Natural language processing applied to incident reports in radiation oncology–determination of completeness and classification

Hui Wang[1] and John Kildea[1]

[1]Affiliation not available

January 9, 2018

Abstract

Purpose: Incident learning in radiotherapy may improve the quality of care by reducing the recurrence of incidents. Incident reports are often descriptions in free-text format, which makes it difficult to extract information efficiently and determine if a report is complete. Natural language processing (NLP) is a proven technique in extracting information from clinical texts and classifying safety reports, but the utility of NLP in classifying radiotherapy incident reports is untested. This project illustrates an NLP approach to classifying radiotherapy incident reports by the process step of incident occurrence, and determining if a report is complete.

Methods: A preliminary analysis that consisted of tagging of parts-of-speech (POS) and annotation with the Unified Medical Language System (UMLS) was conducted on 519 reports. The summary statistics of the preliminary analysis was used to compare fictitious reports deemed complete and reports of real incidents. Multinomial classification by the process step of incident occurrence was transformed into multiple binomial classifications by one-hot encoding. The reports were subsequently cleaned and vectorized into a bag of words, on which the naive binomial bayesian classifiers were trained and tested. Receiver-operating-characteristic (ROC) curves were then plotted to evaluate each binomial classifier.

Results: The number of occurrence of several POS were statistically different in the fictitious and real reports, and the performance of the binary classifiers varied from poor to excellent with an average area under the curve (AUC) of 0.85.

Conclusion: Determining if a report is complete requires domain knowledge outside the scope of statistical NLP. However, NLP can classify incidents by process step with fair performance, and demonstrates the potential to classify incidents by other data elements of the Canadian National System for Incident Reporting - Radiation Treatment (NSIR-RT).

Introduction

Radiotherapy is the use of ionizing radiation to treat cancer, and is used in almost two thirds of cancer treatments (Kaur et al., 2011; Montgomery, 2016). The procedure is multi-step and highly complex, and involves many different staff groups. The complexity of the procedure arises from the wide range of cancers treated, technologies used, and expertise required. Although errors only occur approximately once per 600 treatments (Ford et al., 2012; Mans et al., 2010), the consequence to the patient can be significant.

Incident learning consists of identifying, reporting, and investigating actual and potential incidents, and installing protocols to prevent the recurrence of incidents. Since radiotherapy incidents may cause irreversible and severe damage to the patient, incident learning must include precursors and near misses rather than consisting exclusively of trial-and-error at great cost to the patients (Milosevic 2016a). The advantages of incident learning from both actual and potential incidents is paralleled in many industrial environments where errors are of catastrophic consequence. Figure 1 (Jehring and Heinrich, 1951) illustrates the fact that in industrial environments, for every fatal incident, there are 30 minor events and 300 near misses (Jehring and Heinrich, 1951). Root cause analysis of the minor events and near misses has been shown to yield valuable lessons that prevent serious incidents (Cooke et al., 2007). For example, incident learning has reduced the risk of fatal accidents by 73% in 10 years, and is a basic feature of nuclear operations (Ford et al., 2012).

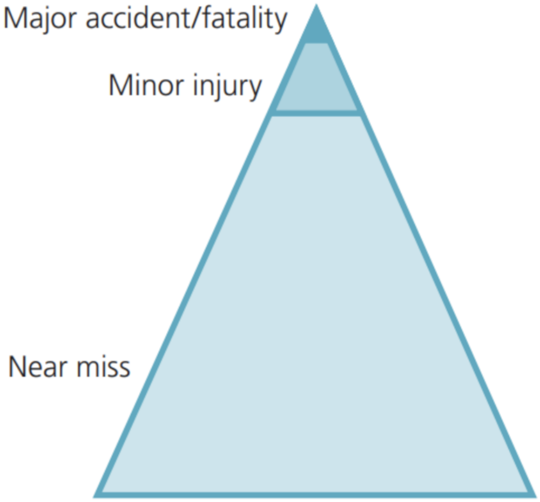
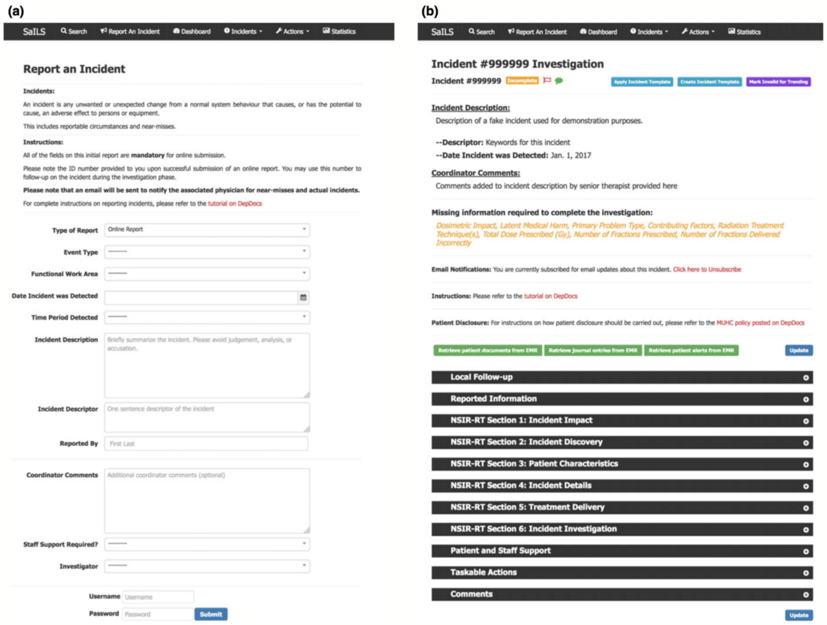


Figure 1: An illustration of the relative occurrence of major accidents, minor injuries, and near misses in industrial environments.

The benefits of recursive self-improvement by incident learning has inspired collaborative incident learning initiatives within institutions as well as across national and international communities (Donaldson 2007). In Canada, the National System for Incident Reporting in Radiation Treatment (NSIR-RT) taxonomy is implemented as an online reporting and analysis system. The taxonomy comprises 33 data elements in 6 categories: impact, discovery, patient, details, treatment delivery, and investigation. The taxonomy has generated a high level of agreement within the Canadian community, and is aligned with the American and European radiation treatment incident classifications (Milosevic et al., 2016).

