, with new areas developing as a result of observing threats to academic integrity. Some identified challenges include cybersecurity threats (Dawson, 2020), contract cheating (Clarke & Lancaster, 2006), study helper websites (Harrison et al., 2020) and paraphrasing tools (Prentice & Kinden, 2018). The positioning of research discussing threats to integrity and opportunities for student misconduct suggest a continuing view of negative integrity. The fast pace of technological change and the need to raise awareness of this further suggest that a certain level of negative integrity research will always be required within the field.

The widening of the academic integrity research field and the growth of technology has brought with it the opportunity for innovation in how academic integrity research is conducted. Methodologies have moved beyond surveys. Social media analysis can be used to investigate why students cheat (Amigud & Lancaster, 2019). Region and sector specific literature reviews are possible (Eaton & Edino, 2018). Internal academic conduct records can be analysed (Atkinson et al., 2019). Others have had success working

Fig. 1 Data set formation pipeline 

around analysing existing policies (Eaton et al., 2020). There is plenty of alternative data available that can be examined.

This paper takes such an alternative and data-driven research approach. It considers existing data relating to published academic integrity research and uses NLP techniques to programmatically examine this data.

For the purpose of this paper, the view that academic integrity applies to everyone is supported, but this is balanced by the observation that papers are most relevant if they fit within an educational setting. As such, the interpretation considers academic integrity as it applies to teaching, learning, pedagogy and education, where students, academics and professional university staff are at the forefront of the conversation. The related field of research integrity is sometimes included with academic integrity, but to avoid muddying the water it is only included in the investigations reported here when this also relates to education.

Although this paper makes no attempt to provide a fresh definition of academic integrity, the approach taken naturally identifies papers with examples of both positive integrity and negative integrity. One side product of looking at both positive and negative views is that it is hoped the range of papers, topics and issues identified will help to inform future definitions of academic integrity so they can both be current and complete.

Investigative Methodology

Formation of the Primary Data Set

The research presented in this paper relies on a data-driven approach. Data was collected in May 2020 to form a primary data set. From this four further secondary data sets were derived.

The procedure through which the data sets were gathered and processed employed standard techniques from the domains of NLP and machine learning. As is customary in this field, experimentation was undertaken on the initial data to determine how best to present it for NLP. Some of the final decisions presented here may appear arbitrary, but they were made to fine tune the results for readability and accuracy. The pipeline is presented to provide enough information for researchers looking to undertake related studies, whilst recognising this paper is aimed at the academic integrity field, an audience who may be unfamiliar with NLP.

Google Scholar was used as the primary data source to identify academic integrity research publications. Data was collected from Google Scholar through an iterative process, with the aim of ensuring data set completeness and consistency. Both manual and automated checks and corrections were made on the resulting data. Excel and Python were used extensively to support data collection and processing with several scripts developed for internal use. The NLP aspects of processing relied heavily on the NLTK platform. Sentiment analysis, a process where the subjective information in a written expression is evaluated to identify the tone of the expression, was completed using the VADER toolkit. Both NLTK and VADER are open source.

Figure 1 provides a high-level overview of the data collection, cleansing and processing pipeline.

