Journal of Biomedical Informatics 90 (2019) 103103

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches

Maxim Topaz[a,b,*], Ludmila Murga[c], Katherine M. Gaddis[d], Margaret V. McDonald[b], Ofrit Bar-Bachar[c], Yoav Goldberg[e], Kathryn H. Bowles[b,d]

[a] School of Nursing & Data Science Institute, Columbia University, New York, NY, USA
[b] The Visiting Nurse Service of New York, New York, NY, USA
[c] Cheryl Spencer Department of Nursing, University of Haifa, Haifa, Israel
[d] School of Nursing, University of Pennsylvania, Philadelphia, PA, USA
[e] Department of Computer Science, Bar Ilan University, Tel Aviv, Israel

ARTICLE INFO

Keywords:
Natural language processing
Word embedding models
Nursing informatics
Text mining
Falls

ABSTRACT

Background: Natural language processing (NLP) of health-related data is still an expertise demanding, and resource expensive process. We created a novel, open source rapid clinical text mining system called NimbleMiner. NimbleMiner combines several machine learning techniques (word embedding models and positive only labels learning) to facilitate the process in which a human rapidly performs text mining of clinical narratives, while being aided by the machine learning components.

Objective: This manuscript describes the general system architecture and user Interface and presents results of a case study aimed at classifying fall-related information (including fall history, fall prevention interventions, and fall risk) in homecare visit notes.

Methods: We extracted a corpus of homecare visit notes (n = 1,149,586) for 89,459 patients from a large US-based homecare agency. We used a gold standard testing dataset of 750 notes annotated by two human reviewers to compare the NimbleMiner's ability to classify documents regarding whether they contain fall-related information with a previously developed rule-based NLP system.

Results: NimbleMiner outperformed the rule-based system in almost all domains. The overall F- score was 85.8% compared to 81% by the rule based-system with the best performance for identifying general fall history (F = 89% vs. F = 85.1% rule-based), followed by fall risk (F = 87% vs. F = 78.7% rule-based), fall prevention interventions (F = 88.1% vs. F = 78.2% rule-based) and fall within 2 days of the note date (F = 83.1% vs. F = 80.6% rule-based). The rule-based system achieved slightly better performance for fall within 2 weeks of the note date (F = 81.9% vs. F = 84% rule-based).

Discussion & conclusions: NimbleMiner outperformed other systems aimed at fall information classification, including our previously developed rule-based approach. These promising results indicate that clinical text mining can be implemented without the need for large labeled datasets necessary for other types of machine learning. This is critical for domains with little NLP developments, like nursing or allied health professions.

1. Background and significance

There is an increasing adoption of health information technology, such as electronic health records (EHRs), across clinical specialties and settings. For example, in the United States, 96% of hospitals and a similar percentage of outpatient settings have adopted EHRs for their everyday clinical care [1]. The amount of health-related data in EHR systems grows exponentially with the increasing adoption of health information technology. However, about 80% of all health data is captured as text (e.g., family history, progress summaries, radiology reports, etc.), which makes them less usable for healthcare practice and research needs [2].

The rapidly growing volumes of text data require new approaches for data processing and analytics. Natural language processing (NLP) is a field that combines computer science, computational linguistics and health expertise to develop automated approaches to extract meaning from clinical narrative data [3]. Traditionally, NLP was implemented by using either rule-based or machine learning approaches. Rule-based

* Corresponding author at: School of Nursing, Columbia University, 560 W 168th St, New York, NY 10032, USA.
E-mail address: mt3315@cumc.columbia.edu (M. Topaz).

https://doi.org/10.1016/j.jbi.2019.103103
Received 4 June 2018; Received in revised form 14 November 2018; Accepted 31 December 2018
Available online 09 January 2019
1532-0464/

NLP systems are often based on pre-defined vocabularies that include complex clinical logic. For example, a NLP system called "Medical Text Extraction, Reasoning and Mapping System" (MTERMS) was recently used to identify wound information in clinical notes [4]. The wound identification algorithm was based on a customized comprehensive clinical vocabulary of more than 200 terms describing wounds, and a set of rules that were used to accurately identify wound size and other wound characteristics. Creating a comprehensive vocabulary of wound-related terms was one of the most elaborate parts of the project [4]. The project team used a combination of literature review, standard terminologies and a diverse range of clinical notes to identify the candidate words that were then reviewed by the study experts.

One the other hand, machine learning NLP systems are often based on the probabilistic statistical approaches. For example, a recent machine learning NLP project developed several probabilistic algorithms (including Support Vector Machines and Decision trees) to evaluate the adequacy of information in radiology orders [5]. Generating training data for machine learning was one of the most time consuming steps of this project. To train the algorithms the project used ~2000 human-labeled radiology orders that were manually reviewed by two experts in medicine and informatics.

1.1. Positive-only labels learning

Recently, several new machine learning NLP approaches showed promising results in extending the current NLP methods. One research direction is to use positive-only labels learning to decrease time spent on text labeling [6]. In essence, positive-only labels learning is conducted with positively-labeled and unknown-labeled training set, as opposed to positively- and negatively-labeled training set in the traditional supervised machine learning approaches. This assumption is appealing, since once some positively-labeled examples are identified; there is no need to label large datasets of negative examples, unlike many other machine learning approaches. Several studies have shown so far that a positive-only label learning approach can potentially outperform other machine learning approaches [6].

