Article

Systematic Design and Evaluation of a Citation Function Classification Scheme in Indonesian Journals

Yaniasih Yaniasih 1,2 and Indra Budi 1,*

1 Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia; yaniasih@ui.ac.id
2 Research Center for Informatics, Indonesian Institute of Sciences (LIPI), Bandung 40135, Indonesia
* Correspondence: indra@cs.ui.ac.id

Abstract: Classifying citations according to function has many benefits when it comes to information retrieval tasks, scholarly communication studies, and ranking metric developments. Many citation function classification schemes have been proposed, but most of them have not been systematically designed for an extensive literature-based compilation process. Many schemes were also not evaluated properly before being used for classification experiments utilizing large datasets. This paper aimed to build and evaluate new citation function categories based upon sufficient scientific evidence. A total of 2153 citation sentences were collected from Indonesian journal articles for our dataset. To identify the new categories, a literature survey was conducted, analyses and groupings of category meanings were carried out, and then categories were selected based on the dataset’s characteristics and the purpose of the classification. The evaluation used five criteria: coherence, ease, utility, balance, and coverage. Fleiss’ kappa and automatic classification metrics using machine learning and deep learning algorithms were used to assess the criteria. These methods resulted in five citation function categories. The scheme’s coherence and ease of use were quite good, as indicated by an inter-annotator agreement value of 0.659 and a Long Short-Term Memory (LSTM) F1-score of 0.93. According to the balance and coverage criteria, the scheme still needs to be improved. This research data was limited to journals in food science published in Indonesia. Future research will involve classifying the citation function using a massive dataset collected from various scientific fields and published from some representative countries, as well as applying improved annotation schemes and deep learning methods.

Keywords: citation function; classification scheme; annotator agreement; machine learning; deep learning

1. Introduction

1.1. Background

Traditional citation analysis calculates the importance of a citation based on its frequency in bibliographies, while failing to pay attention to the citation’s meaning in the text. An analysis of a citation’s meaning can be based on the function and sentiment of the citation. The citation’s function is determined by the author’s purpose in citing an article, including providing information in the introduction, comparing results, using or modifying a method, and so forth. Meanwhile, a citation’s sentiment covers the author’s opinion on the cited article, that is, it can be positive, negative, or neutral. Citation function and sentiment analysis have been carried out on the language used in citation sentences utilizing computer-aided text analysis (CATA) (all acronym list can be seen in Appendix A) as a part of the content analysis method [1,2]. The results of such an analysis can be used for essential and perfunctory citation classification [3], scientific mapping [4], information retrieval [5], document summarization [6], recommendations concerning citations [7,8], and so forth. This topic has received a lot of attention from various disciplines, including information science, computer science, and quantitative studies of science.

In contrast to the fixed citation sentiment categories (positive, negative, and neutral), the citation function categories vary widely [9]. New category schemes are created by the
authors or adopted in modified form from previous researchers influenced by the objectives of the analysis, field of study, dataset’s structure, and schema preparation process [6,10,11]. Many systematic and literature-based compilation processes remain unexplored. New schemes must be evaluated before being used in any experiment involving a large amount of data.

This paper is the result of preliminary research into classifying citations and using such classifications to develop in-text citation metrics. Citations can be classified automatically and robustly by machine across all fields of science and article structures. The results are used as a basis for weighting citations and calculating citation metrics. Based on these final objectives, this study aimed to (1) formulate a systematic citation function categorization scheme and (2) evaluate the scheme with criteria commonly used to evaluate language annotation schemes. The evaluation method used a combination of statistics, linguistics, and computer science. The systematic and evidence-based process used to develop and evaluate the categorization scheme is this paper’s contribution.

This paper consists of five sections. The first section contains an introduction and an overview of the related research. Section 2 describes the stages used in the scheme’s development and evaluation. Section 3 presents the results through a brief narrative, tables, and figures. In Section 4, the results are discussed, and comparisons are made with previous research. The last section concludes.

1.2. Related Works

1.2.1. Citation Function Categories: Development and Utilization

Many papers have reviewed the literature on categorization by citation function. This section focuses more on listing categories, understanding the basis for the arrangement, and the utility of selecting various categories. This information will be helpful in developing a categorical scheme based on previous studies.

Various categories of citation functions have been proposed. The earliest typology of citation function was proposed by Moravcsik and Murugesanin in 1975 [12]. They divided the citation function into four pairing categories: conceptual or operational, organic or perfunctory, evolutionary or juxtapositional, and confirmative or negational. These initial categories were then first adopted by Chubin and Moitra [13] to study the content analysis of references to citation counting. In studies using machine learning for citation function classification, the most straightforward scheme consisted of three categories: background/information, methods, and result comparisons [14]. This scheme was chosen because it is related to the structure of scientific articles, applicable for exploring topics, and easy to implement using machine learning (ML). Another scheme consists of four categories, namely (Weak)ness, Compare and Contrast (CoCo), (Pos)itive, and (Neut)ral [11]. Unfortunately, only the CoCo category, which is defined as comparing methods or results, is easy to comprehend. This category is structured to classify the citation and provenance functions simultaneously. The further objectives of the classification system were not explained. Bakhti et al. [15] proposed a new scheme consisting of five citation functions: useful, contrast, mathematical, correct, and neutral. This general categorization scheme can be used by various fields of science, easily distinguished by human annotators, used to share implementation needs, and used for building automatic classification models.

