Final Report
Elsevier Author Contribution System

Jingwen Bai (jb4608), Chenxi Jiang (cj2706), Yunxiao Wang (yw3711),
Xinyu Huang (xh2511), Chuyang Xiao (cx2274)

Abstract

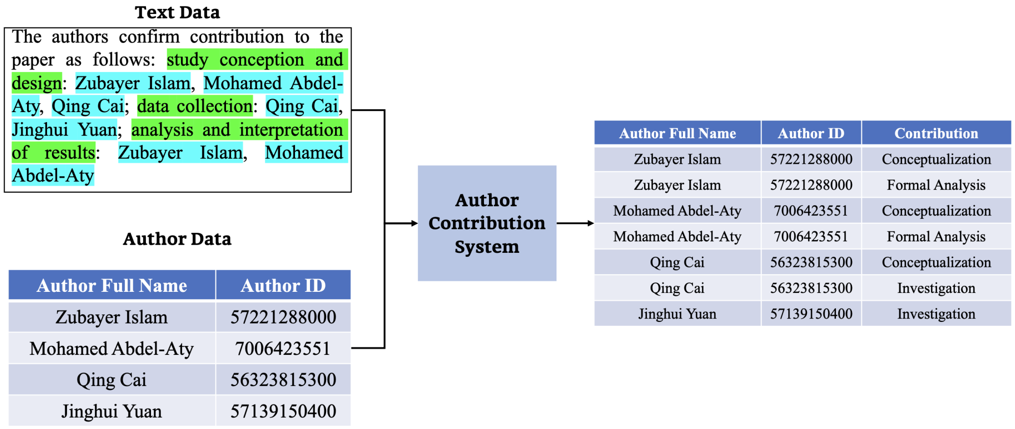
With numerous research papers published and cited at increasing rate nowadays, the CRediT (Contributor Roles Taxonomy) has become particularly significant for all online academic publishers. This function analyzes the author contribution section of all published journals, discriminates the jobs accomplished by each author, and then classifies each author's accomplishments into the 14 contributor roles defined by the CRediT [1]. As Elsevier's current system lacks this function, the following study proposed an accessible approach to launch an author contribution system via NLP (Natural Language Processing)-based algorithms. The final pipeline is composed of dependency analysis (e.g. spaCy), sentence embedding (Sentence – BERT), and classification models (logistic regression model and neural network). The results are validated based on the sample dataset including 20000 observations and indicate satisfactory performance.

Index Terms

Contributor Roles Taxonomy, NLP, Sentence Embedding, Dependency Analysis, Multi-Target Classification

I. Problem Definition and Progress Overview

THE Elsevier capstone project is sponsored by Elsevier, a Netherlands-based academic publishing company specializing in scientific, technical, and medical content. One of the most celebrated electronic publishers, Elsevier is highly evaluated via not only its enormous and diverse data inventory but its agilely evolving inner systems as well. And CRediT (Contributor Roles Taxonomy) is among these desired functions. By criteria, this system categorizes all possible roles of authors into 14 mutually exclusive clusters. And in actual implementation, this system captures the author contribution sections from published journal papers as inputs, analyzes the raw texts to identify authors and their accomplishments, and eventually finds the role(s) each author plays in a study [1]. Hence, this system could integrate the entire online publishing workflow of Elsevier into an automatic pipeline and reduces manual labor in the labeling process. In addition, visualizing the exact contribution is also a way to address several well-described problems with author lists and orders. Among the most important is to provide more accountability to prevent questionable, guest, and ghost authorship on research papers, and therefore eliminate authorship disputes and improves collaboration [1] [21]. Therefore, concerning CRediT's potential and merits, the objective of this capstone project would be to establish an author contribution system for Elsevier via available NLP modules.



The authors confirm contribution to the paper as follows: study conception and design: Zubayer Islam, Mohamed Abdel-Aty, Qing Cai; data collection: Qing Cai, Jinghui Yuan; analysis and interpretation of results: Zubayer Islam, Mohamed Abdel-Aty

Author Full Name	Author ID
Zubayer Islam	57221288000
Mohamed Abdel-Aty	7006423551
Qing Cai	56323815300
Jinghui Yuan	57139150400

Author Full Name	Author ID	Contribution
Zubayer Islam	57221288000	Conceptualization
Zubayer Islam	57221288000	Formal Analysis
Mohamed Abdel-Aty	7006423551	Conceptualization
Mohamed Abdel-Aty	7006423551	Formal Analysis
Qing Cai	56323815300	Conceptualization
Qing Cai	56323815300	Investigation
Jinghui Yuan	57139150400	Investigation

Fig. 1. The workflow of the author contribution system. The text data and author data enter this system, and the participated authors (marked in azure) and their exact contributions (marked in green) are extracted and organized into a relation table.

The example in 1 above visualizes the entire process and the expected outcome of the author contribution system. It initially takes two types of datasets as inputs: text data, the raw text of the author contribution section in the published paper, and author data, the relational table projecting an author's unique author id to his or her full name. This system afterward seeks each author's contributions from the texts, categorizes them into the 14 existing categories defined by CRediT, and stores these author-contribution correlations in the resulting relational map.

Within this study, both exploratory data analysis and supervised modeling are conducted. The sample data is initially cleaned and pre-processed in an ingestible format to allow for convenient analysis regardless of the specific model being used, and the distribution as well as presenting schema of non-numeric qualitative features are also explored. Besides, since the original sample data lacks ground truth labels for further classification, keyword-matching methods, and word vectorizers are also utilized to label the selected samples for further modeling. Established on the processed and labeled data, two supervised classifiers, a logistic regression model, and a neural network are trained and tested, and in the validation process, both model prototypes display their robust predictions.

Regarding these outcomes, the significance of this study is three-folded. To begin with, the classification results of the NLP-based author contribution system are almost as accurate as those of manual reviews of Elsevier. Therefore, its massive implementation is capable of both largely mitigating the stress of human reviewers and reducing Elsevier's computational as well as labor costs. Besides, this system could also associate with the author credit system and therefore make it accessible to generate authors' accomplishment portfolios automatically. In addition, the exploratory methodology and results generated from the sample dataset could also facilitate Elsevier's staff to better comprehend the distribution of working load in regular journal papers.

II. Dataset and Exploratory Data Analysis

At the preprocessing and explorative phase, a sample dataset including 20000 observations are analyzed. With each observation representing a journal paper published on Elsevier, this dataset is composed of 3 valid features:

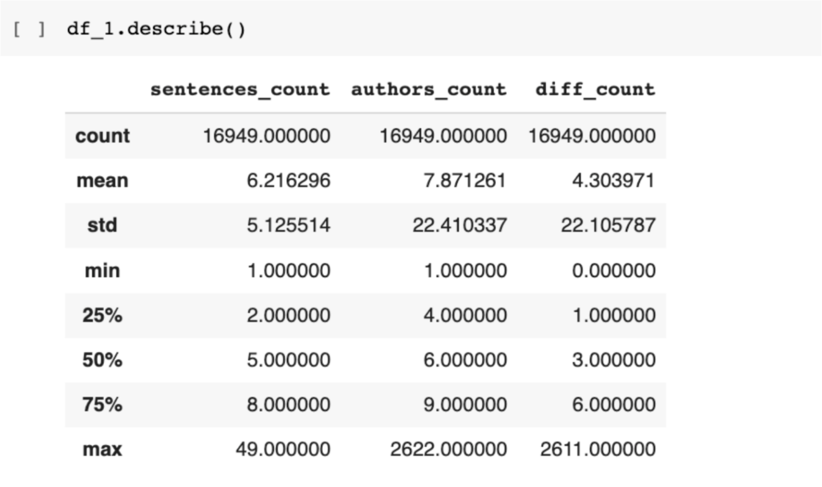
- **PII (Publisher Item Identifier):** the unique ids for the journal paper.
- **Section Content:** the raw text extracted from the author contribution section of the journal paper
- **Authors:** a dictionary containing the author sequence, author id, and author name of each author of the journal paper.

With respect to their characteristics and potential pain points, the authors conducted a related analysis to better interpret them as well as facilitate the modeling process.

A. Basic Data Analysis

At the beginning stage of the data exploration, the fundamental distributions and statistical characteristics of each feature need to be investigated. Thus, the basic analysis shall be conducted from two aspects: the author-wise and content-wise perspectives.

