TSP Trends in Scholarly Publishing

OPEN ACCESS

Time-Bound Investigational Text Analysis of Predatory E-mails Using Pareto Principle

Mostafa Essam Eissa
Pharmaceutical Research Facility, Cairo, Egypt

ABSTRACT

Predatory act in the scientific field is a progressively growing problem that impacts the integrity of the scientific community. While there are several guidelines available to avoid them, young researchers and inexperienced scientists may not come across them unless it is too late. This might be a challenge, especially for those working in the non-academic sector where a lack of clear guide or reference is not uncommon. The objective of this work is to provide a simple and straightforward guide for recognizing predatory patterns through simple and commercial tools in the analysis of the messages which are used as traps to bring the victims into the net of the intruders ruining the efforts of the researchers and destroying their reputation, in addition to the adverse effects on the progression of their career. The E-mail of an industrial researcher was screened for spam messages for those with scientific nature and then they were isolated in an Excel sheet for processing using the MeaningCloud Add-in platform for text analysis followed by further examination using Pareto charting for prioritization of the contributors using Minitab and Excel. Predatory e-mail senders showed an encouraging pattern to the victims with a systematized insistent attack on the inbox. They hide under the general scope of science, even claiming specificity and professionality. Positive sentiment tendency is predominant in the predatory texts to drag the victim into involvement actions with the fraudulent. Phishers tend to use a bunch of specific keywords to drag their prey to their traps with apparently unprofessional messages that could be sighted through text messages.

KEYWORDS

Agreement, disagreement, cluster, objective, subjective, topic extraction, text classification, Pareto, predatory, polarity

INTRODUCTION

The internet era has marked the burst in the scientific activity sector, notably publications, webinars and conferences. However, this does not come without pernicious consequences[1]. There are lurking deceitful communities of predators aimed to gain their own benefits without conscience who have intruded on the scientific community[2,3]. They attempt to drag the prey of inexperienced researchers and young scientists to their traps[4]. An important means they execute is by sending a mass of e-mails to their victims in the hope that they would hook on their hook.

ISSN: 2789-9446 (Online)
ISSN: 2789-9438 (Print)
https://doi.org/10.21124/tsp.2023.1.8

Received: 20 Sep., 2022
Accepted: 21 Nov., 2022
Published: 3 Jan., 2023

In the present study, we aimed to investigate the pattern and behavior of the messages pertaining to the scientific sector, including publications, editorial or review board invitations, conferences and webinars that are sent to the recipient. Text analysis will be used to elucidate the main characteristics of the predatory messages that are delivered to the accounts of the victims.

MATERIALS AND METHODS

The account of an industrial non-academic researcher in the healthcare field was examined for suspicious spam messages embracing invitations to involvement in scientific activities (manuscript submission, book chapter contribution, conference attendance/speaker, editorial/reviewer board join). Science-centered invitation e-mails were screened and grouped together in an Excel sheet. Conversion of the e-mails to the Excel sheet database was achieved using Email Backup Wizard V 6.0[5]. They were analyzed textually using the customizable MeaningCloud Add-in[6]. The examination involved text classification, topic extraction, deep categorization, text clustering and global and topic sentiment analysis. Datasets generated were further investigated for the predominant contributors based on the Pareto principle. Pareto charts could be generated by either Microsoft Excel V 2207 or Minitab software V 17.1.0 using the 60/40 or 80/20 principle[7-9].

RESULTS

In general, scanning spam messages about scientific activities showed the essence of 81.8% subjectivity and only 18.2% objectivity, equivalent to 153 and 34 received e-mails, respectively. The detailed analysis is described in the following sections.