In January 2016, the Department of Radiation Oncology of Cedar Cancer Centre of the McGill University Health Centre (MUHC) in Montreal, Canada deployed the Safety and Incident Learning System (SaILS) that includes incident reporting, investigations, tracking, and data visualization. The MUHC version of SaILS is compatible with NSIR-RT taxonomy, and has processed 519 incidents from January 2016 to September 2017. Reporting an incident requires mainly a description and a one-sentence descriptor of the incident. For each incident, investigators then retrospectively and manually classified the incidents into one of the menu choices for each data element in the NSIR-RT taxonomy (Montgomery et al., 2017). For example, the process step of incident occurrence is one of the data elements of the NSIR-RT taxonomy, and it has the menu choices: patient assessment and consultation, imaging for radiotherapy planning, treatment planning, pre-treatment review and verification, treatment delivery, on-treatment quality management, and other. The reporting interface of SaILS is shown in Figure 2.

The SaILS incident reports are similar to the safety reports of various industrial environments in that they are in an unstructured free-text format, which makes it difficult to determine if a report has described the incident adequately. The lack of detail may present challenges to subsequent investigations, diminishing the benefits of incident learning. Another inherent challenge posed by incident learning from free-text reports is the quantity of incident reports may exceed investigatory resources. For example, aviation and railway companies must process hundreds of incident reports each month (Peter Hughes, 2016; Tanguy et al., 2016).



Report an Incident

Incidents:

An incident is any unwanted or unexpected change from a normal system behaviour that causes, or has the potential to cause, an adverse effect to persons or equipment.

This includes reportable circumstances and near-misses.

Instructions:

All of the fields on this initial report are **mandatory** for online submission.

Please note the ID number provided to you upon successful submission of an online report. You may use this number to follow-up on the incident during the investigation phase.

Please note that an email will be sent to notify the associated physician for near-misses and actual incidents.

For complete instructions on reporting incidents, please refer to the tutorial on DepDocs

Type of Report	Online Report
Event Type	----------
Functional Work Area	----------
Date Incident was Detected	
Time Period Detected	----------
Incident Description	Briefly summarize the incident. Please avoid judgement, analysis, or accusation.
Incident Descriptor	One sentence descriptor of the incident
Reported By	First Last
Coordinator Comments	Additional coordinator comments (optional)
Staff Support Required?	----------
Investigator	----------
Username	Username
Password	Password

Incident #999999 Investigation

Incident #999999

Incident Description:

Description of a fake incident used for demonstration purposes.

--Descriptor: Keywords for this incident

--Date Incident was Detected: Jan. 1, 2017

Coordinator Comments:

Comments added to incident description by senior therapist provided here

Missing information required to complete the investigation:

Dosimetric Impact, Latent Medical Harm, Primary Problem Type, Contributing Factors, Radiation Treatment Technique(s), Total Dose Prescribed (Gy), Number of Fractions Prescribed, Number of Fractions Delivered Incorrectly

Email Notifications: You are currently subscribed for email updates about this incident. Click here to Unsubscribe

Instructions: Please refer to the tutorial on DepDocs

Patient Disclosure: For instructions on how patient disclosure should be carried out, please refer to the MUHC policy posted on DepDocs

- Local Follow-up
- Reported Information
- NSIR-RT Section 1: Incident Impact
- NSIR-RT Section 2: Incident Discovery
- NSIR-RT Section 3: Patient Characteristics
- NSIR-RT Section 4: Incident Details
- NSIR-RT Section 5: Treatment Delivery
- NSIR-RT Section 6: Incident Investigation
- Patient and Staff Support
- Taskable Actions
- Comments

Figure 2: The SaILS interfaces for incident reporting and investigation.

Natural language processing (NLP) is an emerging technique for analyzing large quantities of texts from a global point of view (Nerbonne, 2012). For example, using NLP to classify aviation safety reports have permitted investigators to reserve their time for the most unique or concerning incidents (Tanguy et al., 2016). Furthermore, NLP has also been effective in extracting information pertaining to disease and disease progression, and even classifying tumour status from radiology reports (Cheng et al., 2009). Nonetheless, the utility of using NLP in classifying radiotherapy incident reports and determining if these reports are complete is as yet untested.

Objective

The objective of this study is to assess if NLP can determine the completeness of a radiotherapy incident report, and automate the classification of incident reports by the process step of incident occurrence, which is one of the data elements of NSIR-RT taxonomy.

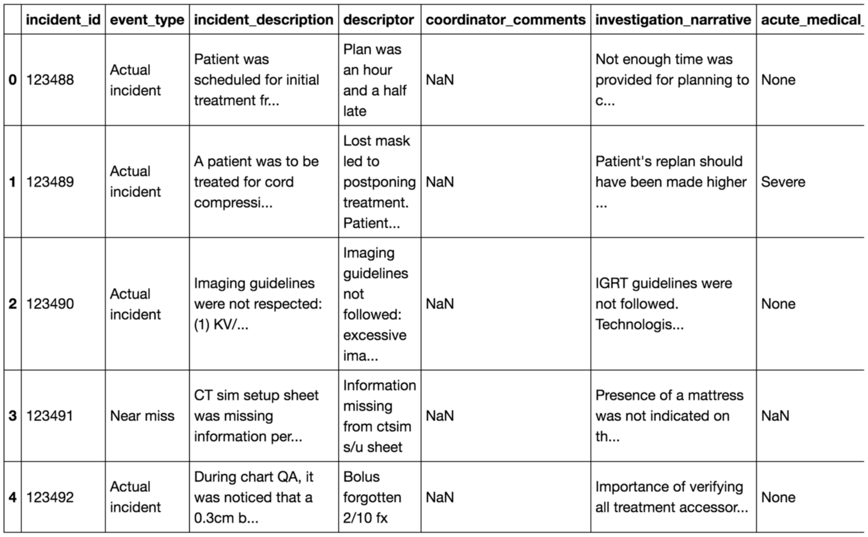
Materials and Methods

Preliminary Analysis

The preliminary analysis was conducted to reveal trends in the linguistic data of the reports. First, each word in each report were tagged with its part-of-speech. Then, each medical terminal was tagged with its semantic type in the Unified Medical Language System ontology. The number of times each part-of-speech and each semantic type appears in a report was counted. From the counts, summary statistics that included the mean, standard deviation, minimum, maximum, and the 25th, 50%, and 75% percentiles, were calculated.