Three studies have divided citation functions into six nearly similar categories. Yousif et al. [7] compiled categories consisting of criticism, comparison, use, substantiation, basis, and neutral. Perier-Camby et al. [10] created a new scheme consisting of background, motivation, uses, extension, comparison or contrast, and future. Zhao et al. [8] divided the functions into use, extend, produce, compare, introduction, and other. Four of the six categories these studies compiled are nearly identical, namely use, compare, extend, and background/introduce. Out of these three studies, only one explains the basis for its selection of categories, that is, it was based on previous literature. However, it did not specify how the selection proceeded. Two of these studies completed the automatic classification process without stating how the resulting classification systems were to be
utilized. One study did explain that the classification scheme was developed to construct a reference recommendation system [7].

Teufel et al. [9] produced a more detailed citation function categorization scheme by translating a general scheme into more specific categories. The scheme consists of four general categories, which are then broken down into 12 further categories. The general categories are weakness, contrast, positive, and neutral. Contrast is then divided into four categories related to the section in which the comparison occurs, such as methods and results. The positive class is divided into six categories that indicate the type of agreement that occurs, such as problems, ideas, and methods. Rachman et al. [6] modified this scheme to create a document summarization system.

The citation function categorization schemes vary according to the purpose of the classification to be carried out. Only a small part of the literature focuses on the basis, development, and evaluation of new schemes. Most of these studies develop schemes used directly in the classification process by humans and machines. Assessments of the results, both in terms of the agreement between annotators and machine accuracy, are not always related to the scheme’s quality, as they can be related to the classification model’s accuracy. The literature [9] that focuses on evaluating schemes first before using them to classify large amounts of data is referred to in this paper.

1.2.2. Classification Scheme Evaluation Methods

Many studies have conducted citation classifications, but not many have focused on evaluating these schemes. In general, the evaluation of a classification scheme in linguistics aims to ensure that the compiled dataset can be used appropriately in the classification process. Seaghdha [16] proposed several criteria for evaluating annotation schemes that are relevant to all classifications related to the meaning of language. The proposed criteria include: (1) the developed categories must take into account, to the greatest degree possible, the characteristics of the dataset; (2) there must be coherence (i.e., clarity) in concepts and category boundaries to prevent overlap; (3) there must be a balanced distribution between classes; (4) there must be ease of use originating from a coherent and simple scheme, supported by the availability of detailed guidelines; and (5) there must be utility, in the sense that categories should be able to provide information that can be used further. Some of the above criteria can be measured quantitatively in a scheme’s evaluation. Coherence and ease of use are evaluated with measurements of the agreement and disagreement between classifiers. Balance can be seen in the resulting distribution of citations, and the effect can be assessed using an automatic classification algorithm. The categories’ coverage can be seen by comparing several schemes ranging from simple to complex against one corpus, but this analysis is rarely done. As for utility, it is subjective because the objectives behind each classification scheme can be different.

Most evaluation studies use measurements that refer to the above criteria. Boldrini et al. [17] evaluated manual categorizations using correlations and agreements between categories. Ritz et al. [18] used a similar method and suggested intensive training of annotators before classifying to reduce the error rate. Teufel et al. [9] evaluated classification schemes using an agreement value between three annotators, calculated and discussed annotation errors in a narrative, and compared machines’ classification results with those of humans. A similar method was used by Palmer et al. [19] and Ovrelid [20] to compare automatic and manual classification results and to check to see which category had the most errors. The results identified which classes are most difficult for humans and machines to identify. Ibanez and Ohtani [21] describe the design, evaluation, and improvement process for classification schemes in more detail. They analyzed the types of disagreements that can arise, improved guidelines, created new schemes to improve reliability, and then created gold standard classifications.
1.2.3. Automatic Classification Evaluation

Automatic classification via ML and deep learning (DL) is a widely used method for evaluating citation classification schemes. DL is increasingly being used, but shallow learning is still used as the primary comparative method. Rachman et al. [6] developed a linear Support Vector Machine (SVM) method to classify citation functions that modified Teufel’s scheme. Using a small amount of data, this method produced an F1-score of 68%. Taskin and Al [22], using an enormous amount of data with the Naïve Bayes Binomial (NB) classifier, achieved an F1-score of 78%. Both these results were better than the results of an experiment conducted by Zhao et al. [23] using the Long Short-Term Memory (LSTM) classifier, which achieved an F1-score of 63%. Bakhti et al. [15] used Convolutional Neural Networks (CNN) to achieve an F1-score of 63%, and Perier-Camby et al. [10], using the BCN model (Biattentive Classification Network) and ELMo (Embeddings from Language Models), achieved an F1-score of 58%. These results show that DL does not always provide superior results; many factors influence the results, such as classification schemes and datasets. However, CNN, LSTM, and their variations are the most widely used classifiers in citation function classification.

DL methods that have recently been used with better results include Recurrent Neural Networks (RNN), CNN, BiLSTM, and a combination of CNN and BiLSTM. Su et al. [11] used CNN and achieved an accuracy of 69%. Cohan et al. [14], using Glove and Elmo on BiLSTM with a scaffolding structure, achieved an F1-score of 67% on a small dataset (1941 sentences). Meanwhile, the score rose to 84% for 11,020 sentences. Experiments with big data (>50,000 sentences) were conducted by Zhao et al. [8], who used the RNN and Bidirectional Encoder Representations from Transformers (BERT) pre-trained models to achieve an F1-score of 78%. Meanwhile, Yousif et al. [7] used CNN and BiLSTM to achieve an F1-score of 88%. This score is among the highest achieved in any reported experiments. This study uses several classifiers used previously to evaluate the proposed classification scheme.