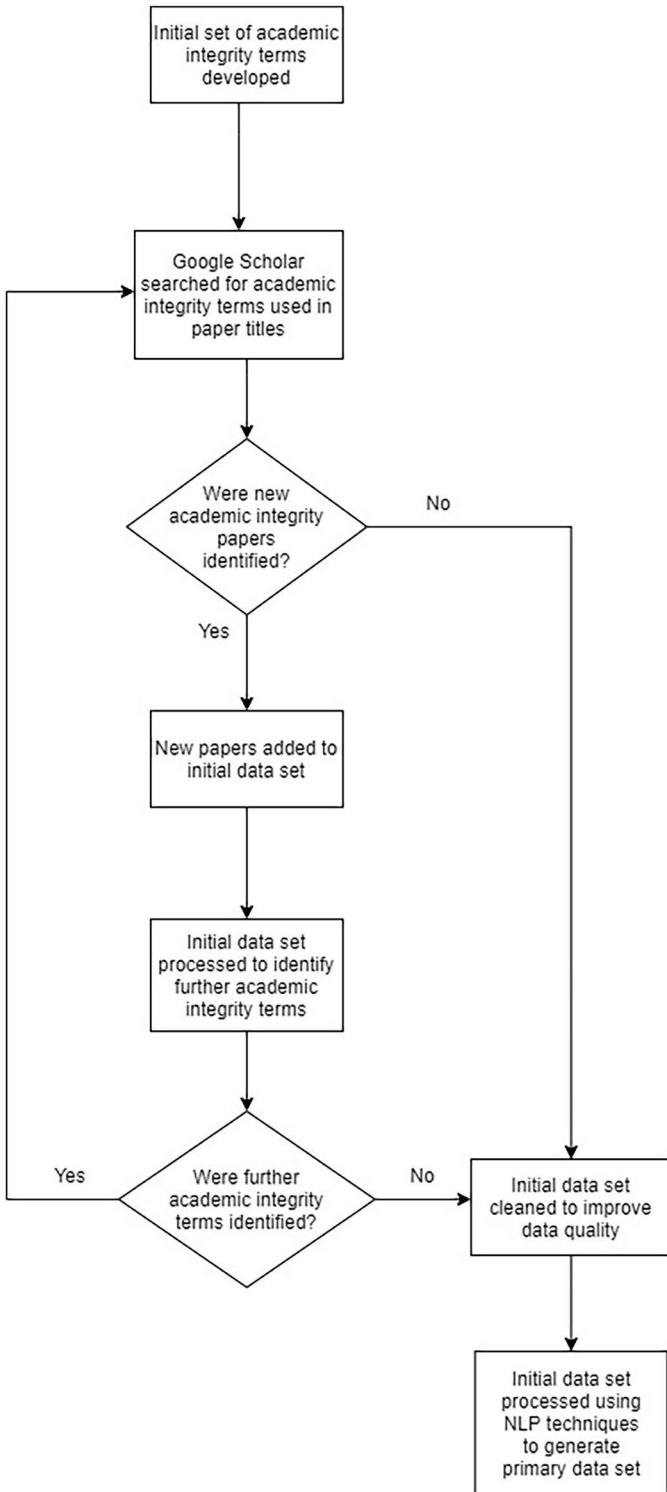


Table 1 Google Scholar search terms used to generate the initial data set

academic dishonesty	educational integrity	student cheating
academic integrity	essay mill	student plagiarism
academic misconduct	exam cheating	teaching cheating
academic outsourcing	exam dishonesty	test cheating
college cheating	exam integrity	test honesty
contract cheating	examination cheating	university cheating
custom essay	parental cheating	
educational corruption	school cheating	

As Fig. 1 indicates, an initial set of search terms for Google Scholar were identified. These included such terms as *academic integrity*, *academic dishonesty*, *student plagiarism* and *contract cheating*. In each case, a search for these terms in the title of documents was conducted. This search type meant that a term like *student plagiarism* would match the word *student* and the word *plagiarism* used anywhere within a title and with the words in either order. The results were manually inspected to identify other possible search terms. Subsequently a list of bigrams (two consecutive words) in the titles was generated to identify more possible search terms. Frequently occurring bigrams related to the wider academic integrity area were also used. This process identified, for example, the term *academic honesty* as an alternative to academic dishonesty. The process also suggested the term *research integrity*, but this was deliberately excluded to avoid adding large quantities of papers to the initial data set that were unrelated to teaching, education or students. Nevertheless, some research integrity papers do appear in the

Table 2 Inclusion and exclusion criteria for the primary data set

Main reasons for inclusion	Main reasons for exclusion
Papers on an academic integrity topic relevant to teaching, learning, students or education	Papers where the title, abstract or summary were not in English
Books with a relevant research focus	Newspaper and magazine articles
Individual chapters from edited books	Book reviews (primarily of academic integrity texts)
Papers published up to the end of 2019	Letters, commentary and editorials
	Internal university documents, such as minutes of meetings
	Presentation slides
	Research posters
	Papers not related to academic integrity (e.g. literary plagiarism, pure research integrity, cheating in college sports, plagiarism law, cheating in relationships, signal integrity)
	Books and guides aimed to teach students how to reference and avoid plagiarism
	Reports authored by organisations (such as QA agencies) rather than individuals
	Blog posts
	Ill-formed sources which could not easily be resolved, particularly where these were only in the form of citations and lacked publisher information or publication years
	Student work, such as essays
	Promotional material by contract cheating services and essay mills

final evaluation where these were identified through other terms and did prove to be relevant to academic integrity. Table 1 shows the final set of search terms that were used.

The initial data set required extensive cleansing. The process which Google Scholar uses to crawl research papers and generate its own records is error prone and so the initial data set contained many duplicate entries, for example where one version had a slightly incorrect title, listed authors in different orders or had author name variants. There were many cases where unsuccessful parsing of research documents had generated incorrect information in the Google Scholar database. In addition, information standing out as potentially suspect was cross-referenced against other sources. One such example was an article about student cheating and the Internet, allegedly published in 1970, whereas a check on the journal's own web pages revealed the correct date.

A further pruning process was necessary to limit the initial data set to only include papers that were research related and on subjects in the wider academic field. Table 2 summarises some of the main criteria applied to identify if papers should be included in or excluded from the primary data set. No direct attempt was made to judge the quality of the papers or exclude those published in predatory journals, although the development of secondary data sets of papers that the academic integrity community considers most important does indirectly address this possible limitation.

The primary data set contained information about 8,507 research sources published between 1904 and 2019. A cumulative frequency graph showing when the papers were published in shown in Fig. 2.

Figure 2 indicates that the rate of increase of publications in the academic integrity field has been exponential. There were approximately the same number of papers published between 1904 and 2011 as there were between 2012 and 2019. Although this may seem like a steep rate of increase, worldwide science and engineering publications were found to have grown at a rate of 4% per year between 2008 and 2018 (White, 2019). The corresponding figure for academic integrity publications is just below 3% per year.