Data Structure

The pandas library (McKinney, 2011) is the most popular data analysis tool in Python. Its data structure, the Dataframe, is similar to an Excel sheet. Throughout this project, the Dataframe is used to store the results, including those of basic linguistic analyses like word count and sentence count, and the occurrence of parts-of-speech and medical terminologies in each incident description and descriptor. An illustration of the Dataframe is illustrated in Figure 3, and the schematic overview of the preliminary analysis is included in that of completeness analysis, illustrated in Fig 4.



	incident_id	event_type	incident_description	descriptor	coordinator_comments	investigation_narrative	acute_medical_
0	123488	Actual incident	Patient was scheduled for initial treatment fr...	Plan was an hour and a half late	NaN	Not enough time was provided for planning to c...	None
1	123489	Actual incident	A patient was to be treated for cord compressi...	Lost mask led to postponing treatment. Patient...	NaN	Patient's replan should have been made higher ...	Severe
2	123490	Actual incident	Imaging guidelines were not respected: (1) KV/...	Imaging guidelines not followed: excessive ima...	NaN	IGRT guidelines were not followed. Technologis...	None
3	123491	Near miss	CT sim setup sheet was missing information per...	Information missing from ctsim s/u sheet	NaN	Presence of a mattress was not indicated on th...	NaN
4	123492	Actual incident	During chart QA, it was noticed that a 0.3cm b...	Bolus forgotten 2/10 fx	NaN	Importance of verifying all treatment accessor...	None

Figure 3: A Pandas Dataframe containing 5 fictitious radiotherapy incident reports

POS Tagging

A part-of-speech (POS) Tagger assigns a POS, such as noun or verb, to each word. POS Tagging in the preliminary analysis is completed with the Natural Language Toolkit (NLTK) (Loper and Bird, 2002), a Python library that offers a comprehensive set of 35 fine grained POS tags, such as singular proper noun and plural proper noun.

NLTK methods are only able to be used on document objects referred to as corpuses, which must be created from plain text files. As a result, the description and descriptor of each incident were converted into text files, and then into corpuses through a PlainTextCorpusReader constructor. Then, for each POS tag in NLTK's tag set, a new column was added to the Dataframe, and the number of occurrence of the POS in the corpus was stored in the column.

UMLS Annotation

The Unified Medical Language System (UMLS) is a comprehensive thesaurus and ontology of medical concepts (Bodenreider, 2004). UMLS has enabled the development of NLP tools that behave as they understand the language of biomedicine and health (Bethesda, 2009). For example, Apache cTakes (Savova et al., 2010), used in the preliminary analysis, is a java program able to annotate medical terminologies recognized by the UMLS. In the preliminary analysis, cTAKES tagged each medical terminology with one of the following UMLS semantic types: diseases/disorders, signs/symptoms, anatomical sites, procedures, or medications.

Since Apache cTAKES is a standalone Java program that offers access from the command line, the descriptions and the descriptors of the incidents in text files were batch processed by a bash script. cTAKES outputs a metadata file, so getter methods were programmed to extract the semantic type annotations from the metadata. Then, for each semantic type, a new column was added to the dataset, and the number of occurrence of each semantic type in the text file was stored into the column.

Summary Statistics

In addition to the number of occurrence of each POS, the fractional POS compositions were calculated by dividing the number of occurrence by the word count. Summary statistics that describes the central tendencies, dispersion, and the shape of a dataset's distribution were then generated for each column that represents the occurrence of POS and semantic type in the Dataframe. Specifically, for each Dataframe column, the mean, standard deviation, minimum, maximum, and the 25th, 50%, and 75% percentiles, were calculated to reveal overall data trends.

Completeness Analysis

20 fictitious incident reports were generated as the "gold standard" of incident reports during the creation of NSIR-RT (Liszewski and Davis, 2015). During the completeness analysis, these 20 fictitious reports were regarded as supreme examples of completeness, and were then subjected to the preliminary analysis described above. The datasets of real and fictitious reports, including the summary statistics of the preliminary analysis, were then concatenated. A column for a new binary variable that indicates if an incident is fictitious or not is added alongside the columns in the summary statistics. Then, the Pearson correlation coefficients were calculated between the "fictitiousness" column and each column in the summary statistics. For each variable that achieved a Pearson correlation coefficient less than 0.05, a comparative histogram of the variable overlaying the real and fictitious data was plotted. The variables with the most different distributions between the real and fictitious incidents were then considered as classifiers for completeness. The Pearson correlations were calculated with the open source Python library, SciPy version 0.19.1 (pyt). The schematic overview of the completeness analysis in illustrated in Figure 4.

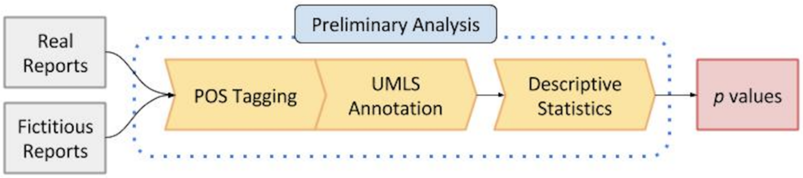


Figure 4: Schematic overview of the completeness analysis. Data objects are represented by rectangular components, and data manipulations by yellow chevrons. Abbreviations: POS: part-of-speech; UMLS: Unified Medical Language System

Incident Classification

Incident classification consisted of tokenizing the reports, cleaning the data, machine learning, and evaluation. Even though there was some overlap between the preliminary analysis and the data preparation steps of tokenizing and data cleaning, completing these steps during the preliminary analysis would interfere with the POS tagging and UMLS annotation. Thus, these steps are detailed in separate sections under incident classification separately from the preliminary analysis. The interferences are also further explained in these separate sections.

In addition, data preparation are completed in spaCy (Honnibal and Johnson, 2015) rather than NLTK. spaCy is a Python NLP alternative of NLTK capable of piping custom tokenization and data cleaning functions, and machine learning classifiers powered by neural networks. In addition to the basic NLP functions included in NLTK, spaCy offers API support which allows integration into the SaILS workflow. Data preparation for classification was done with spaCy for spaCy's potential for integrating the classifier into the SaILS workflow using the API. However, this is reserved for future work due to time constraint. The data preparation described below could have also been completed by NLTK with similar results. A schematic overview of incident classification is illustrated in Figure 5.