2. Materials and Methods

2.1. Data Collection

A dataset consisting of citation sentences was used for developing a new classification scheme and its evaluation. The data came from Indonesian journals on food science that have met Indonesia’s scientific journal quality standards. In 2017, the Indonesian Ministry of Research and Technology compiled the S-score metric to nationally evaluate the quality of researchers, journals, and research institutions. The results of the assessment are disseminated through Sinta’s database (https://sinta.ristekbrin.go.id/, accessed on 11 April 2020). Journals were assessed and ranked into six indexing categories: S1, S2, S3, S4, S5, and S6. S1 is the highest-ranking category, while S6 is the lowest one. There were two assessment variables, namely article quality and journal management. The quality of the articles included the writing format, content, and citation, while the management variables consist of the review process, publishing, and dissemination qualities [24]. This study used articles from seven journals listed in the S2 categories because no food science journals were indexed in the S1 category. Based on the S-score assessment, all journals were considered good quality. However, they were not indexed in Scopus or Web of Science because not all articles are written in English.

Articles in PDF format from all the journals published in 2019 were parsed using the Grobid application [25]. The parsing resulted in XML-formatted documents broken down according to structure, references, and sentences. The resulting XML file could then be used for various citation analyses, including the classification of citation functions. Sentences containing citations and sentences without citations located before or after citations were collected for classification. A total of 2153 sentences were collected. These sentences were then classified according to function by annotators. The scope and statistics of the dataset are shown in Table 1.
Table 1. The scope and statistics of the dataset.

| No. | Journal Title                                      | Indexing Category | Number of Articles | Number of Citation Sentences |
|-----|---------------------------------------------------|-------------------|--------------------|------------------------------|
| 1   | Jurnal Gizi dan Pangan                           | S2                | 21                 | 473                          |
| 2   | Jurnal Teknologi dan Industri Pangan              | S2                | 21                 | 476                          |
| 3   | Jurnal Aplikasi Teknologi Pangan                 | S2                | 24                 | 432                          |
| 4   | Jurnal Penelitian Pascapanen Pertanian            | S2                | 12                 | 217                          |
| 5   | Jurnal Teknologi & Industri Hasil Pertanian       | S2                | 11                 | 131                          |
| 6   | Advances in Food Science, Sustainable             | S2                | 10                 | 178                          |
| 7   | Warta Industri Hasil Pertanian                   | S2                | 14                 | 245                          |
|     | Total                                             |                   | 113                | 2,153                        |

2.2. Proposing a New Scheme

The dataset consisting of citation sentences was used for proposing a new classification scheme. The development of the new scheme was carried out systematically based on previous research in automatic citation function classification. The first phase was a survey of recent publications to produce a list of existing categories for citation functions. The survey turned up 13 publications, and 48 category names were obtained. In the second stage, groupings were carried out based on similarities in meaning, resulting in eight categories. In the third stage, an analysis of the relationships between groups was carried out. It was found that two groups were sub-groups that could be combined into one group. The results of the survey and grouping were six categories of citation function.

Almost all of the datasets used in the previous literature to construct a citation functions category schema have come from international journals in computer science. In comparison, this research used datasets from food science journal articles in Indonesia. To ensure that the proposed categories described the dataset, citation characteristics in Indonesian journal articles were investigated. Then an analysis was carried out by comparing the six categories with Indonesian journals’ characteristics. As a result, the “other” category was removed. The “other” category came from four category names in the surveyed literature: “produce”, “future”, “other”, and “neutral”. “Produce” is not explained in the source literature, so it was difficult for us to interpret its meaning. The “future” category covers conclusions and is scarce in Indonesian journals because no citations are found in conclusions. As for “neutral”, authors considered it to be less of a function category and more of a sentiment category. Based on these considerations, the proposed classification scheme for citation functions consists of five categories, namely, “use”, “extend”, “compare”, “related”, and “background/information”. The classification development scheme, previous literature used, and their dataset scope are shown in Table 2.

Table 2. Categories of citation functions in the literature and proposed scheme.

