A sentence classification framework to identify geometric errors in radiation therapy from relevant literature

Basu, Tanmay; Goldsworthy, Simon; Gkoutos, Georgios V.

DOI:
10.3390/info12040139

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Basu, T, Goldsworthy, S & Gkoutos, GV 2021, 'A sentence classification framework to identify geometric errors in radiation therapy from relevant literature', *Information*, vol. 12, no. 4, 139.
https://doi.org/10.3390/info12040139

information

Article

A Sentence Classification Framework to Identify Geometric Errors in Radiation Therapy from Relevant Literature

Tanmay Basu [1,2,3,4,*], Simon Goldsworthy [5] and Georgios V. Gkoutos [1,2,4,5,6,7]

1 Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK; g.gkoutos@bham.ac.uk
2 University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TH, UK
3 Institute of Translational Medicine, University Hospitals Birmingham, Birmingham B15 2TH, UK
4 MRC Health Data Research UK (HDR UK), Midlands Site, Birmingham B15 2TT, UK
5 Department of Radiotherapy, Somerset NHS Foundation Trust, Somerset TA1 5DA, UK; Simon.Goldsworthy@somersetft.nhs.uk
6 NIHR Experimental Cancer Medicine Centre, Birmingham B15 2TT, UK
7 NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham B15 2TT, UK
* Correspondence: welcometanmay@gmail.com

Abstract: The objective of systematic reviews is to address a research question by summarizing relevant studies following a detailed, comprehensive, and transparent plan and search protocol to reduce bias. Systematic reviews are very useful in the biomedical and healthcare domain; however, the data extraction phase of the systematic review process necessitates substantive expertise and is labour-intensive and time-consuming. The aim of this work is to partially automate the process of building systematic radiotherapy treatment literature reviews by summarizing the required data elements of geometric errors of radiotherapy from relevant literature using machine learning and natural language processing (NLP) approaches. A framework is developed in this study that initially builds a training corpus by extracting sentences containing different types of geometric errors of radiotherapy from relevant publications. The publications are retrieved from PubMed following a given set of rules defined by a domain expert. Subsequently, the method develops a training corpus by extracting relevant sentences using a sentence similarity measure. A support vector machine (SVM) classifier is then trained on this training corpus to extract the sentences from new publications which contain relevant geometric errors. To demonstrate the proposed approach, we have used 60 publications containing geometric errors in radiotherapy to automatically extract the sentences stating the mean and standard deviation of different types of errors between planned and executed radiotherapy. The experimental results show that the recall and precision of the proposed framework are, respectively, 97% and 72%. The results clearly show that the framework is able to extract almost all sentences containing required data of geometric errors.

Keywords: information extraction; health informatics; NLP; text mining; machine learning; radiotherapy; geometric error

Citation: Basu, T.; Goldsworthy, S.; Gkoutos, G.V. A Sentence Classification Framework to Identify Geometric Errors in Radiation Therapy from Relevant Literature. *Information* **2021**, *12*, 139. https://doi.org/10.3390/info12040139

Academic Editor: Wajahat Ali Khan

Received: 11 February 2021
Accepted: 21 March 2021
Published: 24 March 2021

1. Introduction

Systematic reviews are a type of review that uses a rigorous and transparent approach to provide an evidence-based answer for a particular clinical question by summarizing relevant literature [1,2]. Systematic reviews are composed by summarizing different data elements of a particular topic, collected from relevant articles following a detailed, comprehensive, and transparent plan to reduce bias [2]. Some examples of such data elements include the population of an intervention, the inclusion criteria for testing the effect of a drug, etc. The experts manually extract these data elements from the relevant literature, following a predefined protocol, and build a systematic review, a process which typically requires a substantial amount of time [1]. The process of data element extraction

to compose a systematic review is labour-intensive and time-consuming when dealing with large quantities of data [1–3]. The main challenge that such efforts face lies in the fact that the required data elements that need to be identified within a particular article lack particular, defined patterns of occurrence and are typically reported within tables or in plain text. Moreover, the data elements may occur in a variety of contexts within an article, rendering their identification extremely difficult.