The author feature in the dataset is firstly explored due to its comparatively regular structure and readable format: originally, the authors of each paper contain a list of authors of each paper, which is stored in JSON format. Via simple parsing and filtering methods, the count of authors would be visualized.



```
[ ]  df_1.describe()
```

	sentences_count	authors_count	diff_count
count	16949.000000	16949.000000	16949.000000
mean	6.216296	7.871261	4.303971
std	5.125514	22.410337	22.105787
min	1.000000	1.000000	0.000000
25%	2.000000	4.000000	1.000000
50%	5.000000	6.000000	3.000000
75%	8.000000	9.000000	6.000000
max	49.000000	2622.000000	2611.000000

Fig. 2. The basic statistics of the author feature in the original sample dataset

Figure 2 displayed above indicates the basic statistics of the author feature. It could be summarized that the maximal number of authors is 2622, appearing in the record PII S0007091217300181. In addition, there are around 50 papers authors count over 50. This result corresponds with the explanation of our mentor: in some deep professions such as particle and physics, due to the long research period and diverse fields, massive data are collected as well as processed in these studies. Besides physicists, there are also many engineers who maintain experimental machines, engineers who maintain IT hardware, and numerous software workers who do data screening and analysis. Therefore, it is possible that thousands of authors contribute to one study. Nevertheless, in these cases, the author contribution section in the paper would not explain the accomplishment of individual authors in detail yet the collaborations of organizations instead [17]: For example, the observation PII S0007091217300181 describes the contributions of ISOS and Nestle Health Sciences. As a result, these outliers as well as edge cases will not be concerned in this data exploratory phase, and only the first 99.99 percent of the authors' count (less than 46) dataset will be considered.

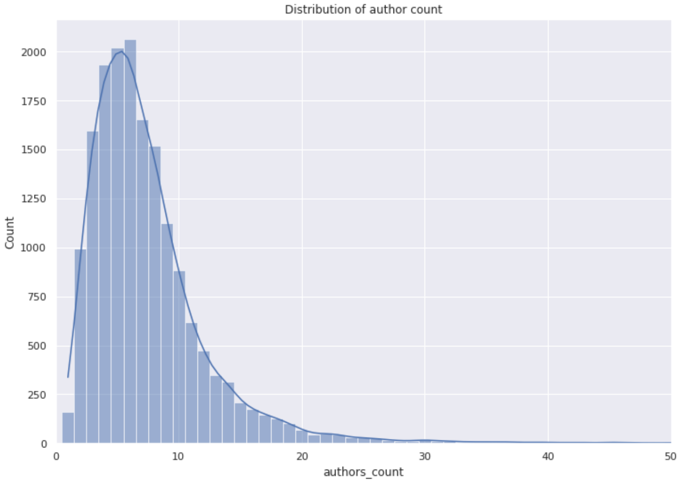


Fig. 3. The distribution of the author feature. It is an obviously right-skewed distribution with 6 having the highest frequency.

Besides the basic statistics, the distribution of the author feature is also visualized. As Figure 3 indicates, the distribution of author number is a right-skewed normal distribution: more than 75 percent of the research papers have 1-9 authors, and papers with 6 authors have the highest popularity. This phenomenon could also be accounted for by the inflating number of authors in recent years, which is triggered by authors' over-specialization and diversification of research topics [15].

With the distribution of the author feature fully understood, the content column, the key feature containing the detailed contribution description, is also investigated. The content column is composed of long strings, which are raw text extracted from the author contribution section of published papers. Since in the coming modeling phase, the texts need to be parsed to facilitate the author-wise keyword extraction, the main goal of this exploratory phase would be to split the content into sentences and visualize the possible sentence counts. In addition, during the visualization, some original raw texts appear to mistakenly employ punctuations, either the content uses a colon or semicolon to separate content, and therefore induce troubles to the parsing operation. Thus, from this vein, the texts with misused punctuations are filtered for the sake of a more accurate conclusion. Therefore, both the distribution of the original dataset and that of the filtered dataset are displayed in Figure 4 and Figure 5 below.

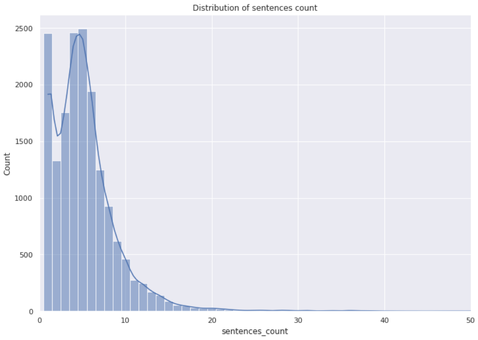


Fig. 4. The distribution of sentence count in the original sample dataset

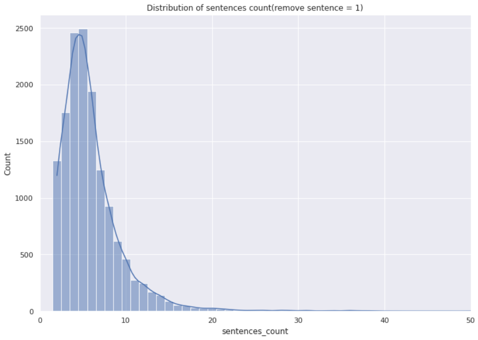


Fig. 5. Interpolation for Data 2

Additionally, a correlational analysis is also conducted between these two core features. According to the correlational map displayed below in Figure 5, it could be concluded that most sentence content counts are very similar compared to the number of authors. And among over 13 percent of the cases, the number of sentences is identical to the author count, and a 50 percent difference is less than 3. Therefore, such a phenomenon as well corresponds to the fact that each sentence in the author contribution section states the accomplished jobs of one author 6. Likewise, in the parsing section later, a one-to-one matching pattern could also be established between these two features.

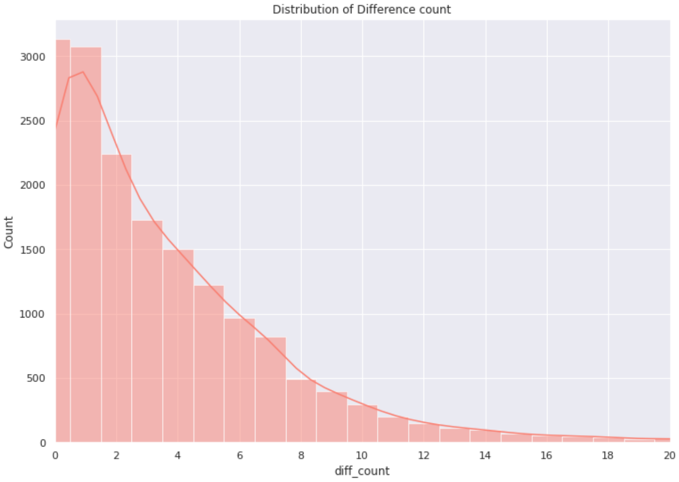


Fig. 6. The correlation between the number of the author and that of the sentence in the content column.

B. Author Name Presenting Schema Analysis

Like most human-centered text mining algorithms practiced in industrial environments, the inconsistent name-ordering, as well as name-presenting schemas, keep hindering the data vectorization [1] [8]. Therefore, since the mining process of this study involves matching the authors' unformatted names in the text section and their full names and their author ids, it is necessary to exhaust all possible name-presenting schema employed in the section content. After numerous filtering and matching operations, the author's name-presenting schema in the sample dataset could be categorized into these mutually exclusive types shown in Table I.

TABLE I
All Author Name Presenting Schemas

Name Presenting Schema	Schema Description
Full Name	Complete given name, middle name, and complete surname.
Initial and Surname	Initials of the given name, middle name, and complete surname. Usually connected by periods.
Initials	Initials of the given name, middle name, and surname. Usually connected by periods, dashes, or commas.
Abbreviations	First letter of the given name, middle name, and surname. No punctuations.
Reversed initials	Initials of surname, middle name, and given name. Usually connected by periods.
Reversed Abbreviations	First letter of the surname, middle name, and given name. No punctuations.

After most possible name-presenting schemas in the sample dataset are detected, a distribution analysis is generated based on their frequency. According to the bar chart Figure 7, it could be concluded that the initials, full names, and abbreviations are the three most popular name-presenting schemas and have much greater frequencies than all other schemas. Such variance could be attributed to the popularity of Harvard, Chicago, and MLE reference formats, which usually employ initials or full names as the in-line citations.