There are some promising examples of using positive-only labels learning for medical text mining. For instance, Halpern et al. have used this approach to design a system for active machine learning with a human expert in the loop [7]. The system helped users create fast learning NLP classifiers in order to predict different patient phenotypes based on structured (e.g., patients' diagnoses and medications) and unstructured (e.g., discharge notes) clinical data. Halpern's approach requires clinical experts to specify "anchor variables" defined as key observations indicating that the patient should have a phenotype of interest. For example, presence of Atazanavir (Reyataz- an anti-retroviral medication) on a patient's medication list would indicate that this patient has high likelihood of having HIV/AIDS [7]. However, active learning systems, such as one suggested by Halpern et al. [7] still require significant clinical expertise and labeling efforts from the user, especially when it comes to identifying "anchor variables" in narrative clinical notes.

1.2. Word embedding

There are several promising emerging approaches that can potentially streamline labels identification for the user. For example, word embedding models help create multi-dimensional vector spaces for text representations, that were shown to boost NLP performance [8]. Word embedding models associate each word in a given vocabulary with a multi-dimensional vector, such that "similar" words receive similar vectors. The similarity of words is determined based on the words that they co-occur with within a specific context. The word embedding vectors are "trained" on large quantities of (unlabeled) text, and capture the word relations within the given text collection. Various algorithms exist for deriving word embeddings from text, among them the skip-gram with negative sampling model is considered to consistently produce good results [11]. The quality of the word embeddings improves with the size of the underlying text corpora they are trained on. Word embedding models are often used as a basis for further NLP tasks, such as named entity recognition.

Once the word embeddings are computed from text, one can use the resulting vectors to compute a list of the most similar words to a given word, where similarity is based on the contexts the words appearing in the training corpus. Similarity is computed based on the cosine-similarity between vectors [9]. Cosine similarity is commonly used to assess high-dimensional positive spaces. Cosine similarity ranges between 0 and 1, and words or phrases that appear in similar contexts have a higher cosine similarity measure. In an emerging body of literature cosine similarity between word vectors is started to be used to identify similar words or phrases in the biomedical domain [10].

1.3. Objective

We aimed to combine the latest NLP approaches to overcome some of the current challenges, and to create a novel rapid clinical text mining system called NimbleMiner. NimbleMiner combines several machine learning techniques that facilitate the process in which a human rapidly labels clinical notes for text classification, while being aided by the machine-learning components. This paper describes NimbleMiner's architecture and presents a case study comparing NimbleMiner's performance with a traditional NLP rule-based system. NimbleMiner includes a user interface (UI) component implemented in R, and we offer it as an open access system here: https://github.com/mtopaz/NimbleMiner.

2. Methods

First, we describe the general architecture of our system NimbleMiner. Then, we present a case study where we applied NimbleMiner to mine homecare clinical notes for presence of fall-related information. Finally, we compare performance between NimbleMiner and a traditional NLP rule-based system.

2.1. NimbleMiner's architecture

We combined several machine learning techniques to develop NimbleMiner. In general, NimbleMiner 's workflow is as follows:

- Stage 1- *Word embedding model creation*: The user selects a large corpus of clinical notes and chooses word embedding model parameters (i.e. word window width and how many similar terms are presented for every term entered by the user).
- Stage 2- *Interactive rapid vocabulary explorer*: The user provides a query term of interest, and the system returns a list of similar terms it identified as relevant. The user selects and saves the relevant terms.
- Stage 3- *Labels assignment and review*: The system uses previously discovered similar terms to assign labels to clinical notes (while excluding notes with negations and other irrelevant terms). The user reviews and updates, when needed, lists of negated similar terms and other irrelevant similar terms. The user reviews the clinical notes with assigned labels for accuracy.
- Stage 4- *Machine learning*: The user chooses a machine learning algorithm to be applied to create a predictive model. The model is then applied to predict which clinical notes might have the concept of interest. The user reviews the predicted notes and can go through stages 2–4 again to add new labels.

In the following paragraphs we provide a rationale and describe the system's architecture, also presented in Fig. 1. Throughout the system description, we provide examples of words and phrases indicating