Within the radiation therapy domain, systematic reviews form extremely useful mechanisms for providing answers to particular radiation therapy questions and tasks—for example, identifying evidence of the effective improvement of geometric discrepancies in the radiation therapy of cancer patients. The geometrical uncertainties are developed from the treatment process of the external beam radiotherapy of tumors [4]. The main sources of uncertainty are tumor delineation inaccuracies of the gross tumor volume, unknown extent of microscopic tumor, organ positional variation within the patient, and setup variations [4]. The deviation between planned and executed radiotherapy indicates geometric error, or discrepancy, even if it is small [4]. Therefore, geometric errors have to be identified and removed for safe radiation therapy. There are many recent studies which address issues of measurement and the reduction in geometrical errors [5–7].

In this work, we report on the development of a framework based on machine learning and natural language processing (NLP) for extracting sentences containing required data elements of geometric discrepancies in radiation therapy from relevant literature. The work was carried out in collaboration with a research radiation therapist at the Somerset NHS Foundation Trust in the UK, who collected the articles containing relevant data elements of geometric errors for experimental analysis. The relevant sentences from a number of articles containing required geometric errors were manually identified by a radiation therapist to evaluate the performance of the proposed framework. We experimentally evaluated the framework and reported on its effectiveness and limitations for the data extraction of geometric errors of radiation therapy from relevant literature. The experimental results delineate that the use of an SVM classifier can extract the sentences containing the required geometric errors with a 97% recall and 72% precision, demonstrating the effectiveness of our approach.

2. Related Works

The purpose of radiation therapy or radiotherapy is to deliver doses of radiation to tumors by minimizing the risk of side effects in healthy tissues. Undeniably, radiotherapy planning and delivery face many uncertainties [8]. Target volume definition, the first step in the treatment planning chain, is associated with substantial uncertainty [8]. The main sources of uncertainty are the tumor delineation inaccuracies of the gross tumor volume, the unknown extent of microscopic tumor, as well as the organ positional variation within a patient and setup variations [4]. The geometric error or discrepancy indicates any deviation between the planned and executed radiotherapy, even if it is small [4]. Hence, it is necessary to have high geometrical accuracy for a safe clinical application of precise radiotherapy. Some recent studies describe the process of measurement and reduction in geometrical errors [5–7]. There are some reviews which discuss different types of errors in radiotherapy and the process to overcome these discrepancies [3,8–10]. Such reviews are extremely valuable and they necessitate research radiation therapists to manually explore the literature to identify case specific geometric errors as well as the corresponding measurements to plan for safe doses. Such tasks though are both resource and time expensive, since the required data elements that need to be extracted do not follow particular or fixed reporting patterns. Machine learning approaches can potentially be used to address these challenges [2,11].

There are a growing number of efforts to identify data elements related to a number of diseases across both the scientific literature as well social media datasets using machine learning and NLP techniques [1,12,13]. Goswami et al. developed a machine learning technique applying a random forest classifier to extract data elements of anxiety outcome measures from relevant literature [11], with potential to assist reviews with large numbers of studies synthesising these measures [3]. RobotReviewer is a web-based system that employs both machine learning and NLP to identify the Risk of Bias (RoB) of how a particular clinical study was performed [14]. Another recent study by Basu et al. describes a machine learning framework to identify relevant data elements of congestive heart failure from literature applying SVM classifier [2]. Several PubMed indexed systematic reviews of congestive heart failure were utilised to generate the training data in this study [2]. Hassanzadeh et al. proposed a framework for quantifying the semantic similarity of clinical evidence in the biomedical literature based on a series of component level generic and domain specific semantic similarity measures [15].

Different workshops on NLP were organised for the de-identification of protected health information from relevant medical records by the Informatics for Integrating Biology and the Bedside (i2b2) research group based at Harvard Medical School [16–20]. Yim et al. developed a sparse annotation method for tumour information extraction and built a conditional random field based system for entity and relation extraction for these characteristics [21]. Recently, Wang et al. published a review article of clinical information extraction applications [22]. They analysed different applications, based on machine learning and NLP techniques, for information extraction from various types of electronic health records [22].

3. Proposed Framework

However, to our knowledge, there is no study that discusses the issue of automatically identifying data elements of geometric errors of radiotherapy from relevant publications. To address this need, a supervised machine learning framework is developed to extract the sentences containing the required geometric errors of radiotherapy from the relevant literature. The framework consists of two major parts, as described below.