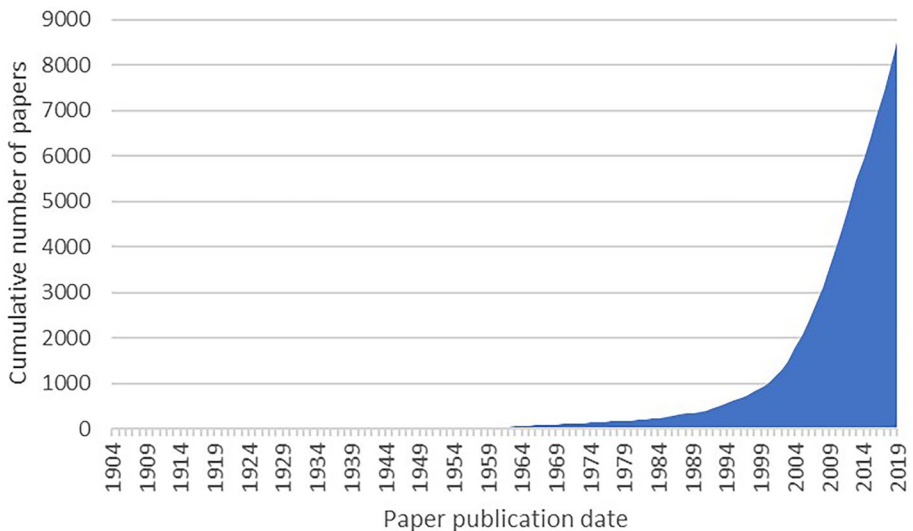


Fig. 2 Cumulative frequency chart of academic integrity paper publications

Formation of the Secondary Data Sets

A further four smaller secondary data sets, all subsets of the primary data set, were developed to allow for a more detailed investigation. These data sets are summarised in Table 3.

Data sets B, C and D consider the most cited papers of all time. These are intended to represent the papers that have overall influence on the academic integrity field. Although these data sets would seem to favour older publications, the 1000 most cited papers data set (D) does contain papers dated as recently as 2019. In general, the primary data set (A) is where most recently published research can be found. This is illustrated in Fig. 3, which shows the relative cumulative frequencies of publications in the five data sets, truncated to start from 1979.

Computation of Individual Primary and Secondary Data Set Records

Individual records were generated for each paper included in the data sets through a combination of continued data cleansing and the application of NLP techniques.

The paper title information obtained from Google Scholar was tokenised to represent paper titles as a series of words of interest. This included the removal of common English language stop words (“and”, “the”, “of” etc.) based off a standard list for the library used.¹ In addition, the word “among” was removed. Two further minor changes were made to improve wording that was not picked up by the standard tokenisation process. This saw the token “student” replaced by “students” and the token “toward” replaced by “towards”. This decision was made to allow these common terms to be clustered together and improve the readability of the final results, rather than see two similar terms occupy two lots of results and confuse matters.

Table 4 shows the information collected and computed for each record in the data sets, along with an indicative example. As well as information collected directly from Google Scholar and subsequently cleansed, this includes information computed using standard NLP techniques of unigrams, bigrams and trigrams. A sentiment analysis score for each title is also calculated to determine if this represents positive, neutral or negative integrity. In the case of the example of Eshet et al. (2014) shown in Table 4, the overall sentiment is considered to be negative, with the machine learning process likely to have made this judgement through the use of the terms “*traits*” and “*academic dishonesty*” in the paper title.

¹ The default stop words provided within the library used are: i, me, my, myself, we, our, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, his, himself, she, he', her, hers, herself, it, t', its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, hat'l, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, on', should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, didn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ma, mightn, mightn't, mustn, mustn't, needn, needn't, shan, shan't, shouldn, shouldn't, wasn, wasn't, weren, weren't, won, won't, wouldn, wouldn't.

Table 3 Summary of primary and secondary data sets

Date Set Identifier	Data Set Name	Inclusion Criteria	Num-ber of Records	Date Range	Average (mean) number of cita-tions	Standard deviation of number of citations
A	Primary data set	All qualifying papers	8507	1904–2019	14.27	48.52
B	10 most cited	10 papers with the most citations. Where papers are tied for inclusion, the earliest published paper is included, followed by the paper whose authors appear first alphabetically	10	1986–2006	869.20	247.80
C	100 most cited	100 papers with the most citations. Where papers are tied for inclusion, the earliest published paper is included, followed by the paper whose authors appear first alphabetically	100	1964–2012	341.07	211.51
D	1000 most cited	1000 papers with the most citations. Where papers are tied for inclusion, the earliest published paper is included, followed by the paper whose authors appear first alphabetically	1000	1957–2019	93.33	112.76
E	Most prolific authors	Papers by the 10 authors with the highest publication counts	264	1982–2019	56.32	118.69

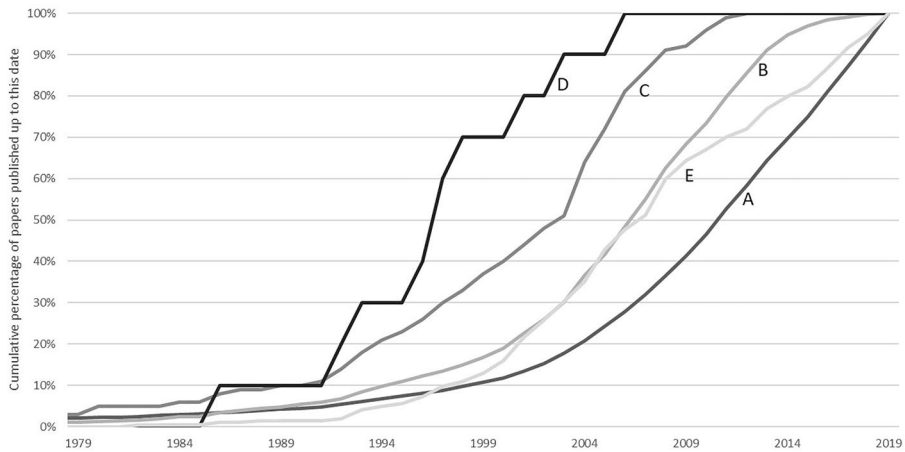


Fig. 3 Relative cumulative frequencies of paper publications in data sets (1979 to 2019)