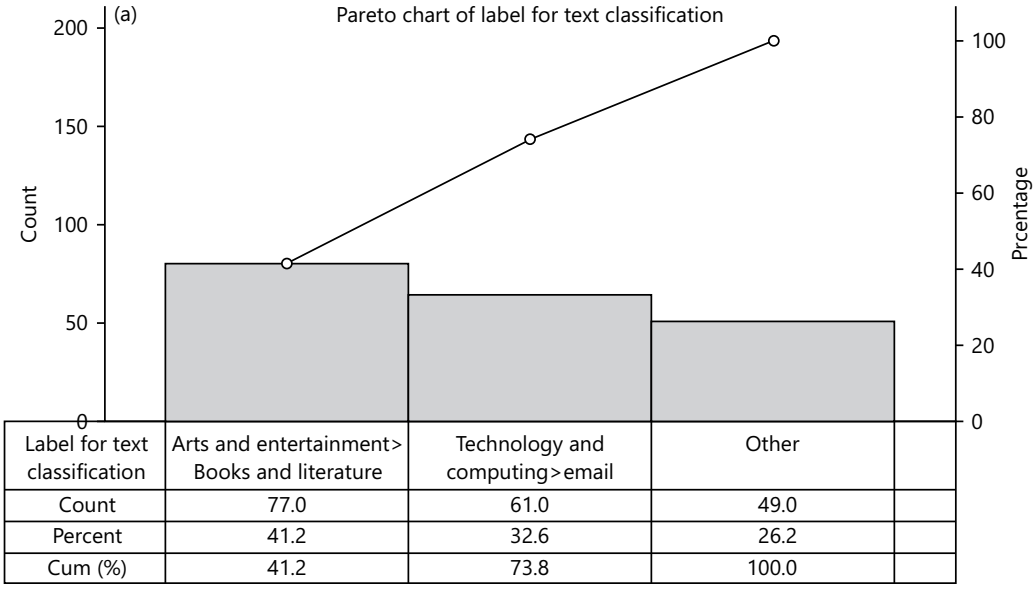
Classification, polarity and agreement of predatory messages: Most messages were non-ironic at their core. However, one e-mail only seemed to smell of irony when overly praising the author's previous work followed by a submission request for a manuscript to drain fees from the researcher.

Messages text body classification: During the course of a one month observation of spam e-mails, two major classes could be identified that label predatory messages in publications and conferences as could be shown in Fig. 1a. Most of the text body of the 187 messages could be labeled with the following codes:

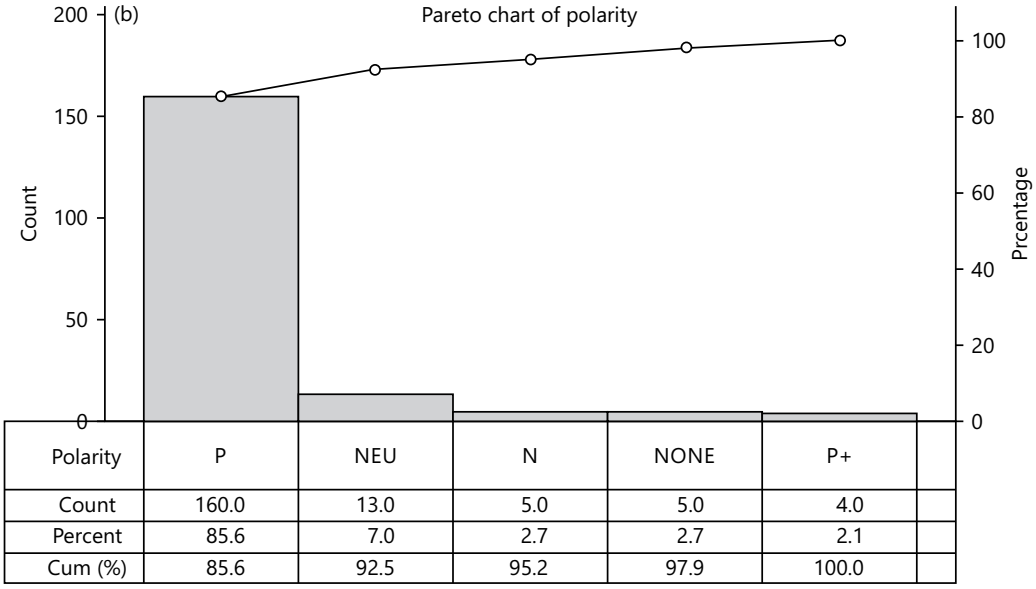
- **Arts and entertainment**
  Books and literature: This label highlights the core of the demand of the book and book chapter contribution, in addition to the most importantly persistent request for article or paper submissions to the claimed organization. Many messages (77) fall within this region, contributing 41.2% of the total e-mails
- **Technology and computing**
  Email: Predatory messages in many instances demonstrate e-mails for sending and/or receiving papers and messages, in addition to the contact persons from the alleged scientific organization. They also direct the recipe for subscription and unsubscription from websites that have not been exposed by the victims in the first place. This might include joining the editorial and review board. This code showed also a significant abundance of 32.6% which is equivalent to 61 messages