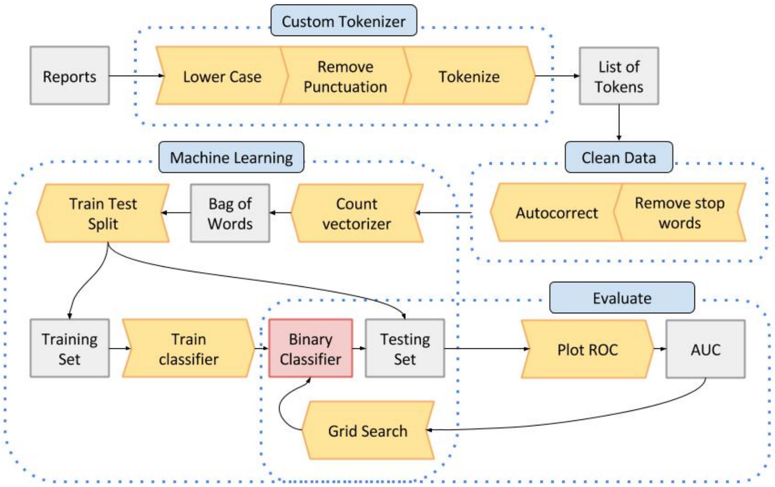


Figure 5: Schematic overview of incident classification. Data objects are represented by rectangular components, and data manipulation by yellow chevrons.

Tokenize

The default tokenizer in both NLTK and spaCy would split "50gy/10fx" into a single token. However, splitting it into "gy" and "fx" is advantageous as another string, i.e. the "gy" and "fx" in "50gy/5fx" can then inform the classifier along with those in "50gy/10fx". Other examples that require a custom tokenizer are "ptv40", and "lost(forgotten)", which are both split into single tokens by the default tokenizer.

Thus, a split function was created to replace that of the default tokenizer. The split function lowercased the reports, and removed all punctuation and numbers, and then tokenized the reports. The custom tokenizer then replaced the default tokenizer in the pipeline. Each report was then processed into a list of tokens.

Clean Data

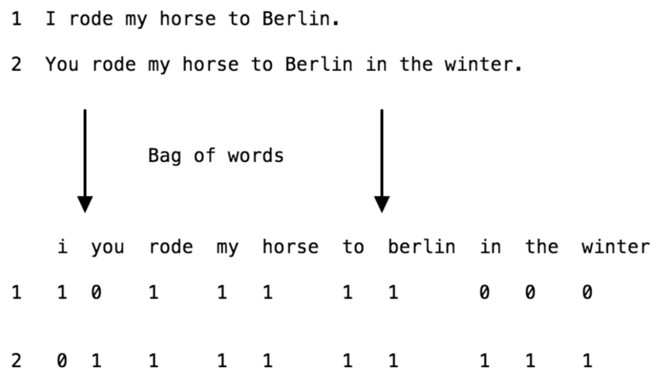
Data cleaning consisted of autocorrecting words and removing common words of little value such as “the” and “an” from the reports. They are common practice in NLP, and would improve the classifiers by improving the information content of the dataset. This was not done in the preliminary analysis because autocorrect often changes medical words such as “cbct” into a words that autocorrect recognizes, thus interfering with the UMLS tagging. This, however, would not inhibit classification as the all of the occurrences of a medical terminology all map to the same autocorrection. In addition, removing stop words was not included in the preliminary analysis because this would interfere with the calculations of the fractional POS compositions. Thus, the data cleaning in the incident classification was conducted separately from the preliminary analysis.

In data cleaning, Autocorrect 0.3.0 (jonasmccallum), a Python 3 spelling corrector, was first added to the pipeline as spelling errors were common in the reports. An example of spelling errors is “wothout” which would not be removed by the subsequent stop word removal.

Stop words are words that occur extremely frequently, and therefore contain little value for NLP. Examples of stop words are “without”, “at”, “the”, “which”, ’is”. A function that removes a token if it is in spaCy’s set of stop words is also added to the pipeline.

Machine Learning

The reports were then converted into numerical data in a bag of words by a count vectorizer. The bag-of-words model treats linguistic data as literally a bag of words, ignoring grammar and word order, but retaining the multiplicity, the number of times each word appears in a report (Zhang et al., 2010). The process of the count vectorizer converting linguistic data into a bag of words is illustrated in Figure 6 (dat). First, the set of unique tokens from all the incidents serve as the column names of a zero matrix, and the incidents serve the row names of the matrix. For each occurrence of a word in an incident, the cell that corresponds to the incident and the word increments by 1.



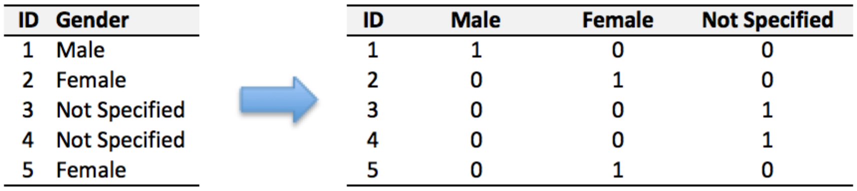
1 I rode my horse to Berlin.

2 You rode my horse to Berlin in the winter.

	i	you	rode	my	horse	to	berlin	in	the	winter
1	1	0	1	1	1	1	1	0	0	0
2	0	1	1	1	1	1	1	1	1	1

Figure 6: A illustration of the bag-of-words model.

One-hot encoding converted the multinomial classification of incidents by process step into multiple binary classifications by each menu choice of the process step. Figure 7 (cla) is an illustration of one-hot encoding. A new column is created for each menu choice of the process step, and the cell that corresponds to a given incident and menu choice is 1 if the incident occurred during the process step of the menu choice, and 0 otherwise. This was not done for post-treatment completion as there was only one incident report of this menu choice in the given dataset, and this is not sufficient for training a classifier.



ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

Figure 7: An illustration of one-hot-encoding.

The Python library for machine learning, sklearn (Pedregosa et al.), was then used to split the data into the training and testing sets, and to carry out the subsequent machine learning steps. A naive bayesian classifier algorithm (Friedman et al., 1997) was selected because the columns in the bag of words are assumed to be independent of each other. The naive bayesian classifier was trained on the training set, and tested on the testing set.

Evaluate

The output of the binary classifier is the probability from 0 to 1 of an incident belonging to themenu choice of the process step. Different thresholds can be set from 0 to 1 for different true positive and false positive rates. A receiver-operating-characteristic (ROC) curve plots the false positive rate on the x-axis and the true positive rate on the y-axis for different thresholds. The area under the curve (AUC) is a conventional method of comparing the performance of binary classifiers. The more the ROC curve hugs the upper left corner, the higher the the AUC, and the better the performance of the classifier. ROC curves were plotted for each binary classifier, and the associated AUCs were calculated using sklearn.