3.1. Building Training Corpus

We used 60 articles in PDF format related to geometric errors of radiotherapy to conduct this study. Fitz (https://pypi.org/project/PyMuPDF/1.9.2, accessed on 17 March 2021), a Python module, was used to convert the PDFs to free text. A total of 52 out of 60 documents were randomly selected to build the training corpus, with two classes, *geometric-errors*, and *non-geometric-errors*. In principle, the geometric errors class should have the sentences that contain the required geometric errors of radiotherapy. Certain keywords related to the geometric errors of radiotherapy—e.g., *geometric organ error*—were used to identify whether a sentence belongs to the geometric errors class. This set of keywords was defined by the domain expert and is reported in Table 1. The sentences that do not contain any of these keywords related to geometric errors were used to form the non-geometric errors class. There may exist some sentences in an article that contain some of the required keywords, but do not contain any required data element—i.e., a decimal number. These sentences were discarded, as they were not relevant to either of the classes.

A sentence similarity measure was used to identify the relevant sentences from a given article that represent individual classes. The similarity measure, termed *sent_sim*, was defined in line with the Jaccard similarity measure [23]. The similarity between a keyword (say, *kw*) that represents the geometric error and every sentence (say, *S*) in a given article or part of the training set is defined as:

$$sent\_sim(kw, S) = \frac{|\mathcal{T}(kw) \cap \mathcal{T}(S)|}{|\mathcal{T}(kw)|} \quad (1)$$

Here, $\mathcal{T}(kw)$ and $\mathcal{T}(S)$ denote the set of words in kw and S, respectively. Note that the values of *sent_sim* range between $[0, 1]$, where 1 denotes highest similarity. The aim of *sent_sim* is to identify how many words of kw exist in S, unlike traditional sentence similarity measures, such as Jaccard, which compute the similarity based on the common words of two sentences. Let us assume that kw is a small phrase and S is a large sentence, but many words of kw exist in S. In that case, the *sent_sim* score of kw and S will be high, indicating the sentence S is relevant to kw.

Table 1. Relevant Keywords of Geometric Errors in Radiation Therapy.

systematic displacement error	random displacement error
systematic random displacement error	rotational random systematic error
translational random systematic setup error	translational error
x y z direction translational	x y z direction rotational
x y z correction translational	x y z correction rotational
translation discrepancies	translational discrepancies
rotational discrepancy	rotational discrepancies
rotation translation error	rotation translation discrepancy
rotation translation discrepancies	rotation translation displacement
mean set up error	geometric organ error
standard deviation of set up error	population systematic error
population random error	organ motion translation
organ motion rotation	set up error translation
set up error rotation	translational correction
rotational correction	vector correction
residual error	total error mean or SD
total error mean or standard deviation	rotational systematic error
translational systematic error	rotational random error
translational random error	systematic and random population error
anterior posterior inferior superior	translation discrepancy

The sentences with *sent_sim* scores greater than or equal to a prefixed threshold α and containing a decimal number were extracted from a given article to construct the geometric-errors class. The value of α was fixed experimentally as described in Section 5. Sentences with a *sent_sim* score of 0, for all the given keywords, were used to form the non-geometric-errors class. The remaining sentences of the document were discarded. Algorithm 1 describes the detailed steps of the training corpus generation.

Algorithm 1 Generation of Training Data

Input : (1) A number of free text documents
(2) *Keywords* ← {geometric organ error, random displacement error etc.}
(3) α ← A threshold on *sent_sim*
Steps:

```
for each document do
  get sentences from document following regular delimiters
  for each sentence in a document do
    flag ← 0
    convert each character to lower case
    for each term in Keywords do
      score ← sent_sim(term, sentence)
      if score > α and sentence contain a decimal number then
        geometric_errors_class ← sentence
        goto step 3
      else if score ≠ 0 then
        flag ← flag + 1
      end if
    end for
    if flag = 0 then
      non_geometric_errors_class ← sentence
    end if
  end for
end for
return geometric_errors_class, non_geometric_errors_class
```

3.2. Extraction of Desired Data Elements

A machine learning framework was developed in the second stage, where the training corpus was used to train a classifier to determine whether a sentence from a test article contains any geometric error. The bag of words model was then applied for generating features from free text. Unigrams, bigrams, and trigrams generated from sentences were used as features with the SVM classifier in the experimental analysis. A unigram considers all unique words in a sentence as features [24]. A bigram or trigram, on the other hand, considers only two or three consecutive words as a feature, respectively [24]. Both bigrams and trigrams were used in this framework, since there were many terms in the training corpus—e.g., rotational discrepancy, random displacement error—which should be conjoined for analysis.