Polarity investigation: Predatory invitations and bogus conferences demonstrated a polarity distribution range (Fig. 1b) as follows:

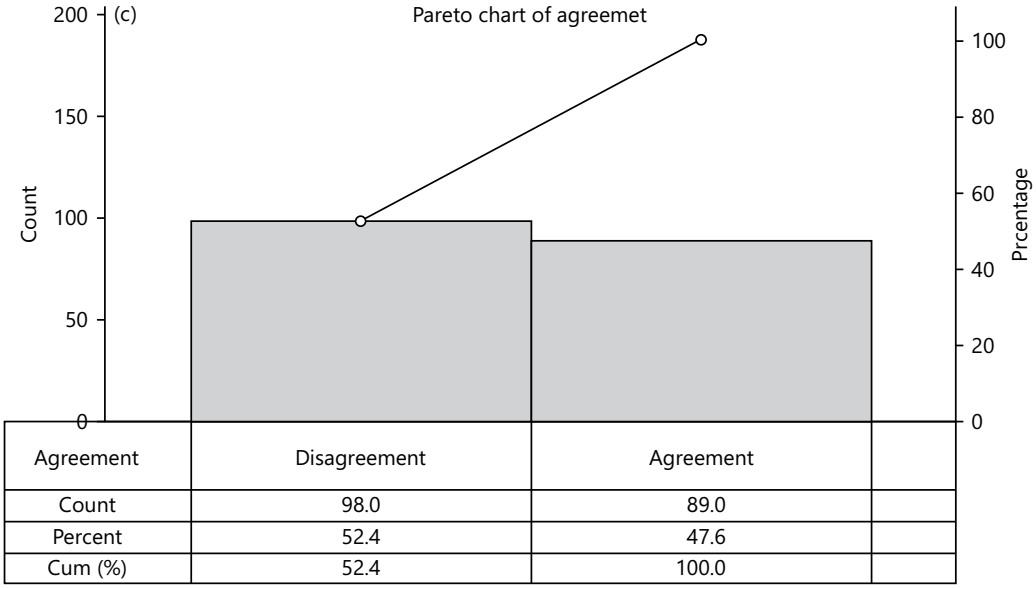
- **Positive polarity (P):** 85.6% which is equivalent to 160 messages
- **Neutral polarity (NEU):** 7.0% which is equivalent to 13 e-mails
- **Negative polarity (N):** 2.7% which is equivalent to 5 messages
- **Non-polar (none):** 2.7% which is equivalent to 5 messages
- **Strongly positive polarity (P+):** 2.1% which is equivalent to only 4 e-mails



Label for text classification	Arts and entertainment> Books and literature	Technology and computing>email	Other	
Count	77.0	61.0	49.0	
Percent	41.2	32.6	26.2	
Cum (%)	41.2	73.8	100.0	



Polarity	P	NEU	N	NONE	P+	
Count	160.0	13.0	5.0	5.0	4.0	
Percent	85.6	7.0	2.7	2.7	2.1	
Cum (%)	85.6	92.5	95.2	97.9	100.0	



Agreement	Disagreement	Agreement	
Count	98.0	89.0	
Percent	52.4	47.6	
Cum (%)	52.4	100.0	