To prevent overfitting the classifier to words with low occurrence in all of the dataset, only words with an occurrence greater than an arbitrary threshold are included in the bag of words, and the columns headed by the other words are removed from the dataset used to train and test the classifiers. The grid search method of sklearn performs an exhaustive search of the threshold possibilities, and outputs the threshold and the associated classifier that offer the highest AUC.

Other Models

In this section, attempts to represent the reports as numerical data with other models other than the bag-of-words model are presented. These models were explored before the objective of classifying incident reports by process step was defined. As a result, the models were compared by their performance in classifying the incident reports by if the incident was a near miss, a menu choice of event type, an arbitrarily chosen data element of the NSIR-RT.

In these other models, the reports were not vectorized into a bag-of-words count vector. Instead, the POS and medical terminology occurrence columns from the preliminary analysis served as the numerical data for the the classifier. A new column for the a variable that is 1 if an incident is a near miss, and 0 otherwise is added to the Dataframe. Pearson correlation coefficients between the “near miss” column and each of the columns from the preliminary analysis were calculated. The occurrence of medical terminology were excluded from the further consideration for weak correlation. The three alternative datasets used to train the classifier were the POS with the strongest correlation, the 3 most strongly correlated POS, and all POS. The trained classifier were then evaluated by the methods described above, and their performance was compared to that of the bag-of-words model.

Results

Preliminary Analysis

Of the 519 incidents, 325 were actual incidents, 103 were near misses, and 91 were reportable circumstances. The length of the reports demonstrates high variance. The average word count of the incident description is 24.83±20.73 words. The corresponding histogram indicates that the average word count of the incident is left-skewed. The average word count of the incident descriptors is 5.13±3.98 words, that of the average sentence count of the incident description is 3.61±2.65 sentences, and that of the incident descriptor is 1.26±0.55 sentences. In the incident descriptions, the average of the noun and verb compositions were 0.42±0.21 nouns and 0.17±0.10 verbs, respectively, and both the associated histograms demonstrated bimodal distribution. The summary statistics of the preliminary analysis are shown in Table 1, and the histograms of the word count, noun composition, and verb composition are shown in Figures 8-10.

Descriptive statistics of the preliminary analysis	Incident description word count	Descriptor word count	Incident description sentence count	Descriptor sentence count	Incident description verb composition	Incident description noun composition
Mean	24.83	5.13	3.61	1.26	0.17	0.42
STD	20.72	3.98	2.65	0.55	0.1	0.21
min	3	1	1	1	0	0
0.25	12	2	2	1	0.12	0.29
0.5	19	4	3	1	0.18	0.36
0.75	31	7	4	1	0.24	0.47
Max	242	28	24	5	0.6	1

Table 1: Summary statistics of the preliminary analysis; Abbreviations: STD: standard deviation; min: minimum value; 25%: 25% percentile

The fractional POS compositions of incident descriptions grouped by the process step is shown in Table 2, and the medical terminology compositions of the incident descriptions are shown in Table 3. Notably, patient assessment/consultation achieved the highest word count of 42.48 words, and on-treatment quality management achieved the lowest word count of 19.78 words. In addition, the patient assessment/consultation was the highest in the composition of POS grouped as "other."

As expected, the procedure composition was the highest in the process steps associated with planning and delivery, and lower in those associated with review and verification. All of the process step groups were low in symptom, disease, and anatomy compositions.

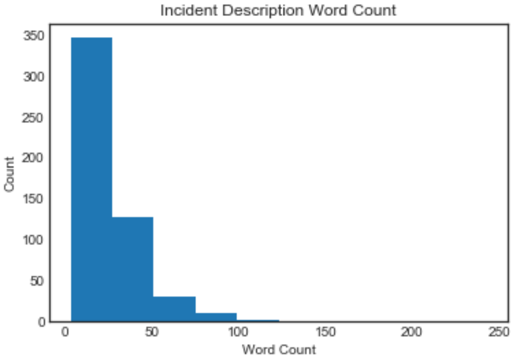


Figure 8: Histogram of the of incident description by word count

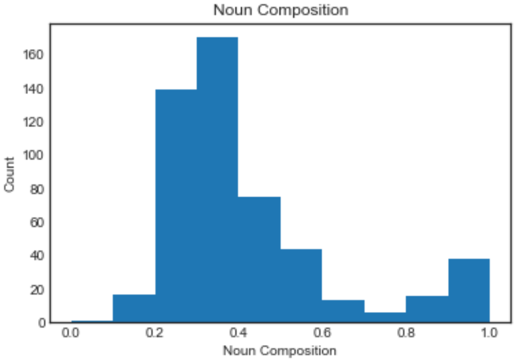


Figure 9: Histogram of the of incident description by noun composition

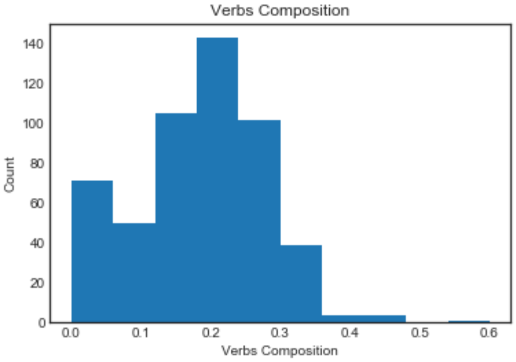


Figure 10: Histogram of the of incident description by verb composition

Completeness Analysis

24 out of 35 POS occurrences achieved a Pearson correlation coefficient with fictitiousness. The Pearson correlation coefficients of the 24 POS are shown in Table 4. The corresponding comparative histograms plotted are too numerous to show here. However, almost all of them are characterized by nearly complete overlaps. The 4 histograms with the most different distributions between the real and fictitious reports belong to the proper nouns, verbs, adverbs, and prepositions, shown in Figures 11-14.