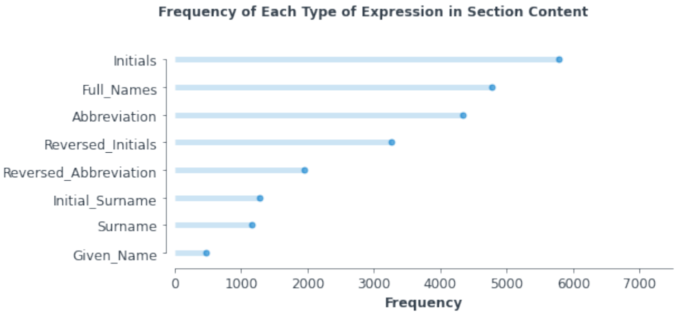


Fig. 7. The frequency of each type of name presenting schema

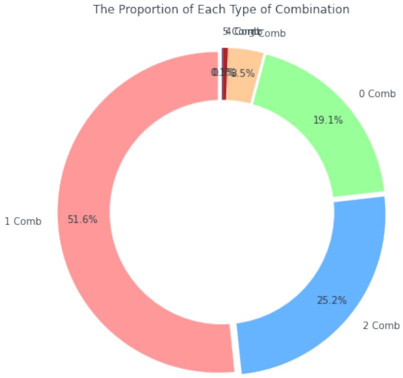


Fig. 8. The proportion of each type of name presenting schema

In addition, regarding the bar chart in Figure 7, it could also be summarized that the total frequency exceeds 20 thousand, the number of observations in the sample dataset since multiple name-presenting schemas could be utilized in one journal paper's author contribution content. Thus, the number of schemas employed in each paper is calculated and visualized in Figure 8. According to this pie chart, most journal papers only exploit 1 or 2 name-presenting schemas, while few of them employ more than 2. Besides, there exists around 20 percent of journal papers do not state any author name in their author contribution section, which could be filtered out in a later process.

After most name-presenting schemas are detected, the names in journal papers are correspondingly replaced by their unique author ids, to improve the text parsing as well as further text mining between observations.

C. Text Analysis

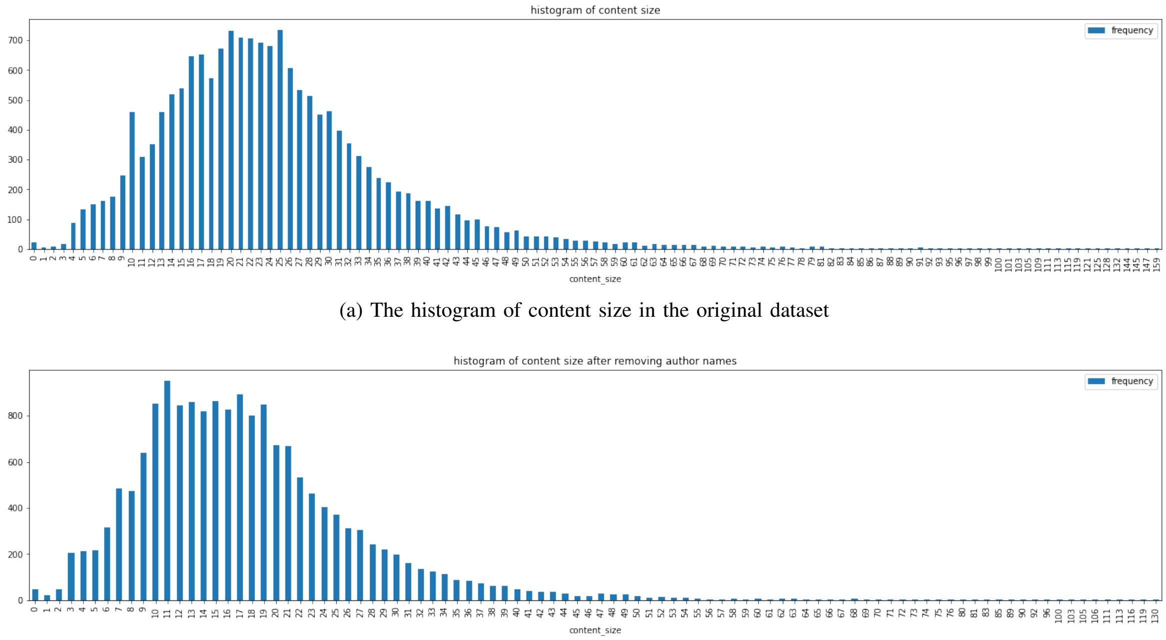
After the basic statistics are fully visualized, the content feature including the raw text data from published papers should also be understood. Before the data visualization, some preprocessing operations would be necessary. First, to preserve the purity of the data and eliminate possibly misleading observations, all the rows containing null content, which is about 15 percent of all data, are filtered from the sample dataset. In addition, to convert the raw texts presented in long strings into a readable and vectorizable format, the content strings are sliced by punctuation. Besides, some unrelated sentences (which do not include a specific relationship between the author and tasks), uninformative terms (some simple and ambiguous acknowledgments), and stop words are also removed. Furthermore, for the sake of consistency, all characters are converted into lowercase letters.

Figure 9 below is a histogram of the content size (number of sentences) in each observation. After removing the punctuation and stop words, the becomes to be a limited vocabulary and the maximum content size is 159. The content size is roughly normally distributed and positively skewed, and generally clusters in the range of 10 32. Consisting with this distribution, in the later parsing and author-identifying process, if the number of resulted sentences exceeds this range, the outcome may be considered problematic.

Likewise, Figure 10 above also indicates the top 30 frequent words in the content feature. It could be seen that the word ’data’ has the highest frequency which appears 21283 times. The top three words are data, manuscript, and analyze, which directly correspond with three clusters of author contributions: investigation, data curation, writing – original draft, and writing – review and editing [19]. Thus, since the key to interpreting author contributions is from the detailed description in the content, the frequency of the terms could also reflect the popularity of each type of author contribution.

To further enhance the entity resolution and remove repetitive words, the content feature is also stemmed, normalized, and author names removed. And the distribution and common words' frequency are displayed in Figure 9 and Figure 11. It could be concluded that the top 30 common words did not alter a lot. However, the content size distribution is much more different and the maximum content size falls to 130 (a decrease of 17.7 percent). Also, there is no interval between 120-129.

Likewise, consistent with the mentor's suggestions, the content feature is also corrected in their spelling via the Contextual Spell Check package in the NLTK library. With respect to the distribution displayed in Figure 12, only 50 percent of the sample content only has 0-1 suggested spelling error. And 75 percent of sample contents' suggested spelling error is within 4. Only a few of the sample has over 10 suggested error. Regarding these contents, most suggested corrections are related to authors' names. These spelling suggestions are not affecting the main content. Therefore, it could be concluded that the spelling error is not an important factor for our project.



(a) The histogram of content size in the original dataset

(b) The histogram of content size in the stemmed and normalized dataset

Fig. 9. These histograms reflect the content size in the original and pre-processed dataset. It could be summarized that in both conditions, the content size is normally distributed and positively skewed. And the main cluster in around 10 to 30.

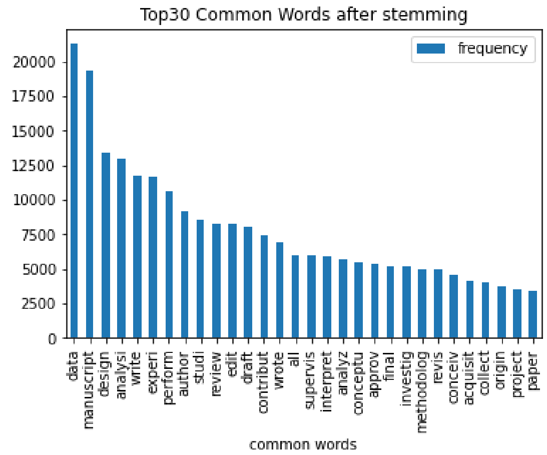


Fig. 10. The distribution of sentence count in the original sample dataset

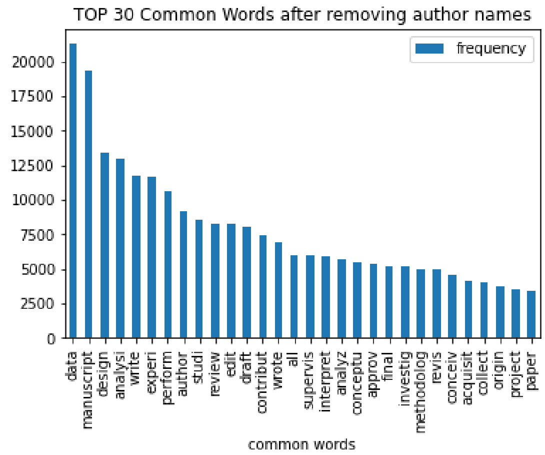


Fig. 11. Interpolation for Data 2

III. Related Work

During the data exploration phase of this project, research on related work has been conducted to examine the feasibility of various future approaches. Consisting with the outcomes of this retrospective research, sufficient instances have exemplified the maturity and advances of keyword extraction techniques and the efficiency of semantic analysis in industrial applications.