| No. | Category Name | Article Sources and Dataset Scope                                                                 | Group                          | Proposed Category        |
|-----|---------------|--------------------------------------------------------------------------------------------------|--------------------------------|--------------------------|
| 1   | background    | [10] 5338 sentences about computational linguistics indexed in The Association for Computational Linguistics (ACL); [14] 52,705 sentences about computational linguistics, biomedical, and life sciences indexed in ACL and Pubmed; [26] 548 sentences from 26 articles about computational linguistics indexed in CmpLg (Computation and Language archive); [27] 1969 sentences from 185 articles about computational linguistics indexed in ACL | background/information         | background/information   |
| 2   | based on/supply | [28] 1432 sentences about computational linguistics indexed in ACL                                                                                   |                                |                          |
| 3   | basis         | [9] 9159 sentences about computer science and medical indexed in Semantic scholar; [29] 2000 sentences from 80 articles about computational linguistics indexed in ACL |                                |                          |
| No. | Category Name          | Article Sources and Dataset Scope                                                                 | Group      | Proposed Category |
|-----|------------------------|--------------------------------------------------------------------------------------------------|------------|-------------------|
| 4   | introduce              | [12] 1562 sentences about computational linguistics indexed in ACL                                |            |                   |
| 5   | present                | [4] 1153 sentences from 50 articles about computational linguistics indexed in ACL                |            |                   |
| 6   | idea                   | [6]                                                                                              |            |                   |
| 7   | problem                | [5] 1153 sentences from 50 articles about computational linguistics indexed in ACL                |            |                   |
| 8   | hedges                 | [24] 384 sentences about computational linguistics indexed in ACL                                 |            |                   |
| 9   | report                 | [4]                                                                                              |            |                   |
| 10  | support                | [9]                                                                                              |            |                   |
| 11  | cocogm                 |                                                                                                  | [9]        |                   |
| 12  | cocor0                 |                                                                                                  | [9]        |                   |
| 13  | cocoxy                 |                                                                                                  | [9]        |                   |
| 14  | compare                |                                                                                                  | [7]        |                   |
| 15  | compare and contrast   | [8, 11] 134, 127 sentences from 20,000 article about natural language processing; [29] 395 articles (science field and indexing information are not available) | compare    | compare           |
| 16  | comparison             | [6, 14, 23] 134, 127 sentences from 20,000 article about natural language processing; [29] 395 articles (science field and indexing information are not available) |            |                   |
| 17  | compare or contrast    |                                                                                                  | [9]        |                   |
| 18  | result comparison      |                                                                                                  | [14]       |                   |
| 19  | contrast               |                                                                                                  | [24]       |                   |
| 20  | criticism              |                                                                                                  | [6]        | criticism         |
| 21  | weak/weakness          |                                                                                                  | [9, 28]    |                   |
| 22  | correct                |                                                                                                  | [10]       |                   |
| 23  | positive               |                                                                                                  | [12]       |                   |
| 24  | similar                |                                                                                                  | [9]        |                   |
| 25  | continuation           |                                                                                                  | [27]       |                   |
| 26  | extend                 |                                                                                                  | [7]        |                   |
| 27  | extension              |                                                                                                  | [10, 27]   |                   |
| 28  | follow                 |                                                                                                  | [4]        |                   |
| 29  | motivation             |                                                                                                  | [10, 27]   |                   |
| 30  | modification           |                                                                                                  | [9]        |                   |
| 31  | proposed-by            |                                                                                                  | [4]        |                   |
| 32  | approach               | [22] 101,019 sentences from 423 articles in library and information science literature in Turkish journals | [4]        |                   |
| 33  | data                   |                                                                                                  |            |                   |
| 34  | data validation        | [22]                                                                                              | use        | use              |
| 35  | definition             | [22]                                                                                              |            |                   |
| 36  | mathematical           | [10]                                                                                                |            |                   |
| 37  | method                 | [4, 14, 22]                                                                                      |            |                   |
| 38  | use                    | [4, 7, 8, 10]                                                                                     |            |                   |
| 39  | useful                 | [10]                                                                                                |            |                   |
| 40  | substantiation         | [6]                                                                                               |            |                   |
| 41  | literature             | [22]                                                                                                |            |                   |
| 42  | see-for detail         | [4]                                                                                               |            |                   |
| 43  | related                | [6]                                                                                               |            | related           |
| 44  | acknowledge            | [28]                                                                                                |            | related           |
| 45  | produce                | [7]                                                                                               |            | other             |
| 46  | future                 | [10]                                                                                                |            | other             |
| 47  | other                  | [7]                                                                                               |            |                   |
| 48  | neutral                | [9]                                                                                               |            |                   |

Dataset scope information is only described once for the literature that appears multiple times.
2.3. Annotation Process

The annotations were carried out by three postgraduate students in the field of food science. They were given annotation guidance, a short training session on category descriptions, and examples of how to do labeling, then they were given two weeks to complete the labeling. The annotation guide that they used contained an overview of citation functions, the definitions of the categories, and a sample table of citation sentences and their labels.

2.4. Evaluation of the Scheme

The evaluation of a classification scheme involves five criteria [16], namely coherence, ease, utility, balance, and coverage. The agreement between annotators and the F1-scores of several machine classifiers was used to assess the coherence and ease of annotation. The balance, coverage, and utility measures were linked qualitatively with the citation patterns in Indonesian journals and the plan to use the classification results.

The agreement between annotators was calculated using Fleiss’ kappa. Fleiss’ kappa is a method for determining the level of agreement between two or more annotators. The kappa value is a coefficient to evaluate the reliability of human coders [2,30]. All kappa values are between \(-1\) and \(+1\), where a value of \(-1\) indicates no agreement at all. In the range from 0 to 1, the strength of agreement increases from low agreement to perfect agreement [31]. The confidence interval for kappa and values indicating statistical significance or the lack thereof (Z and \(p\)-values) can be used as companion measures.

Automatic citation classifications use ML and DL. Both learning algorithms are based on mathematical models that study patterns in data to perform the classification. Current ML models are easy to use in data analyses related to natural language. NB and SVM are two algorithms widely used for language classification, such as sentence sentiment analyses involving news, social media, and product reviews. These two methods have also been used to classify citation functions [6,22].