The conventional vector space model was used to represent the vector corresponding to each sentence [24,25], which is widely used by several text classification and clustering techniques [26]. Let us consider the number of sentences in the corpus as n and the number of unique terms—i.e., the number of unigrams, bigrams, and trigrams—as m. Let us also consider that t_i denotes the ith term and the frequency of t_i in the jth sentence is denoted by tf_{ij}, $i = 1, 2, \cdots, m$; $j = 1, 2, \cdots, n$. The entropy-based term weighting technique is used by many researchers to form a term-document matrix from free text data [27,28]. This method reflects the assumption that the more important term is the more frequent one that occurs in fewer documents, taking the distribution of the term over the corpus into account [28]. Thus, the weight of a term t_i in the jth sentence, denoted by W_{ij}, is determined by the entropy-based technique (https://radimrehurek.com/gensim/models/logentropy_model.html, accessed on 17 March 2021) [28] as follows:

$$W_{ij} = \log(tf_{ij} + 1) \times \left(1 + \frac{\sum_{j=1}^{n} P_{ij} \log P_{ij}}{\log(n+1)}\right), \quad \text{where,} \quad P_{ij} = \frac{tf_{ij}}{\sum_{j=1}^{n} tf_{ij}}$$

Let us assume, $\vec{S_j}$ is the vector of a sentence, say S_j, where the ith component of the vector is W_{ij}—i.e., $\vec{S_j} = [W_{1j}, W_{2j}, \cdots, W_{mj}]$, $\forall j = 1, 2, \cdots, n$. The cosine similarity is a commonly used measure to find similarity between documents [24–26]. Thus, the similarity between two sentences—say, S_j and S_k—can be defined as:

$$\cos(\vec{S_j}, \vec{S_k}) = \frac{\vec{S_j}.\vec{S_k}}{|\vec{S_j}|\,|\vec{S_k}|} = \frac{\sum_{i=1}^{m}(W_{ij} \times W_{ik})}{\sqrt{\sum_{i=1}^{m} W_{ij}^2 \times \sum_{i=1}^{m} W_{ik}^2}}, \quad \forall j, k = 1, 2, \cdots, n$$

Note that cosine similarity is non-negative and ranges between 0 and 1, both inclusive. $cos(\vec{S_j}, \vec{S_k}) = 1$ indicates that the sentences are exactly similar and the similarity decreases as the value comes nearer to 0.

The SVM classifier is used to classify the sentences of the test documents using the training corpus. Given a set of training documents in a vector space, SVM finds the best decision hyperplane that separates individual documents belonging to two different classes. An SVM classifier extends its applicability on the linearly non-separable data sets either by using soft margin hyperplanes or by mapping the original data vectors to a higher dimensional space in which the vectors are linearly separable. The linear kernel is recommended when a data set has large number of features [29], since it has been reported that mapping the data to a higher dimensional space using a non-linear kernel does not result in substantial performance improvement [29]. Since free text data is high-dimensional, an SVM classifier with linear kernel that improves the performance of text classification [29,30] is used in the experimental analysis.

4. Experimental Evaluation

4.1. Experimental Setup

The performance of logistic regression, random forest, and SVM classifiers was tested to classify the sentences of the test documents. The training set was used to tune the parameters of these classifiers, applying 10-fold cross validation technique. The sentences of the test documents were then classified using the best set of parameters of each classifier. The sentences were either classified to the geometric errors class or to the non-geometric-errors class. The code and data set that were used to implement the proposed framework are available at Github (https://github.com/tanmaybasu/A-Sentence-Classification-Framework, accessed on 17 March 2021).

4.2. Evaluation Measures

The performances of the individual classifiers were evaluated using precision, recall, and f-measure [25]. The precision and recall can be defined as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

Here, true positive represents the number of sentences correctly predicted as belonging to the geometric-errors class. False positive represents the number of sentences that are predicted as geometric errors but are members of the non-geometric-errors class. False negative represents the number of sentences that, while predicted as non-geometric-errors, are members of the geometric-errors class. The f-measure can be defined in the following way:

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

F-measure will be high when the values of precision and recall are close to each other [31]. The value of f-measure is 1 when the values of precision and recall are 1 and becomes 0 when the precision is 0, recall is 0, or both are 0. Thus, the value of the f-measure ranges in between 0 and 1. A high value for f-measure indicates the good performance of a classifier.