To begin with, the effectiveness of the BERT-oriented model could be substantiated by the study of S. Kim [18]. Within this research, authors aim to automatically extract keywords from abstracts of economic papers and therefore proposed a BERT-based protocol. In addition to the conventional bidirectional architecture, authors further improve their understanding of whole sentences via the transfer learning process that involves performing blind tokenization as well as next-sentence prediction. Thus, as keyword extraction is also a critical approach in this project, both the BERT-based model and the transfer learning schema will be inherited.

In addition, some human-centered industrial functions with NLP models implemented moreover suggest some practical approaches. Regarding NLP's applications in the interpretation of patients' medical reports as well as i-Pulse examinations [10] [11], corpus and term dictionary is commonly the key to parsing and recognition. As a result, it will be critical to establish a custom corpus for terms and specific tasks in the authors' contribution.

Furthermore, with the BERT-oriented model the skeleton of the author contribution system, consisting of recent studies, many

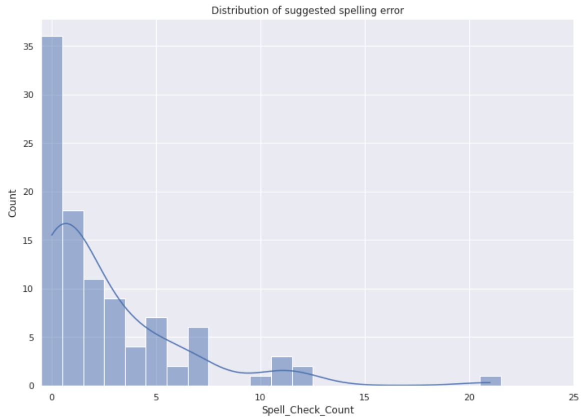


Fig. 12. The distribution of the suggested spelling check

present libraries and algorithms would promote keyword extraction and text comprehension [19] [4]: PageRank algorithm and TextRank algorithm make it available to compare the similarity between sentences or short phrases. Besides, the public spaCy library also facilitates POS (Part of Speech) tagging and component extraction. Thus, these complementary tools could promote the model better generalize already learned patterns to unfamiliar samples and identify the key components in the texts.

To summarize, with respect to previous studies, the effectiveness of the BERT-oriented models, the significance of the customized corpus, and the function of complementary libraries are all validated by plenty of projects, and therefore laid a concrete cornerstone for this author-contribution system.

Nonetheless, this new protocol, rather than simply incorporating all previous accomplishments, still has several pain points. First, while in regular document clustering problems only certain keywords are concerned, as one critical function of this system is to correctly give credit to the contributing authors, the connections between subjects and tasks should not be overlooked and more delicate parsing and extraction schema should be designed. Likewise, compared with the strictly formatted official documents or medical reports, the author contribution sections in an academic paper are highly variable in terminology and diverse in expression, and hence more samples should be manually examined to explore probable expressions and patterns. Consequently, in responding to the unique nature of this problem, more innovative strategies and tools should be applied or invented in later processes to overcome these challenges.

IV. Methodology

A. Overview

The final prototype pipeline of the proposed author contribution system includes the following steps:

- **Data-Preprocessing:** use NLTK packages to clean, stem, and parse the raw texts, and generate the training and testing samples based on the selected contributor roles.
- **Dependency Analysis:** identify the semantic components of the sentences in the datasets and extract the most informative parts with the spaCy library.
- **Word Embedding:** use Sentence-BERT to convert texts into word vectors and numeric values that are model-friendly.
- **Multi-Target Modelling:** employ logistic regression classifiers and neural network classifiers to predict the contributor role of each observation.

The content and format of the output at each stage are displayed in the following figure 13:

B. Data Pre-Processing

With the structure and statistic characteristics of the sample data fully understood in the previous exploratory phase, a series of pre-processing and train-test split operations are conducted as well for the sake of the later modeling phase. The original texts are transformed into model-friendly forms, including vectorized or numeric values, and training and testing datasets are generated. Additionally, to guarantee the generalizability of the final models, the following contributor roles are selected to be the target features of classification:

These contributor roles are selected for multiple reasons, besides the previously collected samples of the “software” cluster, sample texts including “writing – review and editing”, “writing – original draft”, “visualization”, and “funding” are also created. These roles are specially selected for multiple reasons: firstly, the contributor roles “visualization” and “funding” are selected since their keywords are comparatively exclusive and thus easier to search. In addition, as “writing – review and editing” and

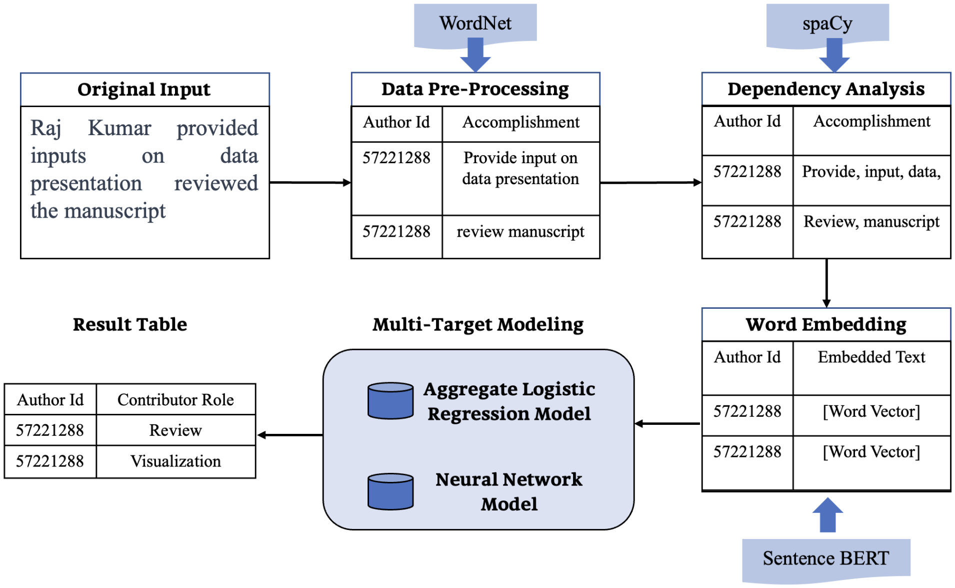


Fig. 13. The flow chart represents the author contribution system's entire working pipeline, which includes the output at each stage.

TABLE II
CONTRIBUTOR ROLE'S KEYWORDS

Contributor Role	Role Description	Keywords
Software	Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.	software, programming, computer, system design, code, algorithms
Visualization	Preparation, creation, and/or presentation of the published work, specifically visualization/data presentation.	plot, graph, fig, image
Writing – original draft	original draft: Preparation, creation, and/or presentation of the published work, specifically writing the initial draft	original draft, original manuscript, cowrite paper
Writing – review and editing	Preparation, creation, and/or presentation of the published work by those from the original research group, specifically critical review, commentary, or revision – including pre- or post-publication stages.	review, edit, critic, censor
Funding acquisition	Acquisition of the financial support for the project leading to this publication	finance, fund, acquisition

"writing – original draft" are semantically similar, analysis of these two contributor roles could also facilitate authors to better understand the data structure and how to discriminate overlapped or similar keywords from the original text.