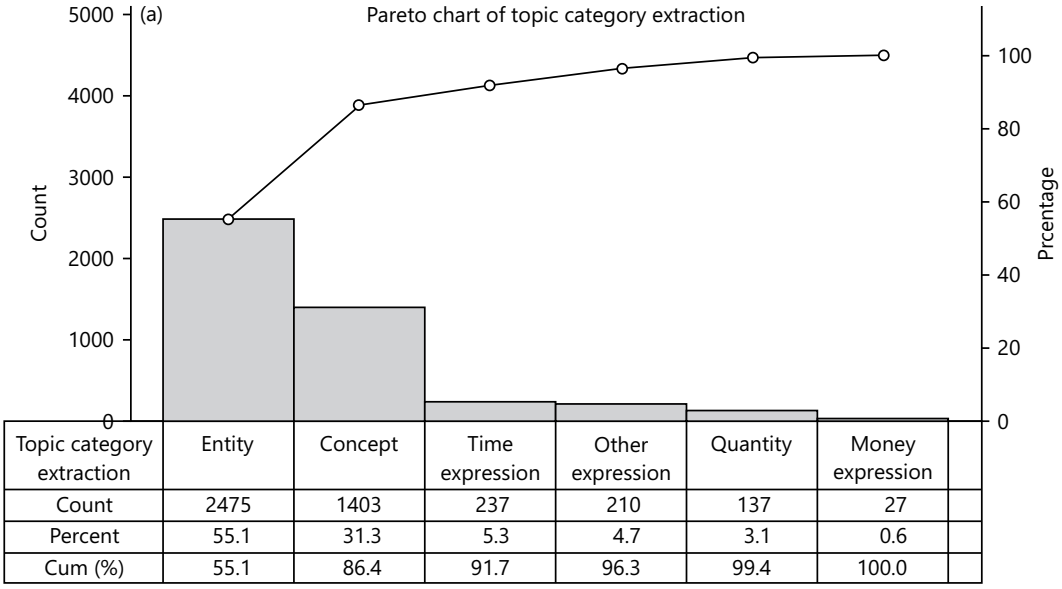
Fig. 1(a-c): Test analysis of the e-mails received from suspect predatory senders (a) Major codes in text classification of e-mail messages received from spam dispatchers, (b) Polarity analysis of e-mail text messages received from spam senders and (c) Agreement and disagreement in polarity

Agreement in polarity: Figure 1c showed a close balance between disagreement and agreement with a slight tilt toward disagreement with values of 98 (52.4%) and 89 (47.6%), respectively.

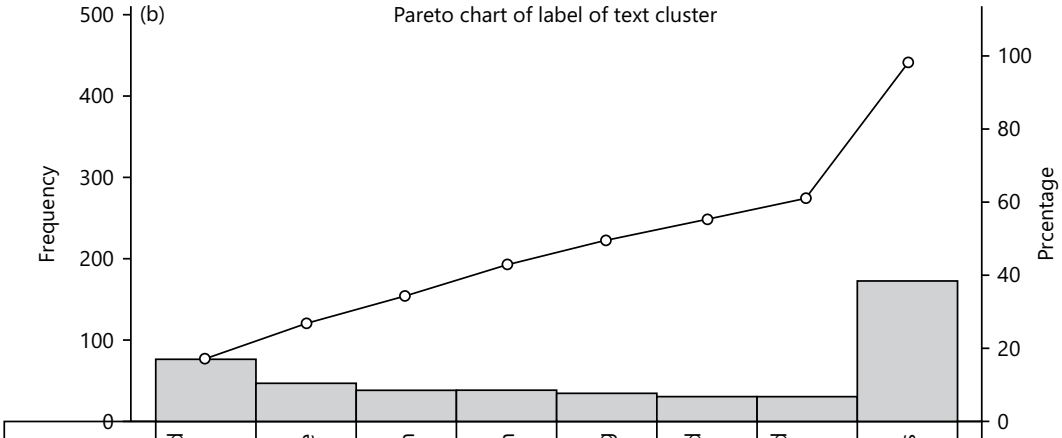
Categorization, labeling and clustering of predatory messages: Suspicious e-mails were analyzed to determine the major and predominant essences of the text core and find the keywords that outline the predatory messages.