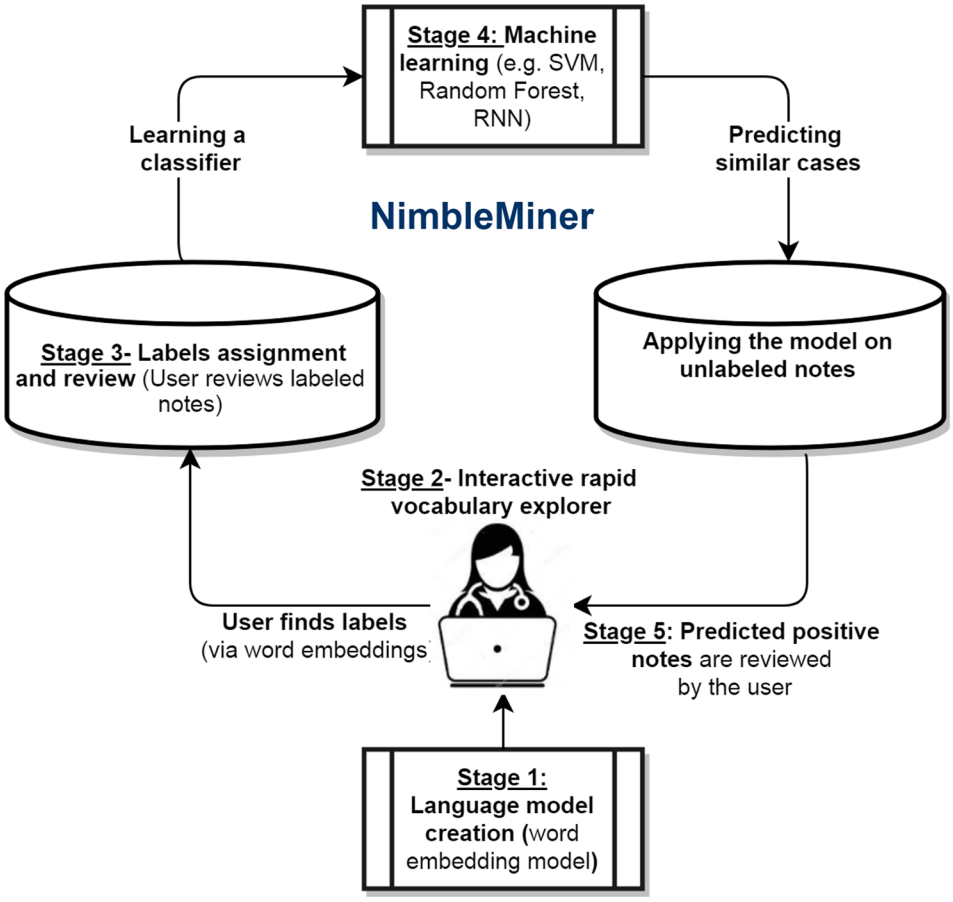

Fig. 1. NimbleMiner system architecture.

patient's fall history, in accordance with our case-study. NimbleMiner's user interface is presented in a series of appendices referred to throughout the methods section.

Stage 1: Word embedding model creation

To prepare clinical notes for the word embedding model training, we pre-process the notes to remove punctuation and lower-case all letters. Additionally, we convert frequently co-occurring words in the clinical notes into phrases with lengths of up to four words (4-grams) [11]. This is a common process in NLP where sets of co-occurring words are combined into phrases. For example, "pt fell yesterday" might be a common 3-gram. We use a phrase2vec algorithm with default settings to implement this in NimbleMiner [11]. Next, NimbleMiner's workflow continues with generating a word embedding vector for all the clinical notes in the corpus. We use a skip-gram model (specifically the word2vec implementation) [9,12] to provide NimbleMiner users with an interactive way to rapidly identify words similar to the concepts they want to find in the text.

Stage 2- Interactive rapid vocabulary explorer

Fig. 2 describes the interactive rapid vocabulary explorer process consisting of four steps.

Step 1: The user starts with inputting one or more keywords to the system, for example "fell" and "fall"(Appendix 2). The system suggests the most similar terms that appear in the same context (Appendix 3), for example "tripped", "fell down", "had fallen", etc. In our approach, potential similar terms are identified automatically based on the cosine similarity [9]. Terms are sorted based on the largest cosine similarity and a list of 50 similar terms for each target word is generated (50 words are the default setting and it can be changed).

In our approach, we refer to the similar terms as "simclins" (SIMilar CLINical terms). Similarly to the definition of a concept called "anchor" suggested by Halpern et al. [7] we define simclins as "words or phrases that have high positive predictive value in identifying the concept of interest". In other words, if a simclin is present, then the patient should almost always have the condition or a problem we are aiming to find. For example, phrases like "pt collapsed" or "she fell down" are considered simclins indicating the presence of fall history (Appendix 3).

Steps 2–3: NimbleMiner suggests potentially similar terms and the user selects simclins. We ask the user to choose only definite synonyms (simclins) from the list of similar terms. For each simclin chosen by the user, the system presents the user with random sentences including this specific simclin. This is done in order to exclude similar but not definite synonyms that would generate false positive matches. For example, one of the frequent similar words to falls is "hit" referring to hitting a head or a wall as a result of fall. However, "hit" is not a simclin since looking for sentences including this word would identify many non-fall-related cases, such as "… the door hits the tiles and prevents pt [patient] to enter the kitchen…". For each chosen simclin, NimbleMiner presents users with another list of 50 terms (while removing duplicates identified in the previous steps) and asks to iteratively choose new simclins until the system cannot suggest any more relevant similar terms.

Step 4: Steps 2–3 are repeated until (1) the user cannot identify any additional simclins based on expertise or literature, and (2) the user finished reviewing all system suggested potential new similar terms. The process of simclin discovery stops at this point.

In sum, interactive rapid vocabulary explorer allows users to create large vocabularies of simclins in a very short time. The output of this step includes a list of simclins reviewed by the user.