The ML classifiers used in this study were the SVM, NB, Logistic Regression (LR), and Random Forest (RF) classifiers. Before being included in the classification process, the citation sentences were pre-processed to eliminate stop words, numbers, punctuation, and stemming. The features used for all the classifiers above were n-gram, weighted term frequency-inverse document frequency (TF-IDF), and vectorization. These three features are used to get at the most meaningful words, which can improve classification accuracy.

Apart from ML, experiments were also carried out with DL methods that other researchers have used for citation classification, namely CNN and LSTM. CNN was developed to process images with a sliding window, which was used to find the most informative local aspects. For text classification, words in sentences are converted into word vectors in a process known as word embedding. The information is then collected in the pooling layer. At the end stage, information is passed through the layer to perform tasks, such as classification [32].

This experiment used the Conv1D model, which has a fully connected layer, ReLu activation, and an output layer with a softmax activation function. The parameters values used included embedding_dim = 64, max_length = 50, filter = 50, kernel = 5, batch_size = 64, and epoch = 10. The second method used was the LSTM model, a variation of RNN in which the model has a feedback connection and memory to recall information during the training process. LSTM is suitable for sequential data, such as words in sentences. The LSTM layer model used 64 memory units and an output layer with a softmax activation function. The parameters values used were embedding_dim = 5000, batch_size = 52, and epoch = 5. The model evaluation used the K-fold cross-validation method with accuracy, precision, recall, and F1-score metrics.

3. Results

Many citation function classification schemes are available for adoption. To determine which one was the most appropriate for our use required some consideration. First, the existing schemes had to be studied to obtain an overview of, and inspiration concerning, the
available schemes. Second, the characteristics of the dataset had to be checked against the categories to determine whether the categories were suitable or not. Third, the categories had to match up with the intended purpose of the classification scheme. Considering these three points, a classification scheme with five categories to classify citation functions in journal articles published in Indonesia was decided. Definitions were developed for each category based on the literature and randomly checked using the dataset’s sentences. Table 3 displays the five classes and their definitions.

Table 3. The proposed citation function classification scheme for the Indonesian journal dataset.

| No. | Category          | Definition                                                                                                                                 |
|-----|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | Use               | Articles that are cited are used directly in research, such as concepts, methods, data, formulas, and algorithms. Generally indicated by the words use, used, based, and so forth. |
| 2   | Extend            | The cited article is used as a basis and is added to in the research. Characterized by the words continue, develop, follow, adopt, modify, be motivated, and so on. |
| 3   | Compare           | The articles cited are used as a comparison (especially the results) to the current information. The citations are characterized by the words compared, appropriate, similar, similarity, confirmation, different, weaknesses, advantages, contrast, inline, contradicting, higher, lower, similar, and so forth. |
| 4   | Related           | Articles that are cited have links or discuss the materials, methods, processes, or results specifically/in detail. Words that identify these citations include research (-previous, -other), related, whereas, have been done, reported, defined, according to, literature, and so on. |
| 5   | Background/Information | Articles cited provide general information or theory (not specific to the field of science) to introduce, exemplify, provide background on, and explain the context, problems, objectives, and so forth. |

Citation sentences were labeled using the above five categories. The results of the human classifications were tabulated and the numbers of citations in each category recorded (Figure 1). Then the reliability of agreement between the three annotators was measured using Fleiss’ kappa and a statistical significance threshold of 0.05 (Table 4). The results show that the overall value of kappa was 0.64. This value illustrates that the strength of the agreement among the three annotators was quite good. The individual kappa values per category were above 0.80 for the “use”, “extend”, and “compare” categories. This finding shows that the agreement for these three categories was very good. Meanwhile, for the “related” and “background/information” categories, the kappa values were 0.566 and 0.598, which means that the agreement among the annotators was moderate [31]. Overall and per-category $p$-values < 0.05 reflect that the kappa coefficient was statistically different from 0 (the agreement was conclusively different from 0). The kappa value fell between the lower and upper $\alpha$ cut-offs of 0.05, indicating that the agreement fell within the 95% confidence interval.

Figure 1. The number of citations in each category.
Classification using ML obtained almost the same F1-score for all classifiers. Multinomial NB produced an F1-score of 0.61, SVM achieved 0.63, and LR and RF both reached 0.65, where the accuracy of RF at 0.66 was 0.1 higher than that at LR (0.65). The experimental results using CNN did not provide much better results than those using traditional ML, which had an F1-score of 0.66. Meanwhile, LSTM classification provided excellent results, with an F1-score reaching 93%, much higher than the scores obtained by CNN and all other methods. The automatic classification results are shown in Table 5.

A more detailed analysis per category of the classifiers with the highest F1-scores for ML and DL can be seen in Table 6. Low precision and recall values may indicate a misclassification in a particular category. Low precision means a high false-positive rate, while low recall means a high false-negative rate. The RF classification results show that “background/information” and “related” are the two categories with the lowest precision. Thus, many sentences were placed in these categories when they should have been placed elsewhere. The lowest recall values were obtained for “use” and “compare”, meaning that a large percentage of sentences in these categories were classified in other categories. As for the LSTM classification results, the lowest precision value obtained was for “use”, and the lowest recall value was registered for “compare”.