Research Methodology Limitations

Some natural limitations of the approach used within this investigation are worthy of mention. The data sets represent a snapshot of content on a live source of data, one that continually receives updates and corrections. The volume of citations observed in May 2020 will be different to that which would have been seen at the end of 2019, the cut-off point for including papers in the data sets. Even then, official publication dates can differ from the date papers were first available to be read and cited. This stems from the advent of papers being published online first before they are assigned to a journal issue.

The analysis presented focuses on paper titles, rather than paper abstracts or their contents. This assumes that titles accurately reflect the contents of the papers. There will be exceptions to this, for instance when a title is written for shock value or to encourage readership, in much the same way that newspaper headlines can be written to draw attention. The tendency for authors to think about search engine optimisation when organising papers is also a relatively recent change that may have influenced the choice of titles within the field.

Only a single source, Google Scholar, is used for data collection. The quality of the data sets is limited to the quality of the underlying source. The resulting data sets did require manual clean up and it is likely that a small number of errors remain. The time afforded for data cleansing and consistency checking was used strategically, focusing most on the secondary data sets since these are likely to have the greatest influence on future practice. Small errors in the primary data set (A) of 8,507 items should have no discernible effect on the accuracy of the overall results. These results are still of importance to the wider academic integrity research field.

Table 4 Data attributes used in this study for each record in the data sets

Attribute	Source	Attribute Description	Example Based on Eshet, Grinautski and Peled (2014)
<i>Paper title</i>	Google Scholar	The unedited title of the paper	“No More Excuses—Personality Traits and Academic Dishonesty in Online Courses”
<i>Publication year</i>	Google Scholar	The stated year of publication	2014
<i>Number of authors</i>	Google Scholar	The stated number of authors	3
<i>First three authors</i>	Google Scholar	The set of names of paper authors, truncated at a maximum of three authors	{“Y Eshet”, “K Grinautski”, “Y Peled”}
<i>Citations</i>	Google Scholar	Number of citations for the paper	8
<i>Unigrams</i>	Computed	Set of individual words in paper title following pre-processing	{“excuses”, “personality”, “traits”, “academic”, “dishonesty”, “online”, “courses”}
<i>Bigrams</i>	Computed	Set of pairs of consecutive words in paper title following pre-processing	{“excuses personality”, “personality traits”, “traits academic”, “academic dishonesty”, “dishonesty online”, “online courses”}
<i>Trigrams</i>	Computed	Sets of three consecutive words in paper title following pre-processing	{“excuses personality traits”, “personality traits academic”, “traits academic dishonesty”, “academic dishonesty online”, “dishonesty online courses”}
<i>Sentiment analysis score</i>	Computed	Compound sentiment analysis score for paper title, normalised between -1.00 (most negative) and +1.00 (most positive)	-0.296
<i>Sentiment</i>		Defined as: Positive, for sentiment score between 0.05 and 1.00 Negative, for sentiment score between -1.00 and -0.05 Neutral otherwise	Negative

Table 5 Most frequent unigrams, bigrams and trigrams in primary data set

Rank	Unigrams	Bigrams	Trigrams
1	academic	academic integrity	college student cheating
2	plagiarism	academic dishonesty	source code plagiarism
3	student	plagiarism detection	high school student
4	integrity	academic misconduct	perceptions academic dishonesty
5	cheating	college student	code plagiarism detection
6	education	higher education	student academic integrity
7	dishonesty	student cheating	plagiarism higher education
8	detection	integrity education	integrity higher education
9	college	student plagiarism	promoting academic integrity
10	university	university student	plagiarism detection using

Results and Discussion

Most Frequently Occurring Unigrams, Bigrams and Trigrams in Paper Titles

Table 5 summarises the 10 unigrams, bigrams and trigrams seen most frequently in the primary data set (A). Each of these measures provides insight into academic integrity research at different levels of granularity.

The unigram data indicates that 7,161 unique unigrams were observed in the primary data set, with a total of 60,402 occurrences. This shows a mean of 8.43 occurrences per unigram and a standard deviation of 477.63. The top 10 ranked unigrams covered 17,203 of those occurrences between then (28.48%). The unigram list does not appear to be particularly insightful.

The bigram data provides more useful level of granularity, with 30,169 unique bigrams and a total of 51,898 occurrences. That is a mean of 1.72 occurrences per bigram, with standard deviation of 11.22. The top 10 ranked bigrams cover 4,620 occurrences (8.90%). The first and second ranked bigrams “*academic integrity*” (seen in 1,234 occurrences, that is 2.37%) and “*academic dishonesty*” (seen in 1,068 occurrences or 2.06%) indicate the close relationship between the use of these two terms.