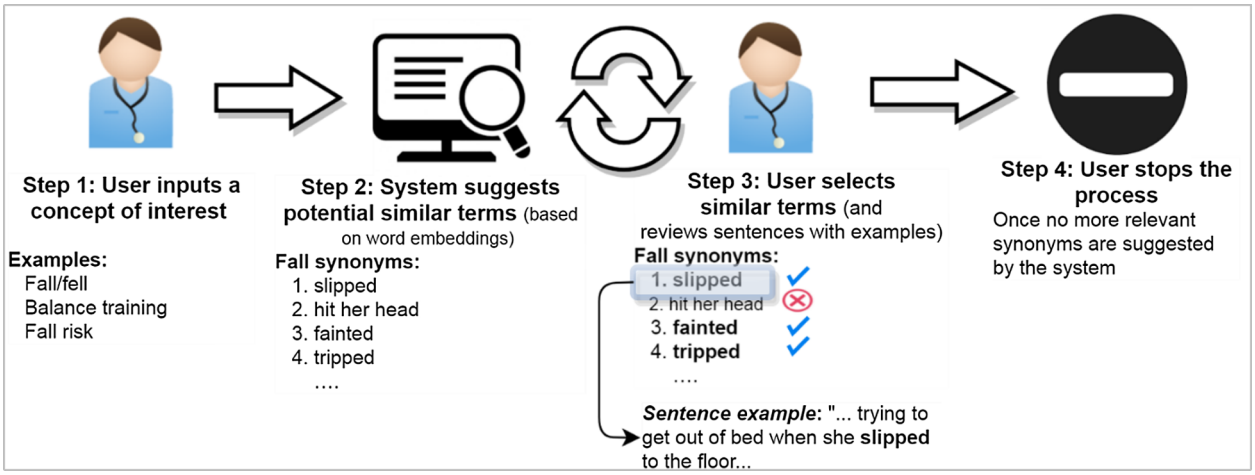

Fig. 2. Interactive rapid vocabulary explorer process.

Stage 3- Labels assignment and review

Next, NimbleMiner follows the positive-only labels learning framework [6] to prepare the data for the machine learning stage. In practice, NimbleMiner uses simclins to identify and label all the clinical notes where the phenomenon of interest is described. Regular expressions are applied to identify vocabulary terms in the notes, labeling notes as "positive" when the term is present, and "unknown" when the term is absent. For example, notes that include terms like "fallen down", or "pt collapsed" will be marked as positively labeled notes. Notes that contain only negated simclins are also labeled as "unknown" using a vocabulary of negations extracted from Negex (e.g., notes with expressions like "no falls"). Users are presented with frequencies of each specific negation in the clinical notes. Negated simclins identified in the notes can be edited, added or removed by the user (Appendix 4).

In addition, sometimes simclins are included in other irrelevant similar terms that should be removed from the positively-labeled notes. For example, simclins related to fall history like "fall" or "fell" are sometimes included in a larger irrelevant term, like "fall risk", "fall prevention intervention", or "fell off". These irrelevant similar terms are identified by the user during the vocabulary discovery phase (these terms were not selected as simclins and discarded by the user in the "Simclin explorer" tab) and are presented by the NimbleMiner for review (Appendix 5). Finally, the user can review information and examples of all the clinical notes with simclins, negations and irrelevant terms. NimbleMiner also offers a summary statistics, including frequencies and percentages of each type of notes (Appendix 6). The output of this stage is a list of clinical notes labeled as either positive or unknown.

Stage 4- Machine learning

To create a full training set for machine learning, NimbleMiner extracts all positively-labeled clinical notes and an additional randomly selected corpus of clinical notes without the positive labels (labeled as "unknown"). The default size of the unknown-labeled cases is set to be equal to the size of the positively-labeled cases. Both positive- and unknown-labeled corpora are then combined to create a training set that is processed by a machine learning classification algorithm of the user's choice (Appendix 7). Our preliminary experiments showed that the Random Forest algorithm outperforms other approaches (e.g., J48 Decision trees, Support Vector Machines) and we use this algorithm by default. The output of this stage is a trained machine learning model that can be used to predict wether a clinical note has a phenomenon of interest, for example information about patient's fall history.

Stage 5- Results review and further system refinement

This step is inspired by machine learning approaches where users interact with machines to implement rapid machine learning [13,14]. Based on the resulting machine learning model, the system generates predictions for the remaining ("unknown") notes and the user can review the notes predicted as "positive" to identify additional simclins and then specify them using the interactive rapid vocabulary explorer (stage 2). New positively-labeled notes (if any) are added to the previous positively-labeled notes and the system goes through another machine learning step with an updated positively-labeled corpus. Users can decide to go through vocabulary explorer and machine learning phases until saturation is achieved (i.e., no additional training is needed, as perceived by the user). When learning is completed, the user can export the positively predicted/labeled cases for further research or clinical purposes.

2.2. Case study

To test our proposed methods, we used a specific case-study aimed at classifying clinical notes regarding whether they contain fall-related information, namely fall history, fall prevention interventions, and fall risk, from narrative homecare notes. We defined "fall" as "An event which results in an individual coming to rest inadvertently on the ground or lower object" [15]; "fall prevention interventions" were defined as "instituting special precautions with the patient at risk for injury from falling" [16]. "Fall risk" in this project was defined as "presence of any balance issues, or presence of a nursing diagnosis of fall risk". We chose to divide fall history into three time-frames: (1) general fall history; (2) falls within two weeks; and (3) falls within two days of the note's date, in an attempt to estimate long-term and short-term fall incidence, ideally to determine whether falls happened within the current homecare episode or before that (an average homecare episode lasts about 30 days).

In our prior work, we built a traditional rule-based system to identify fall-related information. For the purposes of this case study, we compared the two systems (rule-based system and NimbleMiner) in terms of performance and time spent in implementing each system. Of note, both systems were co-developed and implemented by one research team.