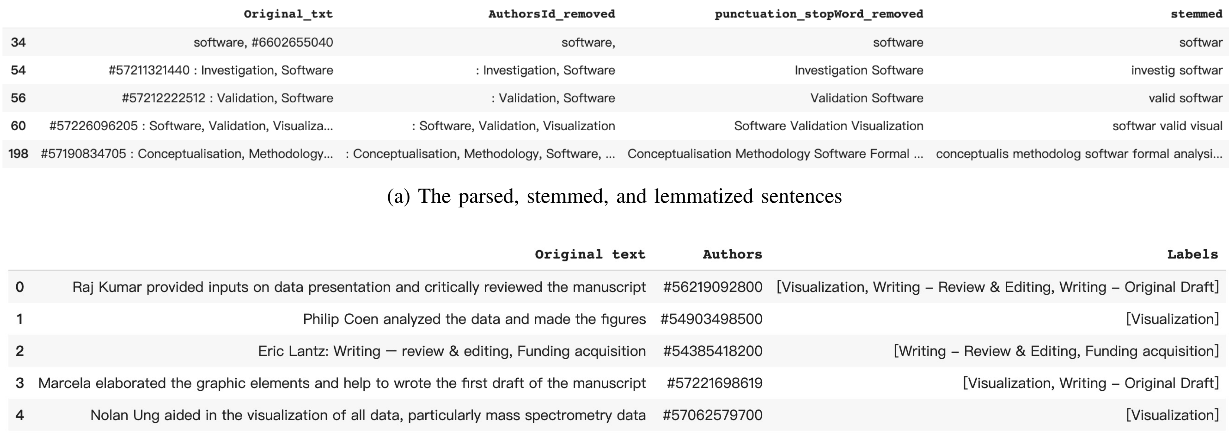
After the key contributor roles are confirmed, the training and testing datasets are then selected and preprocessed accordingly:

- **Name Representation:** Consisting with the presenting schema of authors' names investigated in the first phase, the author names in the raw texts are replaced by their author IDs to both prevent the punctuations within name abbreviations from undermining the parsing, and better link this CRediT system to the authors' accounts (since the author IDs are universal among the system).
- **Author- and Action-Oriented Parsing:** The raw texts are parsed with both punctuations and author IDs, to ensure that one piece of observation exclusively includes one accomplishment of one individual author.
- **Stemming and Lemmatizing:** The author IDs, stop words, and punctuations, are removed from the observations. Likewise,

the words are also lemmatized to promote keyword matching using the WordNet package.

- **Keyword Matching:** For the sake of accuracy and sample purity, the authors initially viewed the sample texts with bare eyes to seek possible keywords related to the selected contributor roles. After all potential keywords are detected, samples are extracted by keywords accordingly based on the table II above displayed. After each extraction, the samples are examined again to ensure all selected observations include the target contributor roles. Otherwise, more filtering conditions will be added.

The outcomes at each stage are displayed in Figure 14 below. For each selected contributor role, a total of 100 samples are selected and input as the training and testing dataset.



	Original_txt	AuthorsId_removed	punctuation_stopWord_removed	stemmed
34	software, #6602655040	software,	software	softwar
54	#57211321440 : Investigation, Software	: Investigation, Software	Investigation Software	investig softwar
56	#57212222512 : Validation, Software	: Validation, Software	Validation Software	valid softwar
60	#57226096205 : Software, Validation, Visualiza...	: Software, Validation, Visualization	Software Validation Visualization	softwar valid visual
198	#57190834705 : Conceptualisation, Methodology...	: Conceptualisation, Methodology, Software, ...	Conceptualisation Methodology Software Formal ...	conceptualis methodolog softwar formal analysi...

(a) The parsed, stemmed, and lemmatized sentences

	Original text	Authors	Labels
0	Raj Kumar provided inputs on data presentation and critically reviewed the manuscript	#56219092800	[Visualization, Writing – Review & Editing, Writing – Original Draft]
1	Philip Coen analyzed the data and made the figures	#54903498500	[Visualization]
2	Eric Lantz: Writing — review & editing, Funding acquisition	#54385418200	[Writing – Review & Editing, Funding acquisition]
3	Marcela elaborated the graphic elements and help to wrote the first draft of the manuscript	#57221698619	[Visualization, Writing – Original Draft]
4	Nolan Ung aided in the visualization of all data, particularly mass spectrometry data	#57062579700	[Visualization]

(b) The labeled stentences

Fig. 14. The upper: The outcome at each stage of data pre-processing: author ID removed, punctuations and stop words removed, and words stemmed. The keywords related to the target contributor roles are finally sought among the stemmed words. The upper, the labeled sentences used for further training and testing

C. Dependency Analysis

Consisting with the fundamental criteria addressed by the CRediT system, multiple contributor roles have almost identical explanations and related tasks in their task descriptions: supervision and project administration both focus on the research activities planning and execution, and writing-original draft and writing-review and editing also commonly overlap with each other [1]. Due to this lexical similarity, a simple keyword-extracting method is barely capable of distinguishing raw texts stating these contributor roles. As the classification model majorly relies on the cosine similarity between the embedded sentences to determine the presence of different contributor roles, such failure in discrimination could result in misclassification and harshly undermine the models' performance.

As a result, rather than certain keywords, multiple grammatic components from the raw texts should be analyzed as a wholesome component to accurately classify them. Hence, dependency analysis and POS tagging are conducted to identify different components from the text before embedding. To better perform the dependency analysis and to detect potential variabilities, such as disjunction between nouns (instead of disjunction tout-court), agentless passive voices, or pairs of indicators in a subordinate clause, a new NLP tool, based on spaCy, are launched to specifically detect variability and identify each sentence component [5] [9].

SpaCy provides an open-source library and neural network model to perform basic NLP tasks. The tasks include processing linguistic features such as tokenization, part-of-speech tagging, rule-based matching, lemmatization, dependency parsing, sentence boundary detection, named entity recognition, similarity detection, etc. [5] The model is stochastic, and the training process is done based on gradient and loss function. The pre-trained model can be used in transfer learning as well.

According to the architecture of spaCy, the keys are Doc and Vocab. Doc contains a sequence of tokens. NLP is the highest level which involves many objects including Vocab and Doc. It has a pipeline to many tasks such as tagger, dependency parser, entity recognition, and text categorization. spaCy typically requires the loading of trained pipelines, to define linguistic annotations - determining, for example, whether a word is a verb or a noun [6]. A trained pipeline typically consists of a few components that use a statistical model trained on labeled data, which makes it accessible to conclude the semantic dependency between tokenized words, and accurately identify the proper nouns, nouns, verbs, adjectives, and adverbs from the original texts. The structure of the tagged sentence is displayed in Figure 15.

It could be found that despite the misleading structure of the processed sentence, both the sentence components and the dependent relationship are visualized. In this study, basically, 3 sentence components are particularly investigated and input to the classification model:

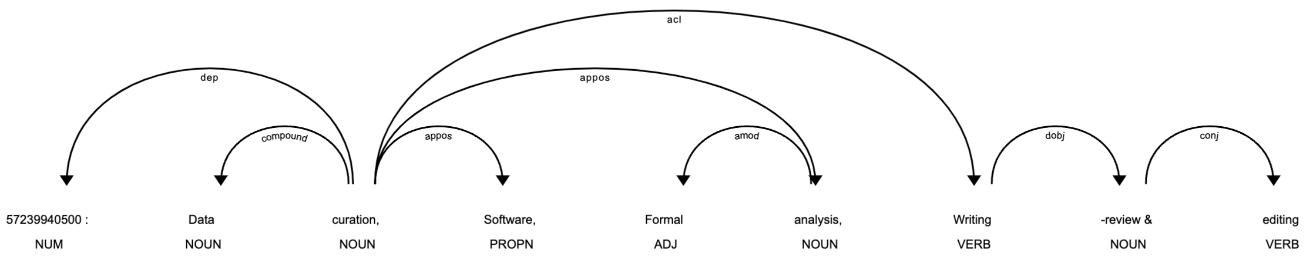


Fig. 15. The output of the tagged sentence. It could be found that after being tagged by the spaCy pipeline, both the semantic components (noun, verb, adj) and the dependent relationship (aci, dobj, etc) are detected.