The trigram data indicates variety across paper titles, with 36,417 unique trigrams observed across 43,420 occurrences, giving a mean of 1.19 occurrences per bigram and a standard deviation of 1.42. Between them, the top 10 ranked trigrams cover only 524 occurrences (1.21%). The terms make intuitive sense and the quadgram “*source code plagiarism detection*” stands out as formable from the second and fifth ranked trigrams, indicating the interest in academic integrity techniques often considered most specific to Computer Science.

Exploration of Bigram Data

The bigram level provides the opportunity to further explore the primary and secondary data sets. Figure 4 shows the data obtained from the primary data set in more detail.

Categorising the bigrams provides an indication of the topics of most interest to academic integrity researchers. The positive integrity terms “*academic integrity*”, “*academic*

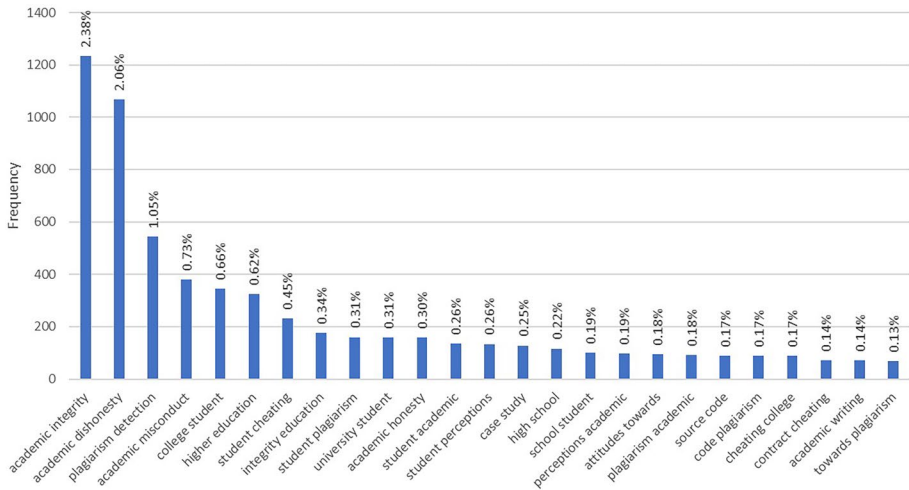


Fig. 4 Top 25 bigrams observed in primary data set

honesty” and “*integrity education*” can be combined to cover 1,568 out of 51,898 occurrences (3.02%). Accordingly, the negative integrity terms “*academic dishonesty*”, “*academic misconduct*” and “*student cheating*” can be combined to give 1,792 out of 51,898 (3.45%) occurrences, suggesting a slight bias towards negativity in paper titles. Other terms suggest wider areas of interest, including the academic level of students (such as university, college and high school), academic integrity challenges (such as plagiarism and contract cheating), methods of addressing challenges (such as through case studies and plagiarism detection), issues of interest to particular research sub groups (academic writing, source code plagiarism and plagiarism detection), as well as the type of data hoped to be gathered in many research projects (perceptions and attitudes).

The data sets were further interrogated to identify how many of the 25 most frequent bigrams occurred in the paper titles, as well as the number of the three positive bigrams (“*academic integrity*”, “*academic honesty*” and “*integrity education*”) and three negative bigrams (“*academic dishonesty*”, “*academic misconduct*” and “*student cheating*”) seen in those titles. Particular attention is paid to the most prolific of those terms, “*academic integrity*” and “*academic dishonesty*” and these are also analysed separately. The results are seen in Tables 6 and 7.

Table 6 suggests that when more of the 26 most frequently occurring bigrams are included in paper titles, those papers are more likely to be cited. They also suggest that the interest in negative integrity is greater than that in positive integrity. For example, in the 1000 most cited papers data set (D), the average paper title contains 0.113 out of the 3 positive bigrams, but contains 0.240 out of the 3 negative bigrams, an increase of 112.39%. A similar finding can be observed when comparing the use of the bigram “*academic integrity*” with “*academic dishonesty*”. The negative bigram is most frequent in all three of the most cited data sets (B, C and D).

There is one clear exception to this finding and that comes from the most prolific authors data set (E). This group uses 0.330 out of 3 positive bigrams per paper title, accompanied by only 0.205 out of 3 negative bigrams, showing that the titles they use contain 60.98% more positive than negative bigrams. Similarly, the prolific authors use the bigram

Table 6 Average (Mean) Numbers of 25 Most Frequent Bigrams Found in Paper Titles

Data Set Identifier	A	B	C	D	E
<i>Data Set Name</i>	Primary data set	10 most cited	100 most cited	1000 most cited	Most prolific authors
<i>Bigrams Per Title</i>	0.712	1.100	0.860	0.724	0.879
<i>Positive Bigrams Per Title</i>	0.183	0.000	0.090	0.113	0.330
<i>Negative Bigrams Per Title</i>	0.196	0.500	0.330	0.240	0.205
<i>Occurrences of "Academic Integrity"</i>	0.144	0.000	0.080	0.099	0.322
<i>Occurrences of "Academic Dishonesty"</i>	0.124	0.500	0.240	0.180	0.136