Messages text body classification: Extraction of the topic categories yields Fig. 2a showing the Pareto chart for the analyzed texts. Entity and concept together contributed by more than 85% of the cumulative category classes that have been extracted from the texts.

Identification of cluster codes: The deep categorization process showed the major contributing labels of the text bodies by more than 60% in Fig. 2b. Generally, science, business industry and technology topics embedded in the messages showed a share of 32.3% of the total contribution in the scanned predatory



Topic category extraction	Entity	Concept	Time expression	Other expression	Quantity	Money expression
Count	2475	1403	237	210	137	27
Percent	55.1	31.3	5.3	4.7	3.1	0.6
Cum (%)	55.1	86.4	91.7	96.3	99.4	100.0



Label of text cluster	Technology and computing> Computing	Scince	Education	Medical health	Healthy leaving	Books and literature	Business and finance> Industries	Others
Frequency	73.0	45.0	36.0	36.0	31.0	28.0	26.0	170.0
Percent	16.4	10.1	8.1	8.1	7.0	6.3	5.8	38.2
Cum (%)	16.4	26.5	34.6	42.7	49.7	56.0	61.8	100.0

Fig. 2(a-c): Continue

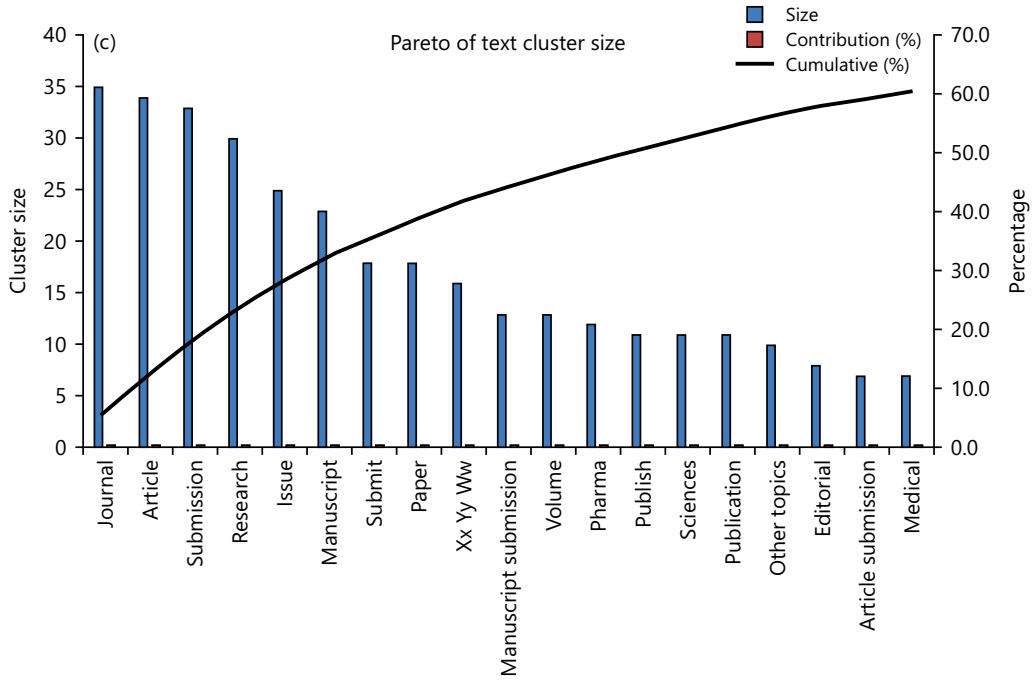


Fig. 2(a-c): Deep investigation of the messages received from the intruder senders (a) Descending order analysis of the topic category extraction, (b) Cluster analysis of the text code using Pareto chart and (c) Cluster size analysis using Pareto diagram showing the major contributing texts by 60%

e-mails. Health field label demonstrated contribution by 15.1%. Text codes that showed the essence of education, book and literature demonstrated 14.4% of the cumulative distribution.