2.2.1. Study data

This study used a large corpus of homecare visit notes (n = 1,149,586) for 89,459 patients treated by clinicians of the largest homecare agency in the United States (located in New York, NY) during

2015. Notes were completed by visiting homecare clinicians (e.g., nurses, physical/occupational therapists, social workers, etc.) using the agency's EHR after a patient visit. This study examined the narrative part of the homecare visit description. Visit notes ranged from lengthy admission notes (often written by a registered nurse) to shorter progress notes (e.g., physical therapy progress notes). The average note length was about 150 words.

2.2.2. Rule-based system development

To develop the rule-based system for this study, we followed a traditional information extraction methodology as follows [17]:

- *Literature review*: First, we conducted a thorough literature search in research databases (e.g., PubMed, Google Scholar, the Cumulative Index to Nursing and Allied Health Literature [CINAHL], etc.) to identify studies of fall incidence, fall prevention, and studies of text processing focused on falls. This helped us to compile a vocabulary of fall-related terms and expressions. We also explored standardized health terminologies (e.g., the Systematized Nomenclature of Medicine - Clinical Terms [SNOMED-CT], the International Statistical Classification of Diseases and Related Health Problems [ICD], the International Classification for Nursing Practice [ICNP®], etc.) to identify fall-related semantically similar expressions. For example, the class “fall” in SNOMED-CT has 61 mappings to similar classes in other terminologies, including “Unspecified fall”, “Multiple falls”, “Falling down”, etc. Finally, we used our team's expertise in nursing and homecare to identify additional words and expressions related to falls.
- *Notes review*: We extracted a random sample of clinical notes for 100 random patients from the study corpus. Overall, there were 1704 notes. Each note was reviewed by the study team (MT and KMG) and annotated for presence of relevant fall-related information, as described above. This step helped us understand the vocabulary used by the clinicians to document fall-related information. We discovered 29 mentions of falls (at different time periods, 16 unique patients), 52 instances of fall prevention interventions (30 unique patients) and 22 mentions of fall risk (18 unique patients). Overall, we found fall-related information in 75 clinical notes, which constituted less than 5% of the notes sample (75 out of 1704 notes).
- *Compiling a vocabulary and building regular expressions for fall information extraction*: Based on the previous steps, we then compiled a comprehensive vocabulary of fall-related terms and expressions and their possible lexical variations, including potential common misspellings (e.g., “fel” or “has fellen”), abbreviations (e.g., “falls prev prog” or “hx [history] of falls”). We iteratively used the vocabulary to create a list of regular expressions to capture fall-related information in the notes. For example, a regular expression “h/?-?o?x?\\s?o?f?\\s?fall[s]?” would capture all instances indicating either “ho”,“hx”,“h/x” of “fall(s)”. We also used a negations vocabulary from an open source negation module (Negex) to remove negated fall instances [18], for example “denies h/x of falls” or “declines fall prevention training”. To maximize system accuracy, we implemented several iterations of regular expression refinement and testing on randomly chosen batches of notes (5 batches of 100 patients) until system performance was deemed acceptable with low false positive/negative rates.

2.2.3. NimbleMiner's specifications

We use a skip-gram model implementation called word2vec and phrase2vec in R statistical package [11,12]. The model was trained on the full corpus of clinical notes. Parameters of the word embedding model were held constant based on parameters suggested in other studies of word embedding [19]. Specifically, we used a model with window width size = 10, vector dimension = 100, minimum word count = 5, negative sample size = 5 and sub-sampling = 1e−3. For each simclin, the user was presented with 50 potentially similar terms based on the cosine distance. At the machine learning stage, we first pre-processed all clinical notes using the following routine: we lowercased all the letters, removed punctuation, removed common English stop words, and removed numbers. We then created a document-term matrix representing each clinical note in the sample. Further, we constructed the training set using all positively labeled clinical notes and an equal size corpus of randomly selected unknown notes. The training set excluded the gold standard testing set described in the next section. We calculated 95% confidence intervals for system’s performance metrics using bootstrapping approach [20]. To accomplish that, we drawn with replacement 100 random samples, each including a random 80% of the training set data, and trained a Random Forest model based on each sample. We then calculated performance metrics (precision, recall and f-measure) for each sample on the test set. The Random Forest algorithm was iteratively trained on the 100 training sets with default settings (number of iterations = 100, minimum number of instances = 1, minimum variance for split = 1e−3, depth = unlimited).

2.2.4. Systems evaluation

To compare the two approaches implemented in this study (rule-based system vs. NimbleMiner), we created a gold standard human annotated testing set of 750 clinical notes. Since fall mentions are rare, we created a maximum fall information likelihood sample of notes with fall information based on the fall history indicator in the structured data (this indicator was not available for all the patients) and fall problem documented by the clinicians on the patients’ problem lists. Each note was annotated by two reviewers who are experts in nursing and health informatics, for a presence of fall information, including fall history, fall prevention interventions, and fall risk. Both reviewers achieved high interrater reliability on the reviewed notes (Kappa statistics = .84 indicating substantial agreement) [21]. Inter-rater agreement was calculated on the note level. Full agreement was achieved on all assigned categories in the gold standard testing set. Both of the systems were applied on the testing set and we calculated precision (defined as the number of true positives out of the total number of predicted positives), recall (defined as the number of true positives out of actual number of positives) and f-score (weighted harmonic mean of the precision and recall) for both approaches. We also calculated the overlap between the regular expressions and simclins lists for each fall-related domain. The degree of overlap was calculated by applying regular expressions on the simclins list and then counting how many simclins were identified as positive matches.