4.3. Analysis and Results

The training and test corpora contained 9545 and 4336 sentences, respectively. The training corpus included 324 sentences belonging to the geometric errors class and 9221 sentences that were members of the non-geometric errors class. The rest of the sentences of the training corpus were discarded, as they were not related to either of the classes. Table 2 shows the performance of the proposed framework using the SVM classifier to classify the sentences of the eight test documents.

Note that the objective of this framework is to achieve a high accuracy in terms of identifying sentences containing measurements of different types of geometric errors from the test documents. Therefore, a high recall is desirable. The sentences of the test documents containing the required data elements were manually identified by a domain expert in radiation therapy to evaluate the performance of the framework.

Thus the recall and precision scores for each test document were computed using the sentences manually identified by the domain expert and the sentences extracted by the proposed framework. The true positive, false positive, false negative, and recall and precision scores of each of the eight test documents are presented in Table 2. Almost all the test documents have zero false negatives, leading to a very good recall score, indicating that the proposed system is able to retrieve relevant information from these documents. Table 3 shows that the aggregate recall of the framework using SVM classifier for the eight test documents is 0.97, while the aggregate precision score is 0.72. A low precision score is still efficient, since potential reviewers would now need to review only $1/0.72 = 1.38$ sentences per document to identify the geometric errors as opposed to reading the entire document, which, on average, contain around 200 sentences.

Table 2. Performance of the proposed framework using SVM classifier.

Test Document	True Positive	False Positive	False Negative	Precision	Recall
Doc 1	12	4	1	0.75	0.92
Doc 2	29	4	1	0.87	0.96
Doc 3	12	13	0	0.48	1
Doc 4	3	1	0	0.75	1
Doc 5	12	3	1	0.8	0.92
Doc 6	6	1	0	0.85	1
Doc 7	7	4	0	0.63	1
Doc 8	6	3	0	0.66	1

Table 3. Performance of the proposed framework using different classifiers.

Classifier	Precision *	Recall *	F-Measure *
Logistic Regression	0.69	0.95	0.80
Random Forest	0.66	0.91	0.77
Support Vector Machine	0.72	0.97	0.83

* Aggregate score of 8 test documents.

5. Discussion

We developed a framework based on bag of words model and SVM classifier that partially automates the process of building systematic radiotherapy literature reviews by extracting relevant sentences from the literature. Logistic regression and random forest classifiers also perform well for text classification [11,32,33]. Hence, the performance of the proposed framework is assessed using these classifiers. The aggregate precision and recall scores of the proposed framework using logistic regression, random forest, and SVM classifiers for the eight test documents are reported in Table 3. The f-measure scores, reported in Table 3, were computed from the aggregate precision and recall scores of the individual classifiers. It can be seen from Table 3 that the SVM classifier obtained the best performance in terms of precision, recall, and f-measure.

To our knowledge, the proposed framework is the first of its kind that can automatically extract geometric errors from relevant publications to expedite the process of systematic literature review. Goswami et al. developed a similar method based on term frequency and inverse document frequency (TF-IDF)-based term weighting scheme [24] to extract anxiety outcome measures for comfort intervention from relevant literature [11]. This approach used different articles collected from Medline, EMBASE, CINAHL, and AHMED related to anxiety outcome measures to build the training corpus [11]. However, a set of keywords, defined by the domain experts, was utilised to assess whether any of these keywords occurred in a sentence so as to generate a training corpus, unlike our method that employs a sentence-matching technique. Furthermore, the keywords that were identified by the domain experts for this study [11] are fairly simple, whereas the keywords used in our approach were much more complex. Let us consider the following sentence from test document 3 [34].

Fuss et al. (5) reported that the translational error at the isocenter was 0.74 ± 0.53, 0.75 ± 0.60, and 0.93 ± 0.78 mm in the RL, CC, and AP directions, respectively.

It may be noted that *'translational error'* is one of the keywords in the proposed study and this sentence is clearly describing a geometric error. However, this sentence is indicating another author's work cited in this paper [34] and hence it is treated as false positive by the domain expert. There are many such sentences in different test documents. As the proposed framework is based on bag of words model and was trained on the sentences that contain the keywords, any similar sentence will be extracted as relevant. Thus, the number of false positives is high for some test documents, which results in a low aggregate precision score.