The higher percentage of errors for ML compared to DL indicated that an analysis of the configuration matrix was necessary (Figure 2). This matrix can indicate unclear concepts or overlaps between categories. Figure 2 shows which sentences categorized as “background/information” were most often confused with those in the “related” class and vice versa. More “use” sentences were classified as “background/information” than “use”, resulting in a low recall value for “use”. Many sentences in the “extend” category were
misclassified and placed in the “background/information” and “related” categories. For the “compare” category, most errors placed these sentences in “extend”.

Figure 2. Configuration matrix.

4. Discussion

The classification of citation sentences in Indonesian journal articles placed citations in five categories. The first evaluation of the proposed scheme examined its utilization. The categories “background/information”, “related”, “use”, “extend”, and “compare” have a close relationship with the structure of journal articles, in general. The citations are used as general background and for providing general information in the introduction. “Related” citations show up in the related work section or introduction, as well as in the discussion sections of articles without literature review chapters, such as articles following the IMRAD structure. The “use” and “extend” citations are found in the methods section if methods and data are used or modified but can also be introduced if the author uses an idea or concept from the article being cited. Citations for comparison are present in the results, although there are also citations in the results that support the results. The relationship between citation function and section location has not been widely explored. When viewed in terms of the linkages with the sections, this scheme’s categories still seem to overlap and need to be analyzed further. However, this linkage also indicates that this categorization scheme supports the argument of previous studies that the structure of the article and the citation location are closely related to the function and value of a citation [33,34].

The human classification results show that categories are very unbalanced (Figure 1). “Background/information” and “related” contain many more citations than the other three categories. It is possible that the imbalance in the categories disrupts an automatic classification process [35]. The large differences among categories can be reduced by combining or splitting categories. The “extend” category could potentially be combined with “use” because extending is a modification of use. The positions of the citations in these categories are also not too different, as both types of citation are found mostly in the methods section and less often in the introduction or related work section. For this reason, the scores for these two categories should not be too different when citations are weighted based on location [36,37].

The second alternative is to break the “background/information” category into two more specialized categories. Background includes citations that provide an introduction to ideas or problems in the introduction, and the article cited is not always directly related to,
or on same topic as, the current article. Meanwhile, information is not always found in the introduction. In Indonesia’s food journal articles, many citations function as information to clarify or support the results in the discussion section. This information can be general and is not always on the same, or close to, the topic of the current article in terms of relevance. For example, when a food research product turned brown, the author stated that “this is also explained in [cited article]; the reactions between reducing sugars and amino acid groups can cause browning in fried foods”. The citation here supports or explains why a product can brown using general theories in the food domain. Although the differences are tiny between background and information, they can be separated into two categories.

Whether or not some categories overlap can be seen more clearly from the human classifier agreement level and the automatic classification. The overall agreement level is good (kappa = 0.65), although it is still below the agreement obtained by Teufel (kappa = 0.72) [9]. The agreement on the “use”, “extend”, and “compare” categories is very good (above 80%, in each case) and medium for “background/information” (below 60%). This human classification shows low overlap, even if there were some difficulties encountered in “related” and “background/information” citations. This result is somewhat different from the automatic classification results. For the best ML results, the overall agreement is still low (0.65%), and is lower than the agreement found in other research [6,22].

Meanwhile, per category, the precision is mostly good, except for the “related” and “background/information”, categories, which have precisions below 60%. The recall values are low, except for the “related” category. These results indicate that there are still many difficulties encountered in classification with traditional ML. In addition to containing definitions, the guide for manual classification also mentions several keywords per category that serve as guidelines. However, not all sentences will contain one of those keywords. Besides, certain verbs appear frequently that are not specific to one particular category. This condition is similar to other studies regarding the lexical distribution of in-text citation, where some verbs are not specific and can appear several times for different functions and locations [34]. For humans, even if certain words are used, the context in which the citation sentence is placed can help determine the category. However, context cannot be used in ML because ML only uses basic features, such as TF-IDF. Very different results were given by LSTM, where the accuracy, precision, and recall values were excellent and higher than the values for all previous studies using DL [7,8,11,14]. These results provide optimism in terms of using the LSTM classifier for the classification of larger amounts of data. The relatively small number of categories used here compared to other classification schemes could have also contributed to ease of classification when using DL.

5. Conclusions

Citation sentence classifications are increasingly being used but have seldom been evaluated in terms of the division into categories. Such an evaluation is essential before humans and machines classify a large amount of data. This study develops a citation function classification scheme using Indonesian food science journals, then evaluates it quantitatively and qualitatively.

The evaluation results are quite good in terms of utility and coherence but still indicate some weaknesses in the proposed scheme. The association of the categories with the structures and patterns of the journal articles in the dataset should be examined for future research. The categories are also not balanced, so it is recommended to combine the “use” and “extend” categories and divide the “background/information” category into two separate categories. However, if these categories are used for automatic classification using DL, the results are outstanding.

The utilization of the findings is still limited because it only uses a dataset from Indonesian food science journals. Further research should include assembling a larger dataset and expanding its scope to include more scientific fields and published journals representing other countries. Larger amounts of data and balanced classes promise good evaluation results.
Author Contributions: Conceptualization, Y.Y. and I.B.; methodology, Y.Y.; software, Y.Y.; validation, Y.Y. and I.B.; formal analysis, Y.Y.; investigation, Y.Y.; resources, Y.Y.; data curation, Y.Y. and I.B.; writing—original draft preparation, Y.Y.; writing—review and editing, Y.Y. and I.B.; visualization, Y.Y.; supervision, I.B.; project administration, Y.Y. and I.B.; funding acquisition, Y.Y. and I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by research grants from Universitas Indonesia (Hibah Publikasi Doktoral Tahun 2021 and Publikasi Terindeks International (PUTI) Doktor 2020 No: BA-733/UN2.RST/PPM.00.03.01/2020).

