Using document ranking to classify clinical trial eligibility criteria

Ali Nemati[1], Masoud Khani[1], Jake Luo[2]

[1] University of Wisconsin-Milwaukee, College of Engineering and applied Science
[2] University of Wisconsin-Milwaukee, College of health Science

Objectives

- For years, the National Organization on Disability has endeavored to ensure fair treatment for people with disabilities in their communities.
- For people with disabilities, fair treatment is vitally important for public health, safety, and dignity.
- People with disabilities must be provided with access to services and opportunities that are equal to those available to people without disabilities.
- The U.S. Department of Justice is committed to achieving fair treatment for people with disabilities. The Department seeks to promote and protect the rights of people with disabilities. This includes the right of people with disabilities to be free from discrimination and to be full participants in the life of the community.

Approach

- Avoiding discriminating against people with disabilities.
- Assessing the efficacy of different document ranking techniques in comparing queries to detect discrimination against disabled people.
- The rise of artificial intelligence, and especially machine learning methods, can have an influence on the fair treatment of people with disabilities in the community.
- Improve the current status of algorithms used in information retrieval in two-class inclusion and exclusion.
- By using a state-of-the-art language model to understand clinical trial eligibility criteria (free-text criteria) and generate judgments of patients’ eligibility for trials.

Methodology

Data Description:
Using data from BHI departments at the University of Wisconsin Milwaukee that are not publicly available. In this data include official public registry of clinical trials, clinical eligibility criteria which are organized as a paragraph, a bullet list, or arbitrary user-specified topic categories.

Preprocessing:
The preprocessing step is necessary to improve the data in a way which is easier to understand and process in the algorithm. This step consists of three major techniques: lower case all the word, remove stop words, and stemming.

TF-IDF:
TF-IDF stands for Term frequency-inverse document frequency. We applied the tf-idf weight which is a weight often used in information retrieval (IR) or text mining for example criteria classification topic.

What is TF-IDF tell us?
A Statistical measurement to evaluation of how relevant a word is to a document in a database.

Applying N-gram:
N-gram stands for a sequence of N words. We applied Apply several ranking techniques and compare results by using TF-IDF include 100, 500, 1000 top term vocabulary with 1-2, 2-2 and 2-3 gram for our TF-IDFectorizer.

Random-Forest
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

N-Gram

Results

To obtain the best model, we used different N-Gram model with different top K Term Frequency. For evaluation we measure different metrics to assure the performance of the model such as:

- Accuracy: accuracy is the fraction of predictions our model got right.
- Recall: What proportion of positive identifications was identified correctly?
- Precision: What proportion of positive identifications was actually correct?
- F1-Score: To compare the performance of two classifiers , the F1-score combines the precision and recall of a classifier into a single metric as mean.

Conclusions

- Finding the inclusion and exclusion based on doctors patient’s text criteria.
- We could detect any inclusion and exclusion for people with disabilities using Machine Learning and categories these groups for fair treatment.
- We achieved a model with 94 percent accuracy and F1-Score.
- Random Forest is recognized as one of the most effective algorithms capable of achieving great performance.

Limitations

- The data did not have any label for training. The annotation has made manually.
- The algorithm is fast to train, but quite slow to make predictions once they are trained.
- The dataset’s text contains hashtags, mentions, hyperlinks, and even characters outside the ASCII range.

Bibliography


Contact Information

1. Nemati, Ali. Nemati@uwm.edu
2. Khani, Masoud. Khanim@uwm.edu
3. Luo, Jake, Ph.D Jakehuo@uwm.edu