REHS Neuroscience Gateway Internship - Electroencephalogram (EEG) testing and design of an algorithm to match preprints with published papers

Athrva Deshpande and Peter Eckmann
Mentors: Ramón Martinez, Anita Bandrowski, Dung Truong, Subhashini Sivagnanam, Amit Majumdar, Francesco Marini (UC San Diego, San Diego, California, 92122)

A novel algorithm to match preprints with published papers

Abstract

To evaluate the effectiveness of tools that alert researchers of problems in their preprints, a tool to match preprints with their corresponding papers was built. This tool combines three similarity metrics: authors, title, and abstract. Tests show that this tool is highly effective and precise with 99.5% accuracy, thereby outperforming the algorithms employed by the preprint servers, BioRxiv and ePMC.

Introduction

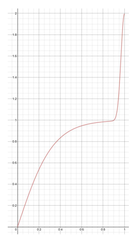
Preprints are papers released to the public that have not been formally peer-reviewed or published in a journal. Preprints often lead to later publications, but the final papers often differ to varying degrees from the original preprints, thus offering a window of opportunities for refining and improving the work. Based on this, we posed the overall question whether it may be possible to have positive impact on the published research by offering advice through automated tools such as SciScore that can detect as methodological or statistical errors and alert researchers of such errors. To test this proposition, it is necessary to determine whether any given preprint is ultimately published. As a first step, preprint servers offer matching tools, but the performance of these algorithms have not been thoroughly tested. To address this challenge, we set out to develop a new decision tool to determine whether a preprint was published at any subsequent time and to retrieve the matching publication, and to compare the performance with those of preprint servers.

Results

(1)

$$f(x,y,z) = \tanh(x) + \tanh(3y) - \frac{1}{e^{50y^2}} + \tanh(3z) + \frac{1}{e^{500(z-1)^2}}$$

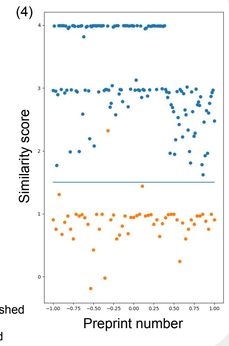
(2)



(3)

	Published (BioRxiv)	Not published (BioRxiv)
Published (our tool)	100	44
Not published (our tool)	0	56

(4)



(1) shows the formula that was used to produce a combined similarity metric using the individual abstract, author, and title metrics (*x, y,* and *z*, respectively). (2) plots the formula with only the variable *z*, to show how an individual metric is scaled. (3) represents the comparison between BioRxiv's algorithm and our algorithm for 200 representative preprints. Finally, (4) shows the total similarity score of these preprints. The score plotted represents the maximum similarity score between the preprint and a paper on PubMed. Blue dots represent preprints that were published, whereas orange points represent preprints that were not published. The published/not published status was confirmed by manual review. The line at $y = 1.5$ represents a possible cutoff value that could be used to determine if any given preprint was published or not with 99.5% accuracy.

Metrics

Title - The title of both the preprint and paper is tokenized and stop words are removed, producing sets A and B, which enter into equation (a).
Authors - The authors (last names only) of the preprints (A) and paper (B) were entered into equation (a).
Abstract - Equation (b) was used. W is the set of the 40 most uncommon words in a preprint's abstract, and a is the abstract of the PubMed paper. The function *occurrences(A, b)* counts the number of occurrences of A in the string b.

$$s(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad \text{(a)}$$

$$s(W,a) = \sum_{n=0}^{|W|} \sqrt{\text{occurrences}(W_n, a)} \quad \text{(b)}$$

Methods

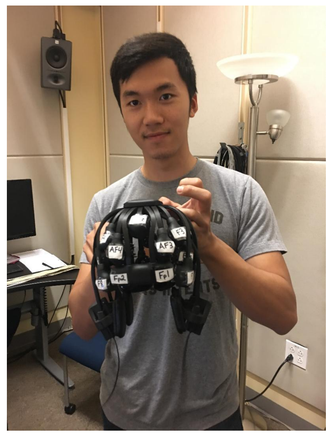
All programs were written in Python. When a preprint is entered, it goes through the following pipeline:

1. The authors, title, and abstract are scanned using a BioRxiv URL
2. Each piece of metadata is scanned against the entire PubMed database
3. Each metric produces a list of the top thousand matches, run on Comet at SDSC
4. For each metric, the best-matching paper is chosen. For at least one of these metrics, the matching paper is ranked top (>99%)
5. For each of the papers top-ranked in one metrics, all metrics are computed
6. The formula shown in (1) is evaluated, and the highest value is chosen
7. If this value is above a certain threshold (1.5), the paper is considered published; if not, it is not considered published.

Conclusions

We have found that common preprint servers, such as BioRxiv or ePMC, are almost always correct in matching preprints with papers in the cases where a preprint is marked as published. However, both servers make incorrect judgements on preprints where no published paper is given when there is, in fact, a published paper (~50% of papers that were not published according to the servers). Our tool avoids this pitfall, while retaining accuracy comparable to that of the servers.

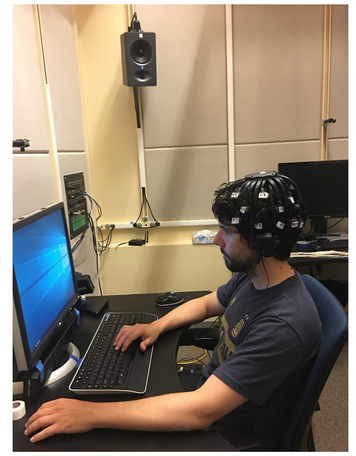
Preprocessing and analyzing electroencephalogram (EEG) data using MATLAB and EEGLAB

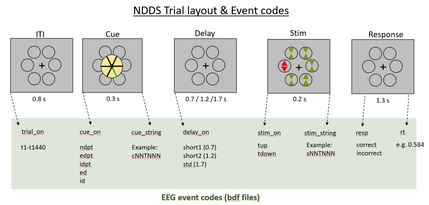


EEG: An electroencephalogram (EEG) is a test used to evaluate the electrical activity in the brain. Brain cells communicate with each other through electrical impulses. An EEG can be used to help detect potential problems associated with this activity.

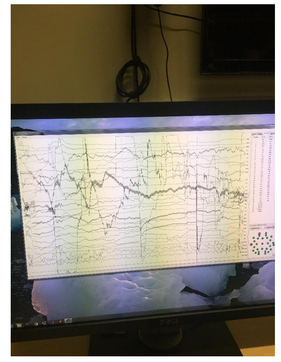
How to Extract Data

1. Collect raw data through your experiment
2. Preprocess the data to create through EEGLAB
3. Perform Subject analysis on the refined data



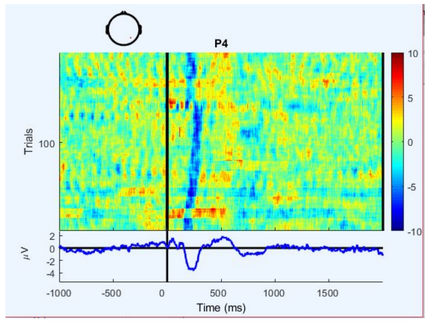


How the Experiment worked



Raw Data from Experiment

EEGLAB Preprocessing



Final Data

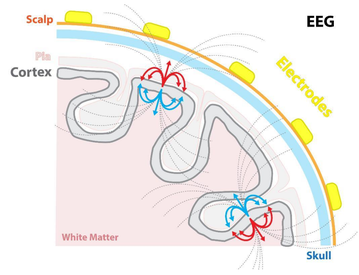


Diagram of the brain with EEGs

Overall, from this analysis, we can see where the visual senses act and where the motor senses come in. We can also see how EEGs work in a more detailed way.