Parts-of-speech compositions of incidents descriptions grouped by process step

Process step	Word count	Verb composition	Noun composition	Preposition composition	Pronoun composition	Adverb composition	Adjective composition	Other composition
Patient assessment/consultation	42.38	0.21	0.35	0.16	0.04	0.03	0.03	0.21
Imaging for radiotherapy planning	29.26	0.2	0.38	0.15	0.03	0.05	0.06	0.13
Treatment planning	26.82	0.2	0.38	0.17	0.01	0.04	0.06	0.15
Pre-treatment review and verification	24.39	0.19	0.4	0.15	0.02	0.05	0.06	0.15
Treatment delivery	27.65	0.18	0.38	0.16	0.03	0.05	0.06	0.16
On-treatment quality management	19.78	0.18	0.38	0.14	0.01	0.06	0.06	0.18
Other	26.11	0.22	0.41	0.18	0.06	0.04	0.03	0.11

Table 2: Fractional parts-of-speech compositions of incident descriptions grouped by process step

Medical terminology compositions of incidents descriptions grouped by process step

Process step	Medical composition	Procedure composition	Symptom composition	Disease composition	Anatomy composition
Patient assessment/consultation	0	0.06	0	0.02	0
Imaging for radiotherapy planning	0.01	0.05	0.01	0.01	0.03
Treatment planning	0	0.01	0.01	0.03	0
Pre-treatment review and verification	0	0.03	0.01	0.02	0.01
Treatment delivery	0.01	0.04	0.01	0.01	0.01
On-treatment quality management	0.01	0.02	0	0.01	0.02
Other	0	0.07	0.01	0.03	0

Table 3: Medical terminology compositions of incident descriptions grouped by process step

Incident Classification

The number of nouns correlated most strongly with "near miss", with the next three being the the number of words, determiners, and adjectives. Common examples of determiners are "the," this," "these," "my," "his," which," "whatever," etc. The classifier for "near miss" that was trained exclusively on the number of nouns achieved an AUC of 0.67; that trained on the top 4 most strongly correlated POS achieved an AUC of 0.70; and that trained on all POS also achieved an AUC of 0.70. The bag of words model achieved the highest AUC of 0.82. The ROC curves of the alternative attempts are shown in Figure 15. Thus, the bag-of-words model was chosen to classify incident reports by process step.

The classifiers achieved very high AUCs of 0.99, 0.99, and 0.98 in classifying imaging for radiotherapy planning (n=81), patient assessment/consultation (n=66), and other (n=15), respectively; a high AUC of 0.86 in classifying treatment planning (n=68); a medium AUC of 0.78 in classifying on-treatment quality

Pearson correlations between POS occurrences and fictitiousness

Word class	p value
Coordinating conjunction	0.01
Cardinal digit	-0.01
Existential there	-0.02
Foreign word	-0.01
Preposition/subordinating conjunction	0.04
Comparative adjective	-0.02
Superlative adjective	-0.02
Modal	0.03
Noun plural	0.04
Proper noun singular	-0.04
Predeterminer	-0.01
Possessive ending	-0.01
Personal pronoun	0
Possessive pronoun	0.03
Adverb	0.01
Comparative adverb	-0.01
Superlative adverb	-0.01
Particle	-0.04
Interjection	-0.01
Verb base form	-0.03
Verb 3rd person singular present	0.04
wh-pronoun	-0.01
wh-adverb	0

Table 4: Pearson correlation coefficients between parts-of-speech occurrences and fictitiousness

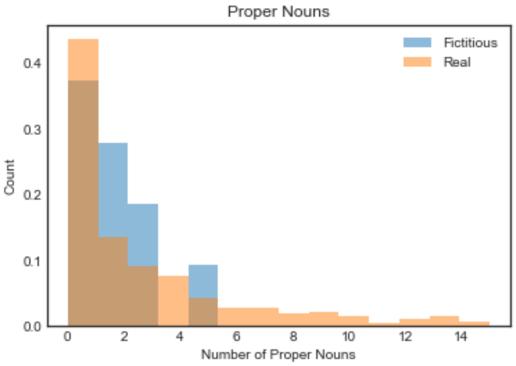


Figure 11: Comparative histogram between real and fictitious reports by proper nouns

management (n=9), and low AUCs of 0.69 and 0.65 in classifying treatment delivery (n=96) and pre-treatment review and verification (n=33). The ROC curves of the classifiers are shown in Figure 16.

Discussion

The number of actual incidents was three times greater than that of near misses and reportable circumstances. If the ratio of major incidents to minor incidents to near misses shown in Figure 1 remotely holds true, then

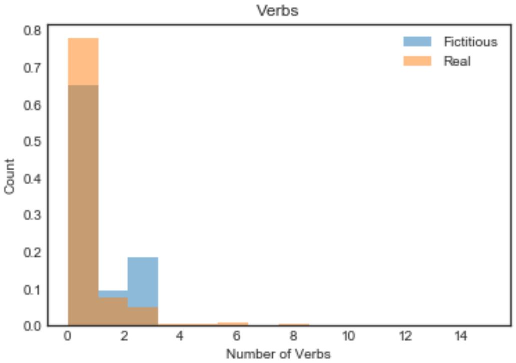


Figure 12: Comparative histogram between real and fictitious reports by verbs

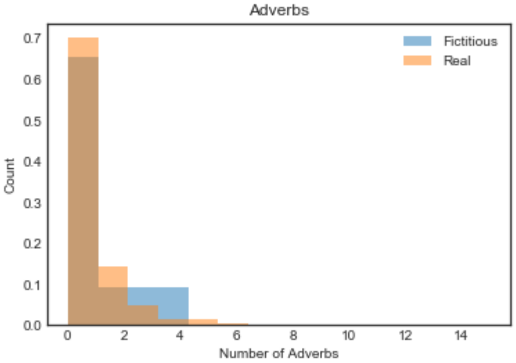


Figure 13: Comparative histogram between real and fictitious reports by adverbs

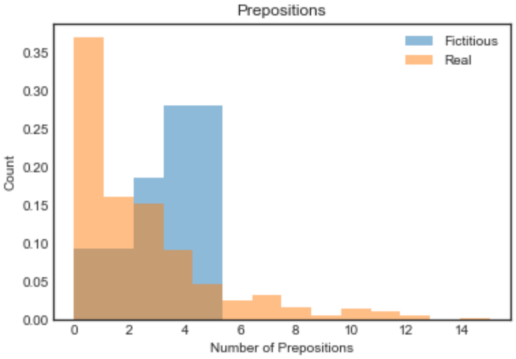


Figure 14: Comparative histogram between real and fictitious reports by propositions

many near misses and reportable circumstances are unreported. It is possible that such potential incidents are regular and have been perceived as acceptable inefficiencies and risks for valid reasons.