Table 7 Standard Deviation of Numbers of 25 Most Frequent Bigrams Found in Paper Titles

Data Set Identifier	A	B	C	D	E
<i>Data Set Name</i>	Primary data set	10 most cited	100 most cited	1000 most cited	Most prolific authors
<i>Bigrams Per Title</i>	0.736	0.831	0.707	0.739	0.597
<i>Positive Bigrams Per Title</i>	0.393	0.000	0.286	0.317	0.470
<i>Negative Bigrams Per Title</i>	0.398	0.500	0.511	0.432	0.403
<i>Occurrences of "Academic Integrity"</i>	0.351	0.000	0.271	0.299	0.467
<i>Occurrences of "Academic Dishonesty"</i>	0.330	0.500	0.427	0.384	0.343

Table 8 Percentage of Paper Titles Containing 25 Most Frequent Bigrams

Data Set Identifier		A	B	C	D	E
<i>Data Set Name</i>		Primary data set	10 most cited	100 most cited	1000 most cited	Most prolific authors
<i>Percentage of Paper Titles Containing</i>	0	43.52%	20.00%	30.00%	42.30%	23.48%
<i>This Number of Bigrams</i>	1	43.80%	60.00%	57.00%	45.30%	66.29%
	2	10.78%	10.00%	10.00%	10.40%	9.09%
	3	1.80%	10.00%	3.00%	1.70%	1.14%
	4	0.11%	0.00%	0.00%	0.30%	0.00%

Table 9 Percentage of Paper Titles Containing Specific Bigrams

Data Set Identifier	A	B	C	D	E
<i>Data Set Name</i>	Primary data set	10 most cited	100 most cited	1000 most cited	Most prolific authors
<i>Title Containing Academic Integrity</i>	14.41%	0.00%	8.00%	9.90%	32.20%
<i>Title Containing Academic Dishonesty</i>	12.45%	50.00%	24.00%	18.00%	13.64%

“academic integrity” in almost one third of their paper titles, 136.76% more of the time than they use “academic misconduct”.

Further analysis show variety in number of the 25 most frequent bigrams included in paper titles. This is summarised in Table 8.

From the primary data set (A), 43.52% of paper titles do not contain any of the 25 most frequent bigrams. For the most prolific authors data set (E), that percentage is only 23.48%, suggested that the researchers writing regularly in this field are familiar with the wider literature, the research base and the terminology to use. In all five data sets, the modal number of the most frequent bigrams used in a paper title is 1. There are 9 cases out of 8,507 records (0.11%) where four bigrams are used, sometimes as part of overlapping bigrams.

Table 9 compares the use of the bigrams “academic integrity” and “academic dishonesty” in paper titles. This indicates a perhaps alarming result, that papers in the field are more likely to be highly cited if they take a negative integrity stance. Once again, the most prolific authors buck this trend. From data set B, none of the 10 most academic integrity papers of all time contain “academic integrity” in their title, or indeed any of the positive keywords that have been identified from the 25 most frequently used bigrams.

Sentiment Analysis

The paper titles in the data sets were analysed using sentiment analysis techniques to determine if they represented positive, neutral or negative sentiment. A summary of the percentage of paper titles falling within each of these sentiments is shown in Table 10.

For data sets A to D, the modal sentiment is neutral, although the most interesting comparisons lie between positive and negative sentiment. The results from the primary data set (A) show that titles are slightly more likely to be viewed as having positive sentiment rather

Table 10 Sentiment Analysis of Data Sets

Data Set Identifier	A	B	C	D	E
<i>Data Set Name</i>	Primary data set	10 most cited	100 most cited	1000 most cited	Most prolific authors
<i>Positive Sentiment</i>	33.96%	10.00%	22.00%	26.20%	45.45%
<i>Neutral Sentiment</i>	41.47%	40.00%	42.00%	46.30%	31.82%
<i>Negative Sentiment</i>	24.57%	50.00%	36.00%	27.50%	22.73%
<i>p</i>	<0.001	0.273	0.042	<0.001	<0.001