3. Results

3.1. Word embedding-based interactive vocabulary explorer results

Overall, our approach resulted in discovering an extensive lexicon of terms (simclins) related to patient falls (Table 1). For example, we found 83 different simclins referring to fall history. These fall history simclins included misspellings (e.g. “felll”) and lexical variants of fall-related terms (e.g., “tipped over” or “fallen backwards”). Larger lexicons of simclins were discovered for fall prevention interventions (n = 234) and fall risk (n = 233). Since only few simclins were discovered for fall history within 2 days/2 weeks, we identified a subset of temporal simclins referring to the two time periods. Specifically, we found 89 simclins referring to 2 days’ time-period (e.g., “1h ago”, “earlier this morning”, “this pm”, “yesterday”, etc.) and 188 simclins referring to 2 weeks’ time-period (e.g., “last week”, “last Fri”, “5d ago”, “1 wk ago”, etc.). We combined fall history simclins with temporal simclins to produce an extensive list of expressions referring to a fall within a certain time period (e.g., “fall this pm”, “collapsed earlier this morning”, “fallen backwards 1 wk ago”, etc.). We manually reviewed the resulting lists of simclins and excluded unlikely terms, which resulted in 612 simclins for falls within 2 days and 783 simclins for falls within 2 weeks.

Table 1
Word embedding-based interactive vocabulary explorer results.

Domain	N unique simclins*	Simclin* examples	Percent overlap between simclins and regular expressions (n of simclins matched by regular expressions)**
General fall history	83	"fell", "felll", "collapsed", "slipped", "tipped over", "fallen backwards"	39.7% (n = 33)
Fall history within 2 weeks (of note's date)	783***	"collapsed this week", "tipped over last Monday ", "slipped a 2d ago"	30.1% (n = 236)
Fall history within 2 days (of note's date)	612***	"collapsed this am", "tipped over 1h ago", "slipped earlier today"	29% (n = 178)
Fall prevention interventions	234	"fall precautions", "falls prevention mgt [management]", "gait training", "balance trg [training]"	52.5% (n = 123)
Fall risk	233	"high fall risk", "balance deficits", "difficulty walking" "generalized muscle weakness", "feeling unsteady"	45% (n = 105)

* Simclins (SIMilar CLINical terms) are defined as words or phrases that have high positive predictive value in identifying concepts in the domain of interest.
** Overlap was calculated as a percentage of simclins that were matched with relevant regular expressions.
*** First, we identified a subset of temporal simclins referring to the two time periods of 2 days/2 weeks (e.g., "1h ago", "earlier this morning", "this pm", etc.). Fall history simclins were combined with temporal simclins to produce an extensive list of expressions referring to a fall within a certain time period (e.g., "fall this pm", " collapsed earlier this morning").

We also calculated the degree of overlap between simclins and regular expressions. Regular expressions matched between 29% and 52.5% of simclins (Table 1). It took about 6 h (1.5–2 h for each fall-related category) to specify fall-related vocabularies using NimbleMiner.

3.2. Systems performance evaluation

Overall, the gold standard testing set (n = 750 clinical notes) contained 71 notes with fall history (including 41 notes with falls within two weeks and 20 notes with falls within two days of the note's date), 186 notes with fall prevention interventions, and 63 notes with fall risk indications. Table 2 presents the results of the system performances on the testing set. Overall, the rule-based system outperformed NimbleMiner in all domains in terms of precision. On the other hand, NimbleMiner outperformed the rule-based system in terms of recall and almost all F-measures (except for falls within 2 weeks). The overall micro averaged F-measure for NimbleMiner was 85.8% compared to 81% by the rule-based system.

Table 2
Systems' performance on the testing set.

	Recall (95% CI)	Precision (95% CI)	F-measure (95% CI)
General fall history			
Rule-based system	78.8	**96.6**	85.1
NimbleMiner*	**90.1 (88.1–91.3)**	88.4 (86.5–89)	**89 (87.9–90)**
Fall history within 2 weeks (of note's date)			
Rule-based system	76.2	**99.3**	**84**
NimbleMiner*	**88.9 (86.5–90.1)**	76.7 (73.2–78.1)	81.9 (79.9–84.3)
Fall history within 2 days (of note's date)			
Rule-based system	72.5	**99.2**	80.6
NimbleMiner*	**82.7 (81.1–84.2)**	83.2 (80.1–84.6)	**83.1 (81.4–84.2)**
Fall prevention interventions			
Rule-based system	74.2	**89.7**	78.3
NimbleMiner*	**86.1 (84–88.1)**	89 (87.7–91.2)	**88.1 (85.9–89.1)**
Fall risk			
Rule-based system	72	**94.2**	78.8
NimbleMiner*	**93.1 (90.1–94)**	82.9 (81.9–85.1)	**87 (85.7–88.9)**

Bold value represents highest metrics.
* For NimbleMiner, we report averaged performance metrics over 100 bootstrapping iterations with 80% of all training set data randomly sampled at each iteration. 95% confidence intervals are reported in parentheses.