Data Availability Statement: Raw data were generated at Sinta (Science and Technology Index) developed by Ministry of Research Technology, Republic of Indonesia (https://sinta.ristekbrin.go.id/). Derived data supporting the findings of this study are available from the corresponding author on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Acronym List.

| No. | Acronym | Explanation |
|-----|---------|-------------|
| 1   | ACL     | The Association For Computational Linguistics |
| 2   | BCN     | Biattentive Classification Network |
| 3   | BERT    | Bidirectional Encoder Representations From Transformers |
| 4   | CATA    | Computer-Aided Text Analysis |
| 5   | CNN     | Convolutional Neural Networks |
| 6   | DL      | Deep Learning |
| 7   | ELMo    | Embeddings From Language Models |
| 8   | IMRAD   | Introduction Method Result and Discussion |
| 9   | LR      | Logistic Regression |
| 10  | LSTM    | Long Short-Term Memory |
| 11  | ML      | Machine Learning |
| 12  | NB      | Naïve Bayes Binomial |
| 13  | RF      | Random Forest |
| 14  | RNN     | Recurrent Neural Networks |
| 15  | SVM     | Support Vector Machine |
| 16  | TF-IDF  | Weighted Term Frequency-Inverse Document Frequency |

References

1. Abu-jbara, A.; Ezra, J.; Radev, D. Purpose and Polarity of Citation: Towards NLP-Based Bibliometrics. In Proceedings of the NAACL-HLT, Atlanta, GA, USA, 9–14 June 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 596–606.

2. Neuendorf, K.A. The Content Analysis Guidebook, 2nd ed.; SAGE Publications: Thousand Oaks, CA, USA, 2017.

3. Hussain, S.J.; Maqsood, S.; Jhanjhi, N.; Khan, A.; Supramaniam, M.; Ahmed, U. A Comprehensive Evaluation of Cue-Word-Based Features and In-Text-Citation-Based Features for Citation Classification. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 209–218. [CrossRef]

4. Hassan, S.U.; Safder, I.; Akram, A.; Faisal, K. A Novel Machine-Learning Approach to Measuring Scientific Knowledge Fows. Scientometrics 2018, 116, 973–996. [CrossRef]

5. Ferrod, R.; Schifanella, C.; Caro, L.D.; Cataldi, M. Disclosing Citation Meanings for Augmented Research Retrieval and Exploration. Lect. Notes Comput. Sci. 2019, 11503, 101–105.

6. Rachman, G.H.; Khodra, M.L.; Widyantoro, D.H. Classification of Citation Sentence for Filtering Scientific References. In Proceedings of the 4th International Conference on Information Technology, ICTISEE, Yogyakarta, Indonesia, 20–21 November 2019; Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada (DTETI UGM): Yogyakarta, Indonesia, 2019; pp. 347–352. [CrossRef]

7. Yousif, A.; Niu, Z.; Chambua, J.; Khan, Z.Y. Multi-Task Learning Model Based on Recurrent Convolutional Neural Networks for Citation Sentiment and Purpose Classification. Neurocomputing 2019, 335, 195–205. [CrossRef]

8. Zhao, H.; Luo, Z.; Feng, C.; Zheng, A.; Liu, X. A Context-Based Framework for Modeling the Role and Function of On-Line Resource Citations in Scientific Literature. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5206–5215.
9. Teufel, S.; Siddharthan, A.; Tidhar, D. An Annotation Scheme for Citation Function. In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, Sydney, Australia, 15–16 July 2006; Association for Computational Linguistics: Stroudsburg, PA, USA, 2006; pp. 80–87.

10. Perier-Camby, J.; Bertin, M.; Atanassova, I.; Armetta, F. A Preliminary Study to Compare Deep Learning with Rule-Based Approaches for Citation Classification. In Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2019), Cologne, Germany, 14 April 2019; Sun SITE Central Europe (CEUR) Technical University of Aachen (RWTH): Aachen, Germany, 2019; pp. 125–131.

11. Su, X.; Prasad, A.; Kan, M.; Sugiyama, K. Neural Multi-Task Learning for Citation Function and Provenance. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2–6 June 2019; IEEE Press: New York, USA, 2019; pp. 394–395.

12. Moravcsik, M.J.; Murugesan, P. Some Results on the Function and Quality of Citations. Soc. Stud. Sci. 1975, 5, 86–92. [CrossRef]

13. Chubin, D.E.; Moitra, S.D. Content Analysis of References: Adjunct or Alternative to Citation Counting? Soc. Stud. Sci. 1975, 4, 423–441. [CrossRef]

14. Cohan, A.; Ammar, W.; Van Zuylen, M.; Cady, F. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3586–3596.

15. Bakhti, K.; Niu, Z.; Nyamawe, A.S. Semi-Automatic Annotation for Citation Function Classification. In Proceedings of the 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Prague, Czech Republic, 19–21 May 2018; IEEE Press: New York, USA, 2019; pp. 43–47.