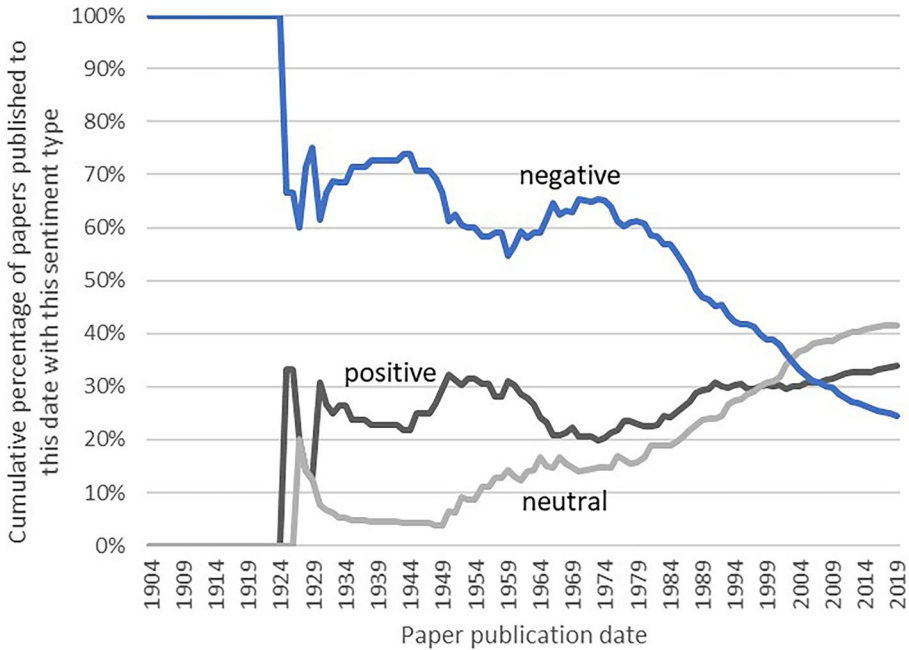


Fig. 5 Cumulative Sentiment Analysis of Paper Titles

than negative sentiment, but this is not consistently the case across all the data sets. In particular, the most cited data sets (B, C and D) show more negative than positive sentiment.

Different results are seen from the most prolific authors data set (E), where 120 out of 264 paper titles (45.45%) are computed to have positive sentiment, making this the modal sentiment group. Since the negative sentiment group contains 60 paper titles, this represents a 100% increase.

Considering a null hypothesis that, if randomly and independently determined, one third of paper titles should each show positive, neutral and negative sentiment, Pearson's chi-squared test shows statistical significance for data sets A, C and D at the 0.001% level and for data set C at the 5% level. Data set B does not show statistical significance, but strictly speaking the sample size is too small for Pearson's test to be valid.

A further element of investigation aims to address how the sentiment of paper titles has developed over time. The results are shown in Fig. 5.

Figure 5 provides a cumulative plot of the percentage of paper titles that were computed to have positive, neutral and negative sentiment. That is, Fig. 5 shows the sentiment results from all the papers published up to a given point in time. This trend shows promise if a move towards positive integrity is considered desirable. Although historically the sentiment of paper titles has been strongly negative, neutral sentiment overtook negative sentiment for the first time in 2003. Positive sentiment then overtook negative sentiment in 2008. The current trend shows a continued decline in the use of negative sentiment.

The answer to one final question may interest researchers in this field. Does having positive sentiment in a paper title affect the number of citations that paper is likely to obtain? Across the primary data set (A) as a whole, papers received an average of 14.27 citations. The paper titles with positive sentiment received 10.60 citations. The papers with

neutral sentiment received 15.39 citations. The papers with negative sentiment received 17.46 citations. It would appear that developing paper titles with negative sentiment affords a good way to get work cited within the academic integrity research field.

Conclusion and Recommendations

This paper represents the first study of its kind in the academic integrity research field, using the largest known data set of academic integrity research publications as its base. The analysis shows that academic integrity research is a field with rapid growth, but citations have been built upon publications with a negative concept of integrity. Both bigram analysis and sentiment analysis show a similar view of negative integrity, but with pockets of positive integrity shining through.

Many opportunities exist to take this research forward. Similar techniques can be applied to other fields, or to specialised subjects within the academic integrity area. The bigram technique has shown the existence of many long-tail keywords that are suitable for literature reviews and more detailed analysis. The sentiment analysis approach used is not specific to academic integrity and could be further optimised through the development of training data sets. In addition, it would be interesting to apply these techniques to paper abstracts and full papers to see if the same results hold. Academic integrity researchers and practitioners may find it useful to develop more NLP and linguistic analysis skills. Many of the techniques applied to research are already akin to those which can be applied to forensic investigation of student work to detect plagiarism and contract cheating (Ison, 2020; Johnson & Davies, 2020).

Although not an intended focus of the investigation, the data serendipitously revealed that there appears to be a question to be posed regarding the value of much academic integrity research. In the primary data set developed for this paper, 2854 out of 8507 papers (33.55%) have never received a single citation. In addition, threats to research integrity were observed when examining the data set. A 2019 paper was found published in three different journals by the same suspect publisher with only slight changes to the paper title and abstract. Paper citation cartels seem to be developing, with single papers having a large group of authors, each of whom then go on to cite as many papers as possible by members of the group. The effect seems to be an artificial bump up the citation metrics for all members. In a field like academic integrity, researchers also need to hold their own practices up to the highest standards.

There are issues that need to be addressed regarding what content should be placed in research repositories and how Google Scholar results are produced. Students can be referred to Google Scholar as a valid starting point for their own research, but not all search results are suitable for this purpose. One university repository contains an archive of blog posts by researchers, but these are now listed by Google Scholar as if they are academic papers. There are also many examples of contract cheating providers finding ways to have their content added to Google Scholar, complete with visible adverts. The promotional methods of the contract cheating industry have already been observed as being highly suspect (Lancaster, 2019) and this is providing yet another method through which students can be brought into their marketing funnel.

One of the biggest disputes in the academic integrity community surfaced continually throughout this paper. Is the best terminology to use in research “academic integrity” or “academic dishonesty”? Should researchers take the opportunity to introduce a more

positive viewpoint of the field? There is much historical interest to the use of the term “academic dishonesty” but this term may no longer be necessary. Despite this, research papers that take a negative view of integrity, using terms such as cheating, dishonesty and misconduct, do drive an emotional response in a manner that integrity does not seem to do. Such papers then benefit from more citations and drive future research. It is something of a vicious circle.