The high variance in the word counts of incident descriptions was partially explained by the variance in the word counts of incident descriptions grouped by process step . The average word count of patient assessment or consultation step was almost a third to two times greater than that of other process steps.

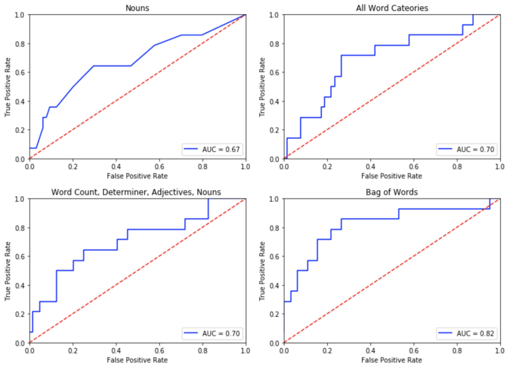


Figure 15: ROC curves of other models used to classify "near miss"

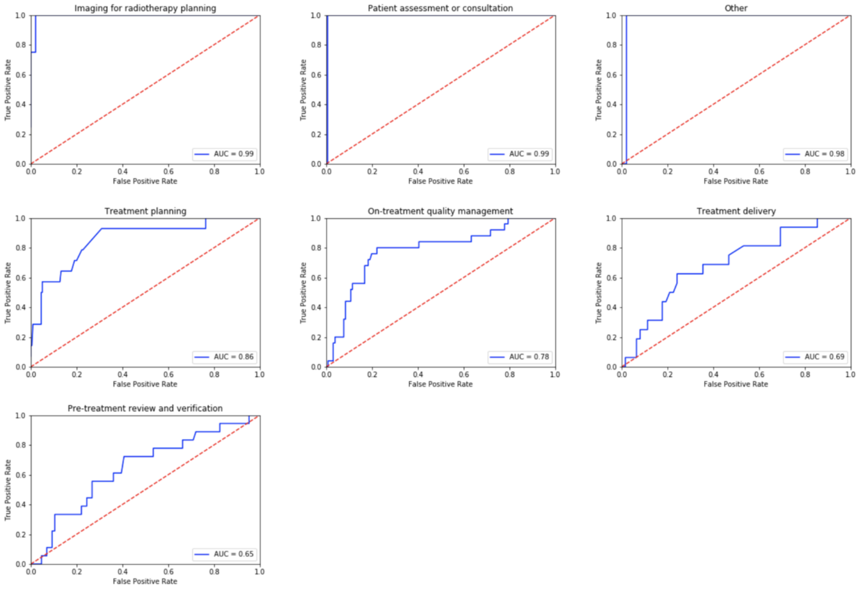


Figure 16: ROC curves of the classifiers of process step of incident occurrence

Incidents in other process steps often can be attributed to a single technical error, and both the problem and the correction in these incidents may be more straightforward. On the other hand, incidents in patient assessment/consultation most often involve scheduling, for which the causes are more varied and complex.

The medical terminology compositions indicate that nearly all of the process step groups contained low compositions of terminologies pertaining to medication, symptom, disease, and anatomy, and the composition of terminology pertaining to procedure is 5-7 times greater than that of other terminology compositions. This indicates a high focus on the procedural safety by the practitioners. A future direction for the project is allowing patients to fill out incident reports, which may contain a higher composition of terminologies pertaining to anatomy and symptoms. This in turn can potentially augment the focus on patient comfort in incident learning, and inspire a patient-centric approach to improving the quality of care.

The completeness analysis revealed a high difference in word counts of the real and fictitious reports, which can explain much of the differences in the number of POS. The near complete overlap in the comparative histograms of the real and fictitious reports may suggest that the context which is lost in reducing words into the number of the POS is an essential component of determining completeness. A review of the reports showed that an extremely short incident report can adequately describe the incident to a reader with domain knowledge, and lead to acceptable classification and investigation.

Imposing criteria that each report must meet to ensure completeness was also considered. However, such measures would not permit the omission of information implicit to participants with domain knowledge. Thus, this may exasperate participants, and diminish participation. On the other hand, integrating domain knowledge and context into NLP techniques to determine completeness is outside of the scope of this project.

The bag-of-words model outperformed the POS-occurrence models because the a bag of words retains the semantics of the reports. In addition, the bag-of-words model may increase in its superiority to the other models as the size of the dataset grows. This is because an increased number of entries may further strengthen the relevance of some words to the process step and weaken that of the words that contribute to overfitting. On the other hand, the increased dataset of POS occurrence will regress to the respective means of POS occurrence of each process step, which in itself is inadequate as a classifier, as revealed in the completeness analysis.

The classifiers for imaging for radiotherapy planning (n=81) and patient assessment or consultation (n=66) are excellent as they achieved AUCs of 0.99, and there were sufficient incidents of these process steps in the dataset for testing. The near perfect AUC of the menu choice of "other" (n=15) can probably be attributed to the low number incidents in the menu choice, leading to insufficient training and testing. The classifier for treatment delivery (n=68) was good as it achieved an AUC for 0.86. The performance of the classifiers for on-treatment quality management (n=9, AUC=0.78) and pre-treatment review and verification (n=33, AUC=0.65) likely suffered from the low number of incidents in these process steps. Finally, the classifier for treatment delivery was poor (n=96, AUC=0.69), in spite of the fact that the classifier had the largest training set as the highest number of incidents occurred in this process step. This may be attributed to the fact that the incidents in treatment delivery are highly varied. It follows that this classifier, as well as the others, can significantly improve from an increased dataset. In the case of aviation safety reports, 600 aviation safety reports were generated each month (Tanguy et al., 2016), providing an enormous and comprehensive training set. The classifiers of this project were trained on only two years of incident reports, and will likely be improved by a larger dataset.

Alternative algorithms are also worthwhile to explore to improve the performance. The algorithm used in the aviation safety reports was the Support Vector Machine algorithm (Tanguy et al., 2016). In addition, spaCy offers a multinomial text classifier powered by neural networks. The classifier object is exportable, and APIs can be easily created for integrating the classifier into the SaILS pipeline. Another algorithm alternative is the multinomial naive bayesian classifier by sklearn, which also offers API support. It is worthwhile to compare the performance of these alternative algorithms before deploying the classifier.

Since the bag-of-words model retains the semantics of the reports, it can in theory be applied to other data elements of the NSIR-RT taxonomy. Automating the classification of incident reports by these data elements is another future direction for the project.