In principle, a high value of the sentence similarity threshold α is desirable in Algorithm 1 so as to avoid a performance degradation. On the other hand, a very high value of α (e.g., $\alpha = 0.95$) may result in the inclusion of very few sentences in the geometric_errors class of the training set. In order to assess the necessary trade-off between these two, the value of α was experimentally determined and different training corpora were generated using different α values. Subsequently, the SVM classifier was performed to classify the sentences of these training corpora following 10-fold cross validation technique. Eventually, the training set for a particular α value, with the highest f-measure, was used to classify the test documents. Thus the value of α was fixed to report the results in Tables 2 and 3.

The performance of the proposed framework was also tested using the conventional TF-IDF based term weighting scheme of the vector space model for text document representation [11,24] instead of entropy-based technique. Additionally, the simple keyword matching technique using *sent_sim* similarity measure to build the training corpus is used to extract relevant sentences from the test documents and the performance is reported in Table 4. The performance of the SVM classifier applying both the entropy-based feature weighting scheme and the TF-IDF-based feature weighting scheme is also reported in Table 4. Moreover, the performance of BioBERT [35], which is a pre-trained language representation model for the biomedical domain is reported in Table 4. BERT (Bidirectional Encoder Representations from Transformers) is a contextualised word representation model that is based on a masked language model and pretrained using bidirectional trans-

formers [36]. This deep learning architecture has been widely used in many NLP tasks over the last few years [35]. BERT was pretrained on general domain corpora—i.e., English Wikipedia and books [36]. BioBERT was initialised with weights from BERT and was pretrained on full text articles and abstracts from PubMed [35]. BioBERT performed very well for certain NLP tasks—e.g., sentence classification for relation extraction [35]. We used the BioBERT pretrained model (https://github.com/naver/biobert-pretrained, accessed on 17 March 2021) and then fine-tuned it on our training corpus. Subsequently, the sentences of the test documents were classified using this pretrained model and the sentence classification framework of BioBERT. It can be seen from Table 4 that the proposed framework—i.e., the entropy-based feature weighting scheme and SVM classifier—performs better than BioBERT and other techniques in terms of aggregate precision, recall, and f-measure scores of eight test documents. It is observed from Table 4 that BioBERT did not perform well on the test documents. We checked the vocabulary built by the pretrained BioBERT model on the PubMED corpus and noticed that it does not contain some useful words from the given keywords in Table 1, which appeared in many documents of the training and test corpora. Hence, it could not capture the semantic interpretation of these keywords from the given texts.

Table 4. Performance of different sentence extraction techniques.

Classifier	Precision *	Recall *	F-Measure *
Keyword match technique based on *sent_sim*	0.76	0.65	0.70
BioBERT pretrained model for sentence classification	0.63	0.84	0.72
TF-IDF based feature weighting scheme + SVM	0.69	0.92	0.79
Entropy based feature weighting scheme + SVM	**0.72**	**0.97**	**0.83**

* Aggregate score of 8 test documents.

The proposed framework has some limitations, although it has performed well empirically. The method extracts required geometric errors from relevant documents, but it cannot make any judgment on the extracted data. This framework works on free text documents and it can not read and extract data from figures or charts. Furthermore, the proposed framework is based on bag of words model and cannot therefore apply any semantic interpretation of the text extracts. A deep learning based document or word embeddings could potentially be employed to generate such semantic features from the documents. In this particular case, however, such deep learning approaches, since they require a large number of documents for training, may not work well, as the size of the corpus used in this work is very small.

6. Conclusions

A machine learning and NLP-based framework is proposed in this study to automatically build a training corpus followed by a sentence classification framework to extract required geometric errors of radiotherapy from relevant literature. The sentence classification framework was developed based on bag of words model for text feature generation, followed by an entropy-based feature weighting scheme and SVM classifier. Although the SVM classifier extracted almost all the relevant sentences containing the measurement of different geometric errors, it extracted some false positive sentences as well from the test documents. In future, we plan to build a deep learning-based embedding by using a substantial number of relevant articles of geometric errors in radiotherapy over PubMED, Scopus, Wikipedia, and other relevant resources to properly derive the semantic interpretation of the contextual information. We also plan to include a direct feed into a systematic review paper and inferences over the extracted data that would be useful for clinical researchers. Finally, we plan to generalise our approach and assess its effectiveness for other diseases and clinical settings.