Cluster size analysis: Text analysis for spotting the major clusters could be illustrated in Fig. 2c. There are only 19 keywords that could be recognized in the predatory and suspicious e-mails. They embrace the victim specialty but also cover non-related fields or scopes of work. Also, the full author name is frequently found to be copied in the text bodies. The basic contributing words by about 70% of the messages are listed:

- Journal (6.32%)
- Article (6.14%)
- Submission (5.96%)
- Research (5.42%)
- Issue (4.51%)
- Manuscript (4.15%)
- Submit (3.25%)
- Paper (3.25%)
- Author full name (Xx Yy Ww) (2.89%)
- Manuscript submission (2.35%)
- Volume (2.35%)
- Pharma (2.17%)
- Publish (1.99%)
- Sciences (1.99%)
- Publication (1.99%)
- Other topics (1.81%)
- Editorial (1.44%)
- Article submission (1.26%)

- Medical (1.26%)
- Impact factor (1.26%)
- Trends (1.08%) (0.90%)
- Follow up e-mail (0.90%)
- Clinical (0.90%)
- Open access (0.90%)
- Research journal (0.90%)
- Journal name 2 (1.80%)
- Volume 9 issue (0.90%)
- Global (0.90%)
- Unpublished (0.72%)

DISCUSSION

Predatory e-mail senders showed attempts of manipulating the emotions of desperate victims from the scientific community, especially those from developed countries[10]. They are trapping the researchers by affecting the basic psychological needs of the scientists and researchers to achieve their goals as fast as possible[11]. A mirage of enjoyable benefits coupled with false professionality of lavish handling of the author's best work was commonly shown in the e-mail text using expressions that encourage the authors to deliver their work[12]. The approached phishing behavior in the name of science was labeled in the text classification as could be seen from Fig. 1a using fostering sentiments through messages as could be seen in Fig. 1b by more than 85% of positive polarity. Slightly more than half of the messages showed disagreement between their different segments in the polarity. Accordingly, these e-mails might demonstrate some ambiguously impressing sentiments (Fig. 1c). This seems to put some pressure on the victims to meet some conditions imposed by the spam e-mail senders such as deadlines and fees.

Interestingly, when extraction of the relevant information from each message text was conducted the money expressions were the least to be obvious whereas, the financial attraction was the least obvious from the addressee. In some cases, time events were the most prominent feature such as conference or webinar dates and submission deadlines for articles and papers. In Fig. 2a, the major topics to be extracted were entities that could be demonstrated by names, locations, institutions and organizations. Moreover, concepts also follow as there are many significant and explicit keywords. On the other hand, text tends to show a clustering pattern that could be visualized by the major contributors in Fig. 2b and c. Many cluster codes refer to the prey specialty or close to it. Nevertheless, in other instances, unrelated fields of research were detected in the text body of the predators' e-mails as they showed reckless behavior towards the scientific scope of research. Finally, there are several words that are common in the predatory messages and they can be summarized collectively by "Inviting the author by name (symbolized by "xx yy zz") that has been transcripted as it is-to contribution by a research manuscript in an issue of the journals in pharma, medical sciences by submission of the article in any topics as the paper will be published in the next volume".

This study is limited by the time of the e-mail database retained in the spam folder for 30 days. It could be expected that predators would evolve themselves to gimmick the victims. However, the basic concepts of marking and detection of predatory behavior and attitude would be the same as those described in this study. The description of the general characteristics herein are indicators of the phishing pattern in the scientific field. Till now, there are no signs looming on the horizon for mitigation of this noxious misconduct nor there are no serious legal actions been proposed to stop this thread. Accordingly, predators have become more bold and annoying due to confidence in doing their actions without bearing consequences from the prosecution. Currently, the hope resides in protecting vulnerable populations from the scientific community by increasing their awareness through this investigational study.

