MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190558

NimbleMiner: A Novel Multi-Lingual Text Mining Application

Maxim Topaz [a,b]

[a] *School of Nursing & Data Science Institute, Columbia University, New York City, NY, USA*
[b] *Visiting Nurse Service of New York, New York City, NY, USA*

Abstract

This demonstration showcase will present a novel open access text mining application called NimbleMiner. NimbleMiner's architecture is language agnostic and it can be potentially applied in multiple languages. The system was applied in a series of recent studies in several languages, including English and Hebrew. The system showed good results in terms of text classification performance when compared to other natural language processing approaches.

Keywords:

Data mining; Medical informatics; Natural language processing,

Introduction

Natural language processing (NLP), is an important set of techniques aimed at extracting insignias from narrative health data. However, the vast majority of published NLP studies and existing NLP systems focus on processing information in English, while most of the languages around the world are described as low-resource languages in terms of NLP. More tools and approaches are urgently needed to overcome the resource barrier so advances in NLP can deliver more widespread benefits to health providers, researchers and patients.

Through this poster we will present a novel open access text mining application called NimbleMiner. NimbleMiner's architecture is language agnostic and can be potentially applied in multiple languages. So far, the system was applied in a series of studies in several languages. Examples include identifying fall in formation in clinical notes in English [1] and finding diabetes in clinical documents in Hebrew [2]. The system showed good results in terms of text classification performance when compared to other natural language processing rule-based or machine learning approaches. We will present the system and demonstrate the necessary steps in applying our pipeline. Examples in English and Hebrew will be provided throughout the presentation.

NimbleMiner's Architecture Overview

In general, NimbleMiner 's workflow is as follows (see Figure 1 for graphical depiction): **Stage 1- Language model creation** (Word embedding): The user selects a large corpus of clinical notes and defines word embedding model parameters including word window width (ranges between 1-10) and how many similar terms to present for every term entered by the user (ranges between 5-200). **Stage 2- Interactive rapid vocabulary explorer:** The user provides a query of terms of interest, and the system returns a list of similar terms identified as relevant (based on a combination of a cosine and Levenshtein distances) [1,2]. 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 narrative documents (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 narrative documents with assigned labels for accuracy. **Stage 4 - Machine learning:** The user choses a machine learning algorithm that will be applied to create a predictive model. The model is then applied to predict which narrative document might contain the concept of interest. The user reviews the predicted documents and can repeat stages 2-4 to add new labels.

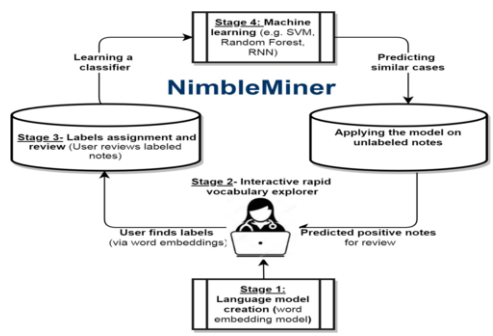

Figure 1– NimbleMiner's Architecture Overview

Follwing are further details on the system's architecture. Throughout the system description, we provide examples of words and phrases indicating patient's fall history, in accordance with a recent case-study [1]. The case-study aimed at identifying whether patients who received treatments from nurses in their homes had a history of falls. The narrative documents were nurses' clinical notes from each patient visit.

Stage 1: Language Model Creation (Word Embedding)

To prepare clinical notes for the word embedding model training, NimbleMiner pre-processes 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). 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) 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

The interactive rapid vocabulary explorer process consists of four steps.

Step 1: The user starts with inputting one or more keywords to the system, for example "fell" and "fall". The system suggests the most similar terms that appear in the same context, for example "tripped", "fell down", "had fallen", etc. In our approach, part of the potential similar terms are identified automatically based on the cosine similarity. Cosine similarity ranges between 0-1, with higher similarity being indicative of terms that are more similar. Terms are sorted based on the largest cosine similarity and a list of similar terms for each target word is generated. In our approach, the similar terms are called "simclins" (SIMilar CLINical terms). Simclins are defined as "words or phrases that have high positive predictive value in identifying the concept of interest". If a simclin is present, then the patient should almost always have the condition or a problem we are trying to find. For example, phrases like "pt collapsed" or "she fell down" are considered simclins indicating the presence of fall history. ***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. For each chosen simclin, NimbleMiner presents users with a list of additional 50 terms and asks to iteratively choose new simclins until the system cannot suggest any more new potentially 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 summary, NimbleMiner's interactive rapid vocabulary explorer allows users to create large vocabularies of simclins in a very short time. The output of this stage includes a list of simclins curated and selected by the user.

Stage 3: Labels Assignment and Review

Next, NimbleMiner implements the positive-only labels learning framework to prepare the data for the machine learning stage. Previously discovered simclins are treated as positive labels. 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 cases as "positive" when the term is present, and "unknown" when the term is absent. Negated simclins are labeled as "unknown" using the Negex vocabulary (e.g., notes with expressions like "no falls"). NimbleMiner allows users to add new negations beyond the exiting vocabulary and specify whether the negation should be searched for before or after the term (or both).

Finally, the user can review information and examples of all the clinical notes with simclins and negations. NimbleMiner also offers a summary statistics, including frequencies and percentages of each type of notes. The output of this stage is a list of clinical notes labeled as either positive or unknown. Based on the the positive-only labels learning framework, unknown narrative notes might contain negated simclins.

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"). Both positive-labeled 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. The output of this stage is a trained machine learning model that can be used to predict weather a narrative document, such as a clinical note, has a phenomenon of interest, for example information about patient's fall history.

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 cases (if any) are added to the previous positively-labeled cases 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.

Method Validation

NimbleMiner was validated and compared against other natural language processing approaches in a series of recent experiments. For example, NimbleMiner outperformed a well performing rule-based system in fall-related information classification tasks. NimbleMiner's overall F- score was 85.8% compared to 81% by the rule based-system [1]. In an additional ongoing project NimbleMiner was applied on a large sample of inpatient clinical notes in Hebrew to identify instances of diabetes mellitus (both insulin dependent type I diabetes and insulin resistant type II diabetes). NimbleMiner achieved high performance measures (average F-score = .94)[2].

Acknowledgements

This work would not be possible without Dr. Ludmila Murga, who created NimbleMiner's user interface. I am also thankful to Dr. Yoav Goldberg (Department of Computer Science, Bar Ilan University, Tel Aviv, Israel) who provided insights into improving NimbleMiner.

References

[1] M. Topaz, L. Murga, K.M. Gaddis, M. V. McDonald, O. Bar-Bachar, Y. Goldberg, and K. Bowles, Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches, *J. Biomed. Inform.* (2019) 103103. doi:10.1016/J.JBI.2019.103103.

[2] M. Topaz, L. Murga, C. Grossman, D. Daliyot, S. Jacobson, S. Rozendorn, E. Zimlichman, and N. Furie, Identifying diabetes in clinical notes in Hebrew: A novel text classification approach based on word embedding [IN PRESS], in: *Medinfo Proccedings*, 2019.

Address for correspondence

Max Topaz, mt3315@cumc.columbia.edu.