- **Noun:** since the contribution sections of academic papers are composed of professional terms, proper nouns within them directly link to the exact accomplished tasks and therefore could be especially a concern. In addition if a compound exists, the main component (major objects) will have higher weight in future classification.
- **Verb:** in addition to nouns, verbs in contribution statements are also critical to the disambiguation of contribution roles. For instance, investigation and formal both are related to data, yet the former involves data collection, and the latter involves data analysis. Hence, the verb indicating the exact action could promote the discrimination of these descriptively similar contributor roles.
- **Adjective:** like verbs, adjectives auxiliary to the nouns also play roles in the content disambiguation, such as discriminating the writing – review and editing and writing – original draft. Nevertheless, the weight of the adjective is not as high as those of verbs and nouns.

As a qualitative result, the instance displayed in 15 could be interpreted as the following:

Statement: #57239940500: data curation, software, formal analysis, writing a review and editing

Extracted Components: data, software, analysis, writing, review, editing

Therefore, it could be concluded that the extracted items are more succinct and more related to the exact contributor roles, which could largely enhance the computational efficiency and accuracy of the classifier.

Thus, after the parsing, tokenization, and dependency analysis, these selected sentence components will be embedded and then enter the classification model as the training and testing datasets.

D. Word Embedding

Two embedding methods, TFIDF (term frequency-inverse document frequency) and SBERT (Sentence-BERT) are utilized in this case to vectorize the sentences in the content column. TFIDF is a classic embedding schema that embeds sentences concerning 2 statistics, TF and IDF: TF gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents [12]. This embedding method is selected mainly due to its relatively simple structure and focus on frequency-related information.

On the other hand, BERT and SBERT selected ideal embedding in this study due to their potential to detect the inner semantic textual similarity (STS) between texts, which could guarantee the generalizability as well as the robustness of the conducted model. BERT uses a cross-encoder, two sentences are passed to the transformer network and the target value is predicted [20]. On the basics of BERT, Sentence-BERT (SBERT) adds a pooling operation to the output of BERT to generate a fixed-size sentence embedding vector. The fine-tuning of SBERT involves the Siamese and triplet networks to update the weight parameters so that the generated sentence vector has semantic information [16]. The embedding vectors' distance of sentences with similar semantics is closer, so it can be used for similarity calculation (Cosine similarity, Manhattan Distance, Euclidean Distance).

As the result, the feature space after the TFIDF embedding includes 847 dimensions, and that after the SBERT method includes 784 dimensions.

E. Multi-Target Classification

With the training and testing datasets selected from the sample dataset, they are embedded via Sentence-BERT as the previous section stated [20], and the technical details and advantages of this embedding method would not be reiterated here. On the other hand, though multiple candidate models including the supervised models (logistic regression model and liner regression model) and unsupervised clustering models (KMeans) are all proposed and tested as baseline models, only the logistic regression and the neural network are selected in the final prototype. Firstly, due to the absence of the ground truth label in the original dataset, the target features (the label of contributor roles) are labeled manually. Thus, the limitation in the sample size and the labels' lack of directivity make the unsupervised model inaccessible in this case. In addition, the comparatively low true-positive rate of the linear regression model is unacceptable for the final prototype. Therefore, as figure 13 displays, the final author contribution model will be established with the aggregate logistic regression classifier and the neural network as the core model.

1) The Logistic Regression Classifier: The logistic regression model is a supervised machine learning algorithm intended for binary classification tasks. It uses a logistic function called the sigmoid function which is in the range of 0 and 1. Eq.1 is the sigmoid equation for logistic regression. The sigmoid function generates probability scores based on input features. Simply saying, this model accepts vector as input and generates binary outputs, which matches the goal of classification in this study. During the previous testing for baseline models, the logistic regression classifier displays a comparatively high precision rate of around 0.92. In addition, according to studies on similar research, logistic regression could be utilized in mainly two ways to classify the vectorized texts into multiple targets.

The first approach involves the calculation of the cosine similarity between the vectorized sentences and vectorized corpus. In Vector Space Model, cosine similarity is defined as: "a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Its calculation is very efficient, especially for sparse vectors, as only the non-zero dimensions need to be considered [2]. The cosine similarity's advantages in comparing vectors with different dimensions make it a particularly effective criterion to detect verbal similarity in many NLP studies, as the tokenized and embedded texts or phrases usually vary by size. Within existing anti-plagiarism systems of publishers [13] as well as movie review analytic systems [3], the cosine similarity is both utilized as the core measurement defining whether two sentences or phrases are identical. Therefore, in this study, corpuses consisting of frequent words and phrases of each contributor's roles are created. Then, the cosine similarity is calculated between the BERT-embedded sentences and these corpora. The resulting cosine similarity distances are then input to one single fine-tuned logistic regression model. If the logistic regression model provides a positive output at one cosine similarity, the original embedded sentence will be considered close to the corresponding corpus and be categorized into the contributor role it belongs to.

The second approach involves the usage of an aggregate logistic regression model. In this prototype, 14 simple logistic regression classifiers generating binary outputs construct a large model producing vector output, which each simple classifier detecting the presence of one contributor role. Instead of creating a new measurement such as the cosine similarity, the embedded sentences are input to each classifier directly. Thus, every logistic regression classifier in the aggregate model is trained and tuned separately, with its target contributor role target labeled as positive samples and the rest negative samples. For instance, the classifier for the contributor role "visualization" will only accept training samples with "visualization" labeled as positive ones, and the ones with "software", "funding", and "project administration" are all interpreted negatively. Regarding the research on related studies, this aggregate model overcomes the major limitation of the conventional ones: while conventional logistic regression classifier can respond to one type of target, the aggregate model makes it accessible to classify multi-target data, such as tweets with different moods or circumstances [22]. Another strength of this model is its flexibility over other multi-target classifiers. As online systems in common update rapidly, the previously defined categories and targets sometimes are altered. Hence, while other multi-target classifiers need to be re-trained if additional categories are added, this aggregate model could be updated with ease by adding a new logistic regression classifier responding to the new class [22].

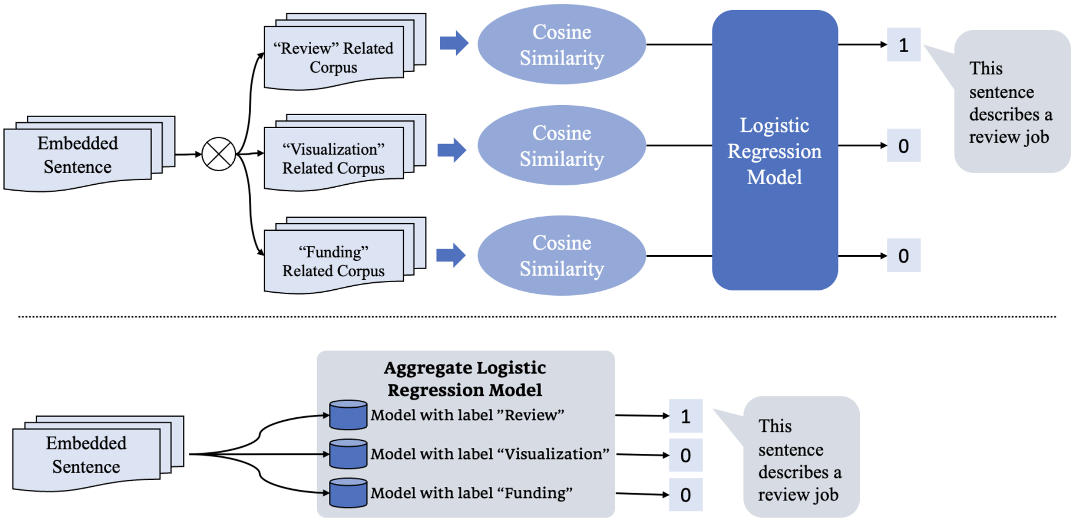

"Review" Related Corpus
"Visualization" Related Corpus
"Funding" Related Corpus
Embedded Sentence
Cosine Similarity
Cosine Similarity
Cosine Similarity
Logistic Regression Model
1
0
0
This sentence describes a review job
Embedded Sentence
Aggregate Logistic Regression Model
Model with label "Review"
Model with label "Visualization"
Model with label "Funding"
1
0
0
This sentence describes a review job

Fig. 16. The illustration of the cosine similarity approach (upper) and the aggregate model approach (lower). Both structures display how an embedded sentence is input to the logistic regression model and classified as a "Review" job.