4. Discussion

This study compared two approaches aimed at identifying fall-related information in narrative homecare notes. The first approach was creating a traditional rule-based system for fall information identification. The rule-based system showed similar or better performance compared to other studies developing NLP tools for fall information extraction [22,23]. For example, a study that examined fall history in inpatient notes developed a rule-based NLP algorithm with similar precision (97%) but lower recall (44%) [24]. Our rule-based system proves that it is possible to use homecare clinical notes to extract not just the fall history as was done previously, but also fall prevention intervention and fall risk information.

Our second NLP approach introduces a novel word embedding-based rapid vocabulary discovery. In other studies word embedding models have been used to automatically generate features for deep learning methods [25], or to perform information extraction from biomedical literature, such as PubMed abstracts [10]. In another recent study word embedding models have been applied to identify similar terms in PubMed abstracts and articles [26]. In the context of health NLP, our study introduces a framework where the human expert is interacting with word embedding models (skip-gram models) to rapidly evaluate and create lexicons of similar terms. Our results suggest that this approach is feasible and promising when applied to real clinical notes.

In NimbleMiner, we refer to similar terms as simclins. Our approach enabled rapid discovery of lexically diverse simclins, including abbreviations, misspellings, and other multi-phrase expressions. Rapid lexicon discovery is very promising, especially for health domains where large vocabularies of similar terms rarely exist (like nursing or allied health professions). In addition, discovered abbreviations and lexical variants that belong to a specific clinical domain or setting, can be used to augment existing vocabularies of similar terms (like standardized health terminologies). Our example of discovering temporal terms (indicators of two time periods) also suggests the possibility of combining different simclins to enrich vocabularies. For example, we were able to rapidly generate large lists of terms indicating falls within two days and two weeks using a combination of fall related and temporal simclins.

Another product of our approach is that during the process of vocabulary discovery, large lists of similar but irrelevant terms are discovered. These terms might look like simclins but they will produce false positive matches when identified in the text. For example, simclins related to fall history like "fall" or "fell" are sometimes part of a larger irrelevant term like "fall risk", "fall prevention intervention", "fell off", etc. These irrelevant similar terms are identified by the user during the

vocabulary discovery phase and their discovery is critical for accurate labeling of positive cases.

Regular expressions developed for the rule-based system matched between 29 and 52.5% of simclins. On the one hand, this suggests that simclin vocabulary was more elaborate and detailed, presenting new expressions and lexical variations, such as misspellings. In terms of positive-label only machine learning applied in this study, these lexical variants provided a useful way to capture additional clinical notes used for machine learning. However, many of the lexical variants discovered by the NimbleMiner might have been words or expressions with low frequency, for example uncommon misspellings. Further work is needed to create a detailed comparison between simclins and standard terminologies or other lists of similar terms or expressions.

Rule-based NLP systems are often crafted to fit a certain domain and type of clinical language. This was also the case in our study: our rule-based system achieved better precision than NimbleMiner in all domains of fall-related information in this study. However, NimbleMiner achieved higher recall in all domains with decent precision (> 73.2%), which resulted in overall high F-measure. For some NLP tasks, high recall is more important than high precision. NLP systems with low recall may miss cases, and the only way to find them is to read all the clinical notes in the corpus. On the other hand, while an NLP system with low precision may identify noisy positive cases; human reviewers only need to review a subsample of the cases the system flagged as positive to filter out the irrelevant cases. In the case of NimbleMiner, our goal was to achieve a higher recall to identify most of the potentially relevant cases and this is the result we achieved.

4.1. Limitations

This study has several important limitations. First, NimbleMiner was evaluated on one type of clinical notes and in one domain and it needs to be tested further to understand the system's generalizability. The use of NimbleMiner might be restricted to classifying documents based on phrases with high positive predictive value, such as fall-related phrases, and further work is needed to explore the generalizability of our approach. We only used one machine learning algorithm (Random Forest) and including other algorithms in further work can produce different results. In addition, both the rule-based system and NimbleMiner approaches were implemented by the same study team, which limits the validity of our results. Further work is needed to understand the desired properties of word embedding models (e.g., optimal word window width) in order to identify simclins in the most effective and efficient manner. Also, further research can explore the use of simclins as regular expressions for text annotation directly without machine learning, similarly to the rule-based system. We will also explore how to best apply user reviewed notes (especially those that were identified as negative notes) to improve our machine learning performance.

5. Conclusions

This paper aims to combine recent NLP approaches to overcome some of the current NLP challenges and create a novel rapid clinical text mining system called NimbleMiner. In this study, using NimbleMiner outperformed our previously developed rule-based system (in terms of F-measure and significant time savings) in identifying fall-related information in clinical notes. We believe that our approach, and an open-access system offered here, are of high interest to health researchers and clinical practitioners, especially for domains with little NLP developments, like nursing or allied health professions. Importantly, our promising results indicate that clinical text mining can be implemented without the need of large labeled datasets typically necessary for machine learning. Finally, our system can be potentially used by almost any clinician without special training in health informatics, which is a limitation of many existing NLP systems.

Conflict of interest

None of the authors have a conflict of interest to report.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103103.

References

[1] The Office of the National Coordinator for Health Information Technology, Health IT Quick Stats, (Website), 2016. < https://dashboard.healthit.gov/quickstats/quickstats.php > (accessed April 12, 2018).

[2] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, JAMA 309 (2013) 1351, https://doi.org/10.1001/jama.2013.393.