Conclusion

Determining if a report is complete requires domain knowledge outside the scope of this project. However, NLP demonstrates potential for classifying radiotherapy incident reports by process step as well as other data elements in NSIR-RT taxonomy.

References

Why using one-hot encoding for training classifier. https://media.licdn.com/mpr/mpr/AAEAAQAAAAAAAAi3AAAAJDczM URL https://media.licdn.com/mpr/mpr/AAEAAQAAAAAAAAi3AAAAJDczM2FjOTBiLWE0OTQtNGJhMi04ODNmLTc4Zjg1OWEwZ png. Accessed on Wed, January 03, 2018.

Dataquest. https://www.dataquest.io/. URL https://www.dataquest.io/. Accessed on Wed, January 03, 2018.

SciPy: Open Source Scientific Tools for Python. https://www.scipy.org/. URL https://www.scipy.org/. Accessed on Fri, December 29, 2017.

Bethesda. UMLS Reference Manual. *National Library of Medicine*, 2009.

O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, jan 2004. doi: 10.1093/nar/gkh061. URL https://doi.org/10.1093%2Fnar%2Fgkh061.

Lionel T. E. Cheng, Jiaping Zheng, Guergana K. Savova, and Bradley J. Erickson. Discerning Tumor Status from Unstructured MRI Reports—Completeness of Information in Existing Reports and Utility of Automated Natural Language Processing. *Journal of Digital Imaging*, 23(2):119–132, may 2009. doi: 10.1007/s10278-009-9215-7. URL https://doi.org/10.1007%2Fs10278-009-9215-7.

D. L Cooke, P. B Dunscombe, and R. C Lee. Using a survey of incident reporting and learning practices to improve organisational learning at a cancer care centre. *Quality and Safety in Health Care*, 16(5):342–348, oct 2007. doi: 10.1136/qshc.2006.018754. URL https://doi.org/10.1136%2Fqshc.2006.018754.

E. C. Ford, L. Fong de Los Santos, T. Pawlicki, S. Sutlief, and P. Dunscombe. Consensus recommendations for incident learning database structures in radiation oncology. *Medical Physics*, 39(12):7272–7290, nov 2012. doi: 10.1118/1.4764914. URL https://doi.org/10.1118%2F1.4764914.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian Network Classifiers. *Mach. Learn.*, 29(2-3): 131–163, nov 1997. ISSN 0885-6125. doi: 10.1023/A:1007465528199. URL https://doi.org/10.1023/A:1007465528199.

Matthew Honnibal and Mark Johnson. An Improved Non-monotonic Transition System for Dependency Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclweb.org/anthology/D/D15/D15-1162.

J. Jehring and H. W. Heinrich. Industrial Accident Prevention: A Scientific Approach. *Industrial and Labor Relations Review*, 4(4):609, jul 1951. doi: 10.2307/2518508. URL https://doi.org/10.2307%2F2518508.

jonasmccallum. autocorrect 0.3.0. https://pypi.python.org/pypi/autocorrect. URL https://pypi.python.org/pypi/autocorrect. Accessed on Fri, December 29, 2017.

Punit Kaur, Mark D. Hurwitz, Sunil Krishnan, and Alexzander Asea. Combined Hyperthermia and Radiotherapy for the Treatment of Cancer. *Cancers*, 3(4):3799–3823, sep 2011. doi: 10.3390/cancers3043799. URL https://doi.org/10.3390%2Fcancers3043799.

Brian Liszewski and Carol-Anne Davis. To Err is Human to Learn is Divine: The National System for Incident Reporting in Radiation Treatment (NSIR-RT). *Journal of Medical Imaging and Radiation Sciences*, 46 (1):S29, mar 2015. doi: 10.1016/j.jmir.2015.01.095. URL https://doi.org/10.1016%2Fj.jmir.2015.01.095.

Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational*

Linguistics - Volume 1, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL https://doi.org/10.3115/1118108.1118117.

A. Mans, M. Wendling, L. N. McDermott, J.-J. Sonke, R. Tielenburg, R. Vijlbrief, B. Mijnheer, M. van Herk, and J. C. Stroom. Catching errors with in vivo EPID dosimetry. *Medical Physics*, 37(6Part2):2638–2644, may 2010. doi: 10.1118/1.3397807. URL https://doi.org/10.1118%2F1.3397807.

Wes McKinney. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011.

Michael Milosevic, Crystal Angers, Brian Liszewski, C. Suzanne Drodge, Eve-Lyne Marchand, Jean Pierre Bissonnette, Erika Brown, Peter Dunscombe, Jordan Hunt, Haiyan Jiang, Krista Louie, Gunita Mitera, Kathryn Moran, Tony Panzarella, Matthew Parliament, Spencer Ross, and Michael Brundage. The Canadian National System for Incident Reporting in Radiation Treatment (NSIR-RT) Taxonomy. *Practical Radiation Oncology*, 6(5):334–341, sep 2016. doi: 10.1016/j.prro.2016.01.013. URL https://doi.org/10.1016%2Fj.prro.2016.01.013.

L Montgomery, P Fava, CR Freeman, T Hijal, C Maietta, W Parker, and J Kildea. Development and implementation of a radiation therapy incident learning system compatible with local workflow and a national taxonomy. *J Appl Clin Med Phys*, Nov 2017.

Logan Montgomery. An evaluation of incident learning using the taxonomy of the Canadian National System for Incident Reporting - Radiation Treatment. Master's thesis, McGill University, 2016.

John Nerbonne. *Natural Language Processing in Computer-Assisted Language Learning*. Oxford University Press, sep 2012. doi: 10.1093/oxfordhb/9780199276349.013.0037. URL https://doi.org/10.1093%2Foxfordhb%2F9780199276349.013.0037.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Miguel Figueres-Esteban & Coen van Gulijk Peter Hughes. Learning from text-based close call data. *Safety and Reliability*, 36, 2016.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, sep 2010. doi: 10.1136/jamia.2009.001560. URL https://doi.org/10.1136%2Fjamia.2009.001560.

Ludovic Tanguy, Nikola Tulechki, Assaf Urieli, Eric Hermann, and Céline Raynal. Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*, 78: 80–95, may 2016. doi: 10.1016/j.compind.2015.09.005. URL https://doi.org/10.1016%2Fj.compind.2015.09.005.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, aug 2010. doi: 10.1007/s13042-010-0001-0. URL https://doi.org/10.1007%2Fs13042-010-0001-0.