The structures of two types of logistic regression classifiers are illustrated in figure 16 above. Nevertheless, though both models are constructed and trained at the beginning, due to the lack of existing corpus as well as explainability, the cosine similarity approach is rejected by our mentor and professor. Consequently, only the results generated by the aggregate model are displayed in this report.

2) The Neural Network Classifier: Besides the logistic regression classifier, the neural network classifier is also established as a candidate model due to its performance during the exploratory phase and its advantages in robustness and accuracy. According to the study of Prof. Yang focusing on enhancing the text classification results, he stated the optimal performance of the neural network classifier with BERT implemented: BERT as a pre-trained transformer could be simply transferred to downstream NLP tasks with fine-tuning, which has refreshed records on multiple NLP tasks. It also takes advantage of the self-attention mechanism and builds a multi-layer self-attention network, which matches the multi-layer and multi-channel structure of the neural network classifier [23] [7]. Therefore, since like this study, we also employ Sentence-BERT as the transformer during the word embedding process, the neural network classifier is tested as well in this case. Since the Sentence-BERT already produced high-quality embeddings, we just append a single dense layer to the embedding layer. For the activation function, we used sigmoid. We are performing a multi-label prediction, and traditional activation function like softmax does not work here. For the same reason, we need to use a hyper-parameter threshold to decide what's the decision boundary, which is 0.5 by default. Another thing is that to make our predicted probabilities more reasonable, we used a Mean Squared Error (MSE) loss function.

V. Validation

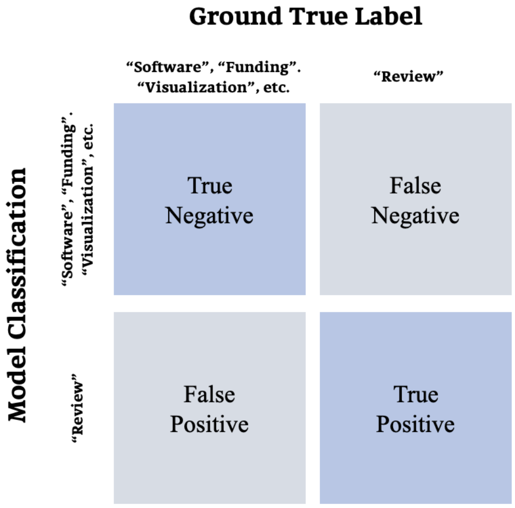
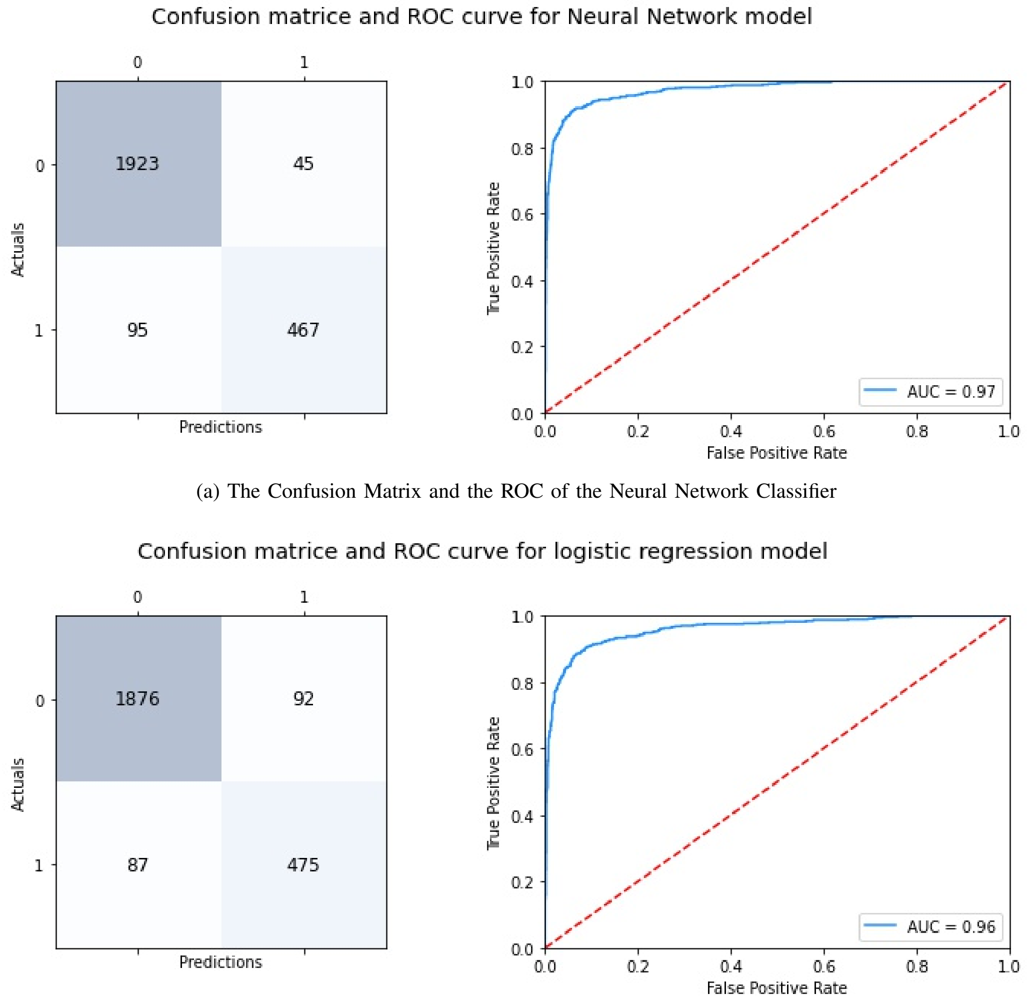


Fig. 17. A sample confusion matrix to evaluate the logistic regression classifier responsible for the contributor role "Review" in the multi-labeled case.

During the previous section, multiple feasible models are proposed to conduct parsing, dependency analysis, word embedding, and classification. Therefore, to better illustrate their effectiveness and robustness, each model is operated end-to-end as a comprehensive system and compared their classification results. Regarding the output structure of the aggregate logistic regression model and the neural network model, the measuring metrics are set to be the true-positive rate and AUC. Nevertheless, established on the data preparation process described in the previous section, the testing dataset including in total of 500 observations from the sample dataset is evenly labeled as "Software", "Visualization", "Writing – original draft", "Writing – review and editing", "Funding acquisition" with 100 entries under each label, and therefore resulted in a multi-labeled and imbalanced dataset. Thus, whereas the AUC and true positive rate could overcome the imbalance issue, they are still not compatible with multiple classes. Thus, as Figure 17 illustrates, a new metric is defined, and the presence of the given label determines the positive and negative label. In addition, due to the limitation in the size of the testing data, the 5-fold cross-validation is employed to guarantee the indiscrimination and reduce possible overfitting.

Figure 18 displays the performance of the aggregate logistic regression model and that of the neural network classifier. According to the label distribution of the testing dataset, the positive samples compose around 20 percent of the testing data, which is supposed to be 500 out of 2500 after the cross-validation. Hence, according to both confusion matrices above, it could be concluded that most of the positive samples are detected, and the rest of the negative samples are also correctly classified. Regarding the illustrated ROC curves, it is as well obvious that both models make exceedingly strong and accurate predictions, with AUC at 0.96 and 0.97.

Concerning these experimental results, the effectiveness and accuracy of both models are validated comprehensively. In addition, it could also be summarized that the neural network classifier has a slightly better outcome than the aggregate logistic



(a) The Confusion Matrix and the ROC of the Neural Network Classifier

(b) The Confusion Matrix and the ROC of the Aggregate Logistic Regression Model

Fig. 18. The confusion matrix and the ROC plot evaluating the performance of the neural network classifier (the upper one) and the aggregate logistic regression model (the lower one)

regression model. Nevertheless, regardless of this overperformance, neither of the two models will be abandoned, since as the previous section stated, the aggregate model possesses the advantages of flexibility and interpretability. Hence, both models will be submitted to the industry for future reference.