This paper has demonstrated that it is possible to present the academic integrity research field using positive terminology. Several of the most prolific authors in the field are doing just that. More publications appear to be taking a positive integrity view than ever before. Emerging academic integrity researchers can and should be encouraged to model their approach on such papers and researchers. But further effort needs to be made by the academic integrity community to promote such papers and to show that research into positive integrity is possible, worthwhile and of value.

Perhaps then a move to purely talk about positive integrity is a step too far. As the sentiment analysis presented in this paper has demonstrated, the most recent trend in paper title construction has been a move towards titles devoid of positive or negative intention. Researchers in related fields talk about ethical neutrality. At the start of this paper, a quote from Barnes (1904) talked about social responsibility. Too often, academic integrity researchers are the same people who are the practitioners working on academic integrity in the classroom, teaching students and often awarding penalties for academic integrity breaches. Due to the nature of the research field, true independence of research from practice and teaching may be impossible. Aiming instead for neutrality as to how research in the academic integrity field is presented may then provide the best future solution for all concerned.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Amigud, A., & Lancaster, T. (2019). 246 reasons to cheat: An analysis of students’ reasons for seeking to outsource academic work. *Computers & Education*, *134*, 98–107.
- Atkinson, D., Nau, S. Z., & Symons, C. (2019). Ten years in the academic integrity trenches: Experiences and issues. *Journal of Information Systems Education*, *27*(3), 5.
- Barnes, E. (1904). Student honor: A study in cheating. *The International Journal of Ethics*, *14*(4), 481–488.
- Clarke, R., & Lancaster, T. (2006). Eliminating the successor to plagiarism? Identifying the usage of contract cheating sites. In: *Proceedings of 2nd International Plagiarism Conference*. JISC Plagiarism Advisory Service, Newcastle, United Kingdom.
- Dawson, P. (2020). *Cybersecurity: the next academic integrity frontier*. Edward Elgar Publishing.
- Eaton, S. E., & Edino, R. I. (2018). Strengthening the research agenda of educational integrity in Canada: A review of the research literature and call to action. *International Journal for Educational Integrity*, *14*(1), 5.
- Eaton, S. E., Stoesz, B. M., Thacker, E. J., & Miron, J. B. (2020). Methodological decisions in undertaking academic integrity policy analysis: Considerations for future research. *Canadian Perspectives on Academic Integrity*, *3*(1), 83–91.

- East, J., & Donnelly, L. (2012). Taking responsibility for academic integrity: A collaborative teaching and learning design. *Journal of University Teaching & Learning Practice*, 9(3), 2.
- Eshet, Y., Grinautski, K., Peled, Y., & Barczyk, C. (2014). No more excuses - personality traits and academic dishonesty in online courses. *Journal of Statistical Science and Application*, 2(3), 111–118. <https://doi.org/10.17265/2328-224X/2014.03.004>
- Fishman, T. (2014). *The fundamental values of academic integrity*. Clemson University.
- Fishman, T. (2016). Academic integrity as an educational concept, concern, and movement in US institutions of higher learning. *Handbook of Academic Integrity*, 7–21.
- Harrison, D., Patch, A., McNally, D., & Harris, L. (2020). Student and faculty perceptions of study helper websites: a new practice in collaborative cheating. *Journal of Academic Ethics*, 1–18.
- Ison, D. C. (2020). Detection of online contract cheating through stylometry: A pilot study. *Online Learning*, 24(2).
- Johnson, C., & Davies, R. (2020). Using digital forensic techniques to identify contract cheating: A case study. *Journal of Academic Ethics*, 1–9.
- Lancaster, T. (2019). Social media enabled contract cheating. *Canadian Perspectives on Academic Integrity*, 2(2), 7–24.
- Macfarlane, B., Zhang, J., & Pun, A. (2014). Academic integrity: a review of the literature. *Studies in Higher Education*, 39(2), 339–358.
- McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: honor codes and other contextual influences. *The Journal of Higher Education*, 64(5), 522–538.
- McCabe, D. L., & Pavela, G. (2004). Ten (updated) principles of academic integrity: how faculty can foster student honesty. *Change: the Magazine of Higher Learning*, 36(3), 10–15.
- Pitt, P., Dullaghan, K., & Sutherland-Smith, W. (2020). 'Mess, stress and trauma': students' experiences of formal contract cheating processes. *Assessment & Evaluation in Higher Education*, 1–14.
- Prentice, F. M., & Kinden, C. E. (2018). Paraphrasing tools, language translation tools and plagiarism: an exploratory study. *International Journal for Educational Integrity*, 14(1), 11.
- Ransome, J., & Newton, P. M. (2018). Are we educating educators about academic integrity? A study of UK higher education textbooks. *Assessment & Evaluation in Higher Education*, 43(1), 126–137.
- Sefcik, L., Striipe, M., & Yorke, J. (2020). Mapping the landscape of academic integrity education programs: what approaches are effective? *Assessment & Evaluation in Higher Education*, 45(1), 30–43.
- White, K. (2019). Publications Output: U.S. Trends and International Comparisons. <https://nces.nsf.gov/pubs/nsb20206>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.