16. Séaghdha, D.O. Designing and Evaluating a Semantic Annotation Scheme for Compound Nouns. In Proceedings of the 4th Corpus Linguistics Conference (CL-07), Birmingham, UK, 27–30 July 2007; University of Birmingham: Birmingham, UK, 2007; pp. 1–17.

17. Boldrini, E.; Balahur, A.; Martinez-Barco, P.; Montoyo, A. EmotiBlog: An Annotation Scheme for Emotion Detection and Analysis in Non-Traditional Textual Genres. In Proceedings of the International Conference on Data Mining, IEEE, Las Vegas, NV, USA, 13–16 July 2009; IEEE Press: New York, NY, USA, 2009; pp. 491–497.

18. Ritz, J.; Dipper, S.; Michael, G. Annotation of Information Structure: An Evaluation across Different Types of Texts. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco, 28–30 May 2008; European Language Resources Association (ELRA): Luxemburg, 2008; pp. 2137–2142.

19. Palmer, A.; Carr, C.; Robinson, M.; Sanders, J. COLD: Annotation Scheme and Evaluation Data Set for Complex Offensive Language in English. J. Lang. Technol. Comput. Linguist. 2020, 34, 1–28.

20. Øvrelid, L. Empirical Evaluations of Animacy Annotation. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, 30 March–3 April 2009; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 630–638.

21. Ibanez, M.P.V.; Ohtani, A. Annotating Article Errors in Spanish Learner Texts: Design and Evaluation of an Annotation Scheme. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, Phuket, Thailand, 12–14 December 2014; Chulalongkorn University: Bangkok, Thailand, 2014; pp. 234–243.

22. Taskin, Z.; Al, U. A Content-Based Citation Analysis Study Based on Text Categorization. Scientometrics 2017, 114, 335–357. [CrossRef]

23. Zhao, H.; Feng, C.; Luo, Z.; Ye, Y. A Context-Based Framework for Resource Citation Classification in Scientific Literatures. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; Association for Computing Machinery: New York, NY, USA; pp. 1041–1044.

24. Lukman, L.; Lukman, L.; Dimyati, M.; Rianto, Y.; Subroto, I.M.; Sutikno, T.; Hidayat, D.S.; Nadhiroh, I.M.; Siastaen, D.; Haviana, S.F.; et al. Proposal of the S-Score for Measuring the Performance of Researchers, Institutions, and Journals in Indonesia. Sci. Ed. 2018, 5, 135–141. [CrossRef]

25. Lopez, P. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In Research and Advanced Technology for Digital Libraries; ECML 2009. Lecture Notes in Computer Science; Agostinelli, F.; Borbinha, J., Kapidakis, S., Papastefanou, C., Tsakonas, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5714, pp. 473–474.

26. Ahmad, R.; Afzal, M.T.; Qadir, M.A. Pattern Analysis of Citation Anchors in Citing Documents for Accurate Identification of In-Text Citations. IEEE Access 2017, 5, 5819–5828. [CrossRef]

27. Jurgens, D.; Hoover, R.; Mcfarland, D. Citation Classification for Behavioral Analysis of a Scientific Field. arXiv 2016, arXiv:1609.00435.

28. Alvarez, M.H.; Martinez-barco, P. Citation Function, Polarity and Influence Classification. Natl. Lang. Eng. 2017, 23, 561–588. [CrossRef]

29. Nanba, H.; Kando, N.; Okumura, M. Classification of Research Papers Using Citation Links and Citation Types: Towards Automatic Review Article Generation. In Proceedings of the 11th ASIS SIG/CR Classification Research Workshop, Chicago, IL, USA, 12–16 November 2000; American Society for Information Science and Technology: Maryland, USA, 2000; pp. 117–134.

30. Krippendorff, K. Content Analysis: An Introduction to Its Methodology, 2nd ed; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2004.
31. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]

32. Lane, H.; Howard, C.; Hapke, H. *Natural Language Processing in Action*; Manning Publication Co.: New York, NY, USA, 2019.

33. Bertin, M.; Atanassova, I.; Gingras, Y. The Invariant Distribution of References in Scientific Articles. *J. Assoc. Inf. Sci. Technol.* **2016**, *67*, 164–177. [CrossRef]

34. Bertin, M.; Atanassova, I. A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. In Proceedings of the 1st Workshop on Bibliometric-Enhanced Information Retrieval Co-located with 36th European Conference on Information Retrieval (ECIR 2014), Amsterdam, The Netherlands, 13–16 April 2014; Sun SITE Central Europe (CEUR) Technical University of Aachen (RWTH): Aachen, Germany, 2014; Volume 1567, pp. 14–25.

35. Ramyachitra, D.; Manikandan, P. Imbalanced Dataset Classification and Solutions: A Review. *Int. J. Comput. Bus. Res.* **2014**, *5*, 4.

36. Maricic, S.; Spaventi, J.; Pavicic, L.; Pifat-mrzljak, G. Citation Context versus the Frequency Counts of Citation History. *J. Am. Soc. Inf. Sci.* **1998**, *49*, 530–540.

37. Khan, A.M.; Shahid, A.; Afzal, M.T.; Nazar, F.; Alotaibi, F.S.; Alyoubi, K.H. SwICS: Section-Wise In-Text Citation Score. *IEEE Access* **2019**, *7*, 137090–137102. [CrossRef]