VI. Conclusion

In conclusion, according to the descriptions and experimental results displayed above, the authors of this study developed an accurate and cohesive end-to-end pipeline for the author contribution statement of Elsevier. And the aggregated logistic regression models and neural networks are both qualified to correctly classify the authors' contributor roles as human reviewers. To achieve this performance, a series of NLP tools as well as semantic analysis (parsing, tokenization, dependency analysis) are also conducted to facilitate the modeling and classification jobs. Besides the two outstanding prototype classification models (the aggregate logistic regression model and the neural network classifier), this research as well produces a novel methodology to study journal-related texts, including the author- and action-oriented parsing as well as extraction of entities from texts, which could provide future developers with sufficient reference and inspirations.

VII. Future Work

Concerning the limitations and weaknesses of this study, a series of future works could also be accomplished and perfected from multiple aspects.

To begin with, a larger and more comprehensive dataset could extraordinarily enhance the robustness and explainability of the dataset. During the modeling phase, one of the most crucial challenges is the lack of a ground truth label, which forced the authors manually label the observations to train the label. Such author-created labels not only undermine the persuasiveness of the established models but also shrink the sizes of training and testing datasets since authors. Thus, if some historical data including human-reviewed results (the labels added by human reviewers) are accessible, they could certainly enhance the validity of the models as well as their classification results.

Besides, from a boarder perspective, this author contribution system could also associate with the author account system. As the sample output displayed above addresses, all contributor roles are attributed to the author id, a unique account number for each named author on Elsevier. Thus, via interpreting an author's roles and accomplishments within his or her past publications, Elsevier could generate author portfolios accordingly with ease. This portfolio can both evaluate the researcher's academic capabilities and develop his or her strength and weaknesses.

VIII. Ethic Considerations

In addition to the technical advances, ethical issues are also considered and discussed in this research. First and foremost, though the raw texts are all extracted from already published papers permitted to disclose, and their authors are named researchers as well, to preserve the authors' privacy throughout the data analysis and modeling phase, all authors' names are replaced by their author ids correspondingly as a means of data masking.

Nevertheless, the author's contribution developed in this study may still be challenged from ethical and humanistic perspectives. Though AI-based customer service and chatting systems are widely exploited nowadays, machine reviewers for academic papers are still under critical doubt. While during a formal defense of academic research, all reviewers are esteemed and experienced scholars willing to come up with profound questions, machine reviewers, lacking such professional training seem unqualified [14]. Therefore, though the author's contribution statement only composed a short paragraph of an essay, it is still condescending for the authors to be judged by a series of codes and systems instead of humans. Likewise, the accountability of the machine reviewers is as well doubtable. As the capability and prestige of a researcher are exclusively evaluated by his or her past publications, the contributor role he or she plays in these projects is therefore crucial. Hence, if a serious misclassification occurs, the accomplished jobs of a researcher might be ignored, or a critical job will be added to a ghosted author. Both are catastrophic consequences. However, while the human reviewers could be legally accused of such misclassification, the author contribution system may be hard to accuse or blame. To summarize, from the ethical perspective, employing machine reviewers to assess authors' accomplishments, despite their accuracy and indiscrimination, is indeed disrespectful to the authors.

IX. Contribution

- **Yunxiao Wang:** Clean and process data for exploration. Descriptive analysis and visualization. Prepare and label sample data for the baseline model and final model. Draft the report.
- **Jingwen Bai:** Generate sample selection filter. Clean, process, and label data. Draft the report.
- **Xinyu Huang:** Clean, process data and analyze with pos taggings. Create word embeddings, train and evaluate baseline models, Logistic regression models, and Neural Network models with analysis and visualization. Draft the report.
- **Chenxi Jiang:** Clean, process, and analyze data for exploration. Analyze, visualize, optimize the training and evaluate the baseline models, Logistic regression model. Draft the report.
- **Chuyang Xiao:** Team captain. Set up meetings and milestones, and manage the project's progress. Review previous studies and conduct the literature review. Seek theoretical basis accounting for the employed methodology. Clean, process, and analyze with pos taggings. Create figures and flowcharts for the final report. Compose, review, and edit the final manuscript of the report.

Appendix

All Jupyter Notebooks as well as related sample datasets generating these plots and statements could be found in the GitHub link: https://github.com/ChuyangXiao/Capstone-Project-Elsevier-1.

References

[1] Liz Allen, Alison O'Connell, and Veronique Kiermer. "How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship". In: *Learned Publishing* 32.1 (2019), pp. 71–74.

[2] Nunik Destria Arianti et al. "Porter stemmer and cosine similarity for automated essay assessment". In: *2019 5th International Conference on Computing Engineering and Design (ICCED)*. IEEE. 2019, pp. 1–6.

[3] Abdessamad Benlahbib, Achraf Boumhidi, and El Habib Nfaoui. "A Logistic Regression Approach for Generating Movies Reputation Based on Mining User Reviews". In: *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*. 2019, pp. 1–7. DOI: 10.1109/ISACS48493.2019.9068916.

[4] David S Carrell et al. "Clinical documentation of patient-reported medical cannabis use in primary care: Toward scalable extraction using natural language processing methods". In: *Substance Abuse* 43.1 (2022), pp. 917–924.

[5] Chantana Chantrapornchai and Aphisit Tunsakul. "Information extraction on tourism domain using SpaCy and BERT". In: *ECTI Transactions on Computer and Information Technology* 15.1 (2021), pp. 108–122.

[6] Saranlita Chotirat and Phayung Meesad. “Part-of-Speech tagging enhancement to natural language processing for Thai wh-question classification with deep learning”. In: *Heliyon* 7.10 (2021), e08216.

[7] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[8] Liran Einav and Leeat Yariv. “What’s in a surname? The effects of surname initials on academic success”. In: *Journal of Economic Perspectives* 20.1 (2006), pp. 175–187.

[9] Alessandro Fantechi et al. “A spaCy-based tool for extracting variability from NL requirements”. In: *Proceedings of the 25th ACM International Systems and Software Product Line Conference-Volume B*. 2021, pp. 32–35.

[10] Rachit Garg et al. “i-Pulse: A NLP based novel approach for employee engagement in logistics organization”. In: *International Journal of Information Management Data Insights* 1.1 (2021), p. 100011.

[11] Israel Griol-Barres et al. “Detecting weak signals of the future: A system implementation based on Text Mining and Natural Language Processing”. In: *Sustainability* 12.19 (2020), p. 7848.

[12] Maarten Grootendorst. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv preprint arXiv:2203.05794* (2022).

[13] Anggit Dwi Hartanto, Yoga Pristyanto, and Andy Saputra. “Document Similarity Detection using Rabin-Karp and Cosine Similarity Algorithms”. In: *2021 International Conference on Computer Science and Engineering (IC2SE)*. Vol. 1. IEEE. 2021, pp. 1–6.

[14] Antonio Hernández-Blanco et al. “A systematic review of deep learning approaches to educational data mining”. In: *Complexity* 2019 (2019).

[15] Mohammad Hosseini et al. “An Ethical Exploration of Increased Average Number of Authors Per Publication”. In: *Science and Engineering Ethics* 28.3 (2022), pp. 1–24.

[16] Zheng Hu et al. “Semantic-Based Multi-Keyword Ranked Search Schemes over Encrypted Cloud Data”. In: *Security and Communication Networks* 2022 (2022).

[17] E Khozeimeh Sarbisheh et al. “2.1. Author Contribution and Relation to the Research Objectives”. In: *Development of Chelators for Enhancing Radiometal-based Radiopharmaceuticals* (2022), p. 39.

[18] Soojeong Kim, Sunho Choi, and Junhee Seok. “Keyword Extraction in Economics Literatures using Natural Language Processing”. In: *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE. 2021, pp. 75–77.

[19] Srikesh Rajesh Nair et al. “Clustering of Research Documents-A Survey on Semantic Analysis and Keyword Extraction”. In: *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE. 2021, pp. 1–6.

[20] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[21] Drummond Rennie and Annette Flanagin. “Authorship! authorship!: Guests, ghosts, grafters, and the two-sided coin”. In: *Jama* 271.6 (1994), pp. 469–471.

[22] Md Taufiqul Haque Khan Tusar and Md. Touhidul Islam. “A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data”. In: *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*. 2021, pp. 1–4. DOI: 10.1109/ICECIT54077.2021.9641336.

[23] Yiping Yang and Xiaohui Cui. “Bert-enhanced text graph neural network for classification”. In: *Entropy* 23.11 (2021), p. 1536.