[3] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, IMIA Yearb, Med. Informatics Methods Inf. Med. 47 (2008) 128–144 (accessed April 12, 2018), <https://www.eecis.udel.edu/~shatkay/Course/papers/UEHROverview2008.pdf>.

[4] M. Topaz, K. Lai, D. Dowding, A. Zisberg, K. Bowles, L. Zhou, Creating a natural language processing algorithm to identify wound information in narrative Clinical Notes, Int. J. Nurs. Stud. [IN Press] (2016).

[5] W. Al Assad, M. Topaz, J. Tu, L. Zhou, The application of machine learning to evaluate the adequacy of information in radiology orders, 2017 IEEE Int. Conf. Bioinforma. Biomed. IEEE, 2017, pp. 305–310, , https://doi.org/10.1109/BIBM.2017.8217668.

[6] C. Elkan, K. Noto, Learning classifiers from only positive and unlabeled data, Proceeding 14th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD 08, 2008, p. 213, , https://doi.org/10.1145/1401890.1401920.

[7] Y. Halpern, S. Horng, Y. Choi, D. Sontag, Electronic medical record phenotyping using the anchor and learn framework, J. Am. Med. Informatics Assoc. 23 (2016) 731–740, https://doi.org/10.1093/jamia/ocw011.

[8] O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization, in: NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, pp. 2177–2185.

[9] T. Mikolov, G. Corrado, K. Chen, J. Dean, Efficient estimation of word representations in vector space, in: Proc. Int. Conf. Learn. Represent. (ICLR 2013), 2013, pp. 1–12. doi:https://doi.org/10.1162/153244303322533223.

[10] J.A. Minarro-Giménez, O. Marín-Alonso, M. Samwald, Exploring the application of deep learning techniques on medical text corpora, Stud. Health Technol. Inform. 205 (2014) 584–588 (accessed April 12, 2018), <http://www.ncbi.nlm.nih.gov/pubmed/25160253>.

[11] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, 2013, pp. 3111–3119. < http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality > (accessed October 2, 2017).

[12] R Core Team, R: A language and environment for statistical computing, 2013. < http://www.r-project.org/ > . (accessed April 13, 2018).

[13] D. Reker, G. Schneider, Active-learning strategies in computer-assisted drug discovery, Drug Discov. Today. 20 (2015) 458–465, https://doi.org/10.1016/j.drudis.2014.12.004.

[14] S. Marsland, Machine Learning: An Algorithmic Perspective (Google eBook), 2014, p. 457. doi:https://doi.org/10.1145/242224.242229.

[15] A.A. Zecevic, A.W. Salmoni, M. Speechley, A.A. Vandervoort, Defining a fall and reasons for falling: comparisons among the views of seniors, health care providers, and the research literature, Gerontologist 46 (2006) 367–376, https://doi.org/10.1093/geront/46.3.367.

[16] B.F. Miller, Encyclopedia & dictionary of medicine, nursing, and allied health., Saunders, 2003. < https://www.elsevier.com/books/miller-keane-encyclopedia-and-dictionary-of-medicine-nursing-and-allied-health/miller-keane/978-0-7216-9791-8 > (accessed April 12, 2018).

[17] J. Jiang, Information extraction from text, Min. Text Data, Springer US, Boston, MA, 2012, pp. 11–41, , https://doi.org/10.1007/978-1-4614-3223-4_2.

[18] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (2001) 301–310, https://doi.org/10.1006/jbin.2001.1029.

[19] B. Chiu, G. Crichton, A. Korhonen, S. Pyysalo, How to train good word embeddings for biomedical NLP, Proc. 15th Work. Biomed. Nat. Lang. Process. 2016, pp. 166–174, , https://doi.org/10.18653/v1/W16-2922.

[20] M. Wood, Statistical inference using bootstrap confidence intervals, Significance 1 (2004) 180–182, https://doi.org/10.1111/j.1740-9713.2004.00067.x.

[21] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Medica. 22 (2012) 276–282 (accessed April 12, 2018), <http://www.ncbi.nlm.nih.gov/pubmed/23092060>.
[22] S. Toyabe, Detecting inpatient falls by using natural language processing of electronic medical records, BMC Health Serv. Res. 12 (2012) 448, https://doi.org/10.1186/1472-6963-12-448.
[23] A.S. Navathe, F. Zhong, V.J. Lei, F.Y. Chang, M. Sordo, M. Topaz, S.B. Navathe, R.A. Rocha, L. Zhou, Hospital readmission and social risk factors identified from physician notes, Health Serv. Res. 53 (2017) 1110–1136, https://doi.org/10.1111/1475-6773.12670.
[24] B. Shiner, J. Neily, P.D. Mills, B.V. Watts, Identification of inpatient falls using automated review of text-based medical records, J. Patient Saf. 1 (2016), https://doi.org/10.1097/PTS.0000000000000275.
[25] M. Kholghi, L. De Vine, L. Sitbon, G. Zuccon, A. Nguyen, The benefits of word embeddings features for active learning in clinical information extraction, 2016. http://arxiv.org/abs/1607.02810 (accessed April 12, 2018).
[26] Y. Zhu, E. Yan, F. Wang, Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec, BMC Med. Inform. Decis. Mak. 17 (2017) 95, https://doi.org/10.1186/s12911-017-0498-1.