Author Contributions: T.B. contributed to the conception and design of the proposed framework and to conduct the experimental analysis. S.G. collected the data and had done the manual annotations of the test documents to evaluate the performance of the framework. T.B., S.G. and G.V.G. took part in writing the manuscript and accountable for the manuscript's contents. All authors have read and agreed to the published version of the manuscript.

Funding: This work was directly supported by the MRC Heath Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health. Georgios V. Gkoutos also acknowledges support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032) and the NIHR Birmingham Biomedical Research Centre.

Data Availability Statement: The code and data set that were used to implement the proposed framework are available at Github (https://github.com/tanmaybasu/A-Sentence-Classification-Framework, accessed on 17 March 2021).

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

References

1. Jonnalagadda, S.R.; Goyal, P.; Huffman, M.D. Automating data extraction in systematic reviews: A systematic review. *Syst. Rev.* **2015**, *4*, 78. [CrossRef]
2. Basu, T.; Kumar, S.; Kalyan, A.; Jayaswal, P.; Goyal, P.; Pettifer, S.; Jonnalagadda, S.R. A Novel Framework to Expedite Systematic Reviews by Automatically Building Information Extraction Training Corpora. *arXiv* **2016**, arXiv:1606.06424.
3. Goldsworthy, S.; Palmer, S.; Latour, J.; McNair, H.; Cramp, M. A systematic review of effectiveness of interventions applicable to radiotherapy that are administered to improve patient comfort, increase patient compliance, and reduce patient distress or anxiety. *Radiography* **2020**. [CrossRef] [PubMed]
4. Van Herk, M. Errors and margins in radiotherapy. *Semin. Radiat. Oncol.* **2004**, *14*, 52–64. [CrossRef]
5. Goldsworthy, S.; Leslie-Dakers, M.; Higgins, S.; Barnes, T.; Jankowska, P.; Dogramadzi, S.; Latour, J.M. A pilot study evaluating the effectiveness of dual-registration image-guided radiotherapy in patients with oropharyngeal cancer. *J. Med. Imaging Radiat. Sci.* **2017**, *48*, 377–384. [CrossRef]
6. Sarkar, B.; Munshi, A.; Ganesh, T.; Manikandan, A.; Krishnankutty, S.; Chitral, L.; Pradhan, A.; Kalyan Mohanti, B. Rotational positional error corrected intrafraction set-up margins in stereotactic radiotherapy: A spatial assessment for coplanar and noncoplanar geometry. *Med. Phys.* **2019**, *46*, 4749–4754. [CrossRef]
7. Caillet, V.; Zwan, B.; Briggs, A.; Hardcastle, N.; Szymura, K.; Podreka, A.; O'Brien, T.R.; Harris, B.; Greer, P.B.; Haddad, C.; et al. Geometric uncertainty analysis of MLC tracking for lung SABR. *Phys. Med. Biol.* **2020**, *65*, 235040. [CrossRef] [PubMed]
8. Unkelbach, J.; Alber, M.; Bangert, M.; Bokrantz, R.; Chan, T.C.; Deasy, J.O.; Fredriksson, A.; Gorissen, B.L.; Van Herk, M.; Liu, W.; et al. Robust radiotherapy planning. *Phys. Med. Biol.* **2018**, *63*, 22TR02. [CrossRef]
9. Fraass, B.A. Errors in radiotherapy: motivation for development of new radiotherapy quality assurance paradigms. *Int. J. Radiat. Oncol. Biol. Phys.* **2008**, *71*, S162–S165. [CrossRef] [PubMed]
10. Mišić, V.V.; Chan, T.C. The perils of adapting to dose errors in radiation therapy. *PLoS ONE* **2015**, *10*, e0125335. [CrossRef]
11. Goswami, S.; Pal, S.; Goldsworthy, S.; Basu, T. An effective machine learning framework for data elements extraction from the literature of anxiety outcome measures to build systematic review. In Proceedings of the International Conference on Business Information Systems, Seville, Spain, 26–28 June 2019; pp. 247–258.
12. Guntuku, S.C.; Yaden, D.B.; Kern, M.L.; Ungar, L.H.; Eichstaedt, J.C. Detecting depression and mental illness on social media: An integrative review. *Curr. Opin. Behav. Sci.* **2017**, *18*, 43–49. [CrossRef]
13. Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.S.; Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia 19–25 August 2017; pp. 3838–3844.
14. Marshall, I.J.; Kuiper, J.; Banner, E.; Wallace, B.C. Automating biomedical evidence synthesis: RobotReviewer. In Proceedings of the Conference Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 2017, p. 7.
15. Hassanzadeh, H.; Nguyen, A.; Verspoor, K. Quantifying semantic similarity of clinical evidence in the biomedical literature to facilitate related evidence synthesis. *J. Biomed. Inform.* **2019**, *100*, 103321. [CrossRef] [PubMed]
16. Uzuner, Ö.; Luo, Y.; Szolovits, P. Evaluating the state-of-the-art in automatic de-identification. *J. Am. Med. Inform. Assoc.* **2007**, *14*, 550–563. [CrossRef] [PubMed]
17. Uzuner, Ö.; Solti, I.; Cadag, E. Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 514–518. [CrossRef]
18. Halgrim, S.R.; Xia, F.; Solti, I.; Cadag, E.; Uzuner, Ö. A cascade of classifiers for extracting medication information from discharge summaries. *J. Biomed. Semant. Biomed. Cent.* **2011**, *2*, S2. [CrossRef]