CONCLUSION

Predatory activity in the scientific field is a progressively growing challenging problem. The legal actions to stop this type of unethical behavior face a dilemma that cannot be resolved easily. Accordingly, the work provided herein delivers a means to study and understand the predatory pattern through textual analysis using inexpensive and easy methods. Moreover, spotting the keywords could aid the researchers to exclude the sender from the normal inbox list. Finally, the impression of the non-professional and the lack of the scientific essence of the messages could be sensed between the lines in the text which should discourage the scientists from engaging in this kind of unfortunate wasteful activity.

SIGNIFICANCE STATEMENT

In the fast-growing world of internet communication, predatory publications and bogus conferences have become an expanding problem that impacts the integrity of the scientific community. Researchers and scientists are facing constant annoying e-mail messages almost every day to drag them into predatory actions. The study herein demonstrated the implementation of text analysis in the investigation of the messages that are received from suspicious senders. It will be of interest for the researchers and scientists to understand the pattern and behavior of these rapacious messages that could drag them into actions that might ruin their careers and efforts, especially for those who work in non-academic sectors with no guide or mentor for publication activity. Herein, the implementation of text analysis in combination with Pareto charting, using a commercial program platform was used to elucidate major characteristics in message texting to the prey to bring them to participate in their efforts in predatory actions. Identification of the common pattern of these spam e-mails would be useful to young researchers and inexperienced scientists to spot and recognize the suspicious messages to abstain from involvement in a series of unfortunate events that might lead to the defamation of their scientific reputation.

REFERENCES

1. Aaderson, J. and L. Rainie, 2022. Stories From Experts About the Impact of Digital Life. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2018/07/03/stories-from-experts-about-the-impact-of-digital-life/
2. Open Science Support Centre 2022. Predatory Conferences. Available online: https://openscience.cuni.cz/OSCIEN-37.html#:~:text=Predatory%20conferences%20(sometimes%20referred%20to,contribution%20to%20the%20scientific%20community
3. Iowa State University, 2022. Library Guides: Understanding Predatory Publishers: What is a Predatory Publisher? Available online: https://instr.iastate.libguides.com/predatory
4. Shaghaei, N., C. Wien, J.P. Holck, A.L. Thiesen, O. Ellegaard, E. Vlachos and T.M. Drachen, 2018. Being a deliberate prey of a predator-researchers' thoughts after having published in a predatory journal. LIBER Q.: J. Assoc. Eur. Res. Lib., 28: 1-17.
5. Email Backup Wizard, 2022. Email Backup Wizard to Download Emails from Webmail, Cloud, Web Email Services. Available online: https://emailbackupwizard.com
6. MeaningCloud, 2022. MeaningCloud Add-in for Excel | MeaningCloud. Available online: https://www.meaningcloud.com/developer/excel-addin
7. Farmer, E., 2022. Office for Windows-Beta Release Notes-Version 2207. Available online at: https://answers.microsoft.com/en-us/officeinsider/forum/all/office-for-windows-beta-release-notes-version-2207/914b39e5-2dd6-46fd-8add-604bcc8a2f56
8. Eissa, M., 2018. Investigation of outbreaks records and contributing conditions in Unites States. South East Asia J. Med. Sci., 2: 1-2.
9. Asian Hospital and Healthcare Management, 2022. SPC and COVID-19 Quantitative Analysis. Available online: https://www.asianhhm.com/articles/application-of-industrial-statistical-tools-in-quantitative-analysis-of-covid-19-pandemic

10. Shamseer, L., 2022. Predatory" journals: An evidence-based approach to characterizing them and considering where research ought to be published. PhD Thesis, University of Ottawa.
11. Krawczyk, F. and E Kulczycki, 2021. How is open access accused of being predatory? The impact of Beall's lists of predatory journals on academic publishing. J. Acad. Librarianship, Vol. 47. 10.1016/j.acalib.2020.102271.
12. Predatory Journals and Conferences, 2022. Why do predatory publishers send emails and what can you do? Available online: https://predatory-publishing.com/why-do-predatory-publishers-send-emails-and-what-can-you-do/