19. Stubbs, A.; Kotfila, C.; Uzuner, Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. *J. Biomed. Inform.* **2015**, *58*, S11–S19. [CrossRef] [PubMed]
20. Stubbs, A.; Filannino, M.; Uzuner, Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. *J. Biomed. Inform.* **2017**, *75*, S4–S18. [CrossRef]
21. Yim, W.W.; Denman, T.; Kwan, S.W.; Yetisgen, M. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. In Proceedings of the AMIA Summits on Translational Science Proceedings, San Francisco, CA, USA, 21–24 March 2016; pp. 455–484.
22. Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. Clinical information extraction applications: a literature review. *J. Biomed. Inform.* **2018**, *77*, 34–49. [CrossRef]
23. Lee, L. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the ACL, College Park, MD, USA, 20–26 June 1999; pp. 25–32.
24. Manning, C.D.; Raghavan, P.; Schutze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
25. Basu, T.; Murthy, C. A supervised term selection technique for effective text categorization. *Int. J. Mach. Learn. Cybern.* **2016**, *7*, 877–892. [CrossRef]
26. Mukherjee, A.; Basu, T. A medoid-based weighting scheme for nearest-neighbor decision rule toward effective text categorization. *SN Appl. Sci.* **2020**, *2*, 1–9. [CrossRef]
27. Selamat, A.; Omatu, S. Web page feature selection and classification using neural networks. *Inf. Sci.* **2004**, *158*, 69–88. [CrossRef]
28. Sabbah, T.; Selamat, A.; Selamat, M.H.; Al-Anzi, F.S.; Viedma, E.H.; Krejcar, O.; Fujita, H. Modified frequency-based term weighting schemes for text classification. *Appl. Soft Comput.* **2017**, *58*, 193–206. [CrossRef]
29. Hsu, C.W.; Chang, C.C.; Lin, C.J. A Practical Guide to Support Vector Classification; 2010. Available online: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed on 17 March 2021).
30. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning, Chemnitz, Germany, 21–23 April 1998; pp. 137–142.
31. Basu, T.; Murthy, C.A. A Feature Selection Method for Improved Document Classification. In Proceedings of the International Conference on Advanced Data Mining and Applications, Nanjing, China, 15–18 Decmber 2012; pp. 296–305.
32. Paul, S.; Jandhyala, S.K.; Basu, T. Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks. In Proceedings of the CLEF (Working Notes), Avignon, France, 10–14 September 2018.
33. Hasan, M.; Rundensteiner, E.; Agu, E. Automatic emotion detection in text streams by analyzing twitter data. *Int. J. Data Sci. Anal.* **2019**, *7*, 35–51. [CrossRef]
34. Oita, M.; Ohmori, K.; Obinata, K.; Kinoshita, R.; Onimaru, R.; Tsuchiya, K.; Suzuki, K.; Nishioka, T.; Ohsaka, H.; Fujita, K.; et al. Uncertainty in treatment of head-and-neck tumors by use of intraoral mouthpiece and embedded fiducials. *Int. J. Radiat. Oncol. Biol. Phys.* **2006**, *64*, 1581–1588. [CrossRef] [PubMed]
35. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.