Semantic Changepoint Detection for Finding Potentially Novel Research Publications

Bhavish Dinakar[1], Mayla R. Boguslav[2], Carsten Görg[2], Deendayal Dinakarpandian[3]

[1]University of California, Berkeley [2]University of Colorado Anschutz Medical Campus [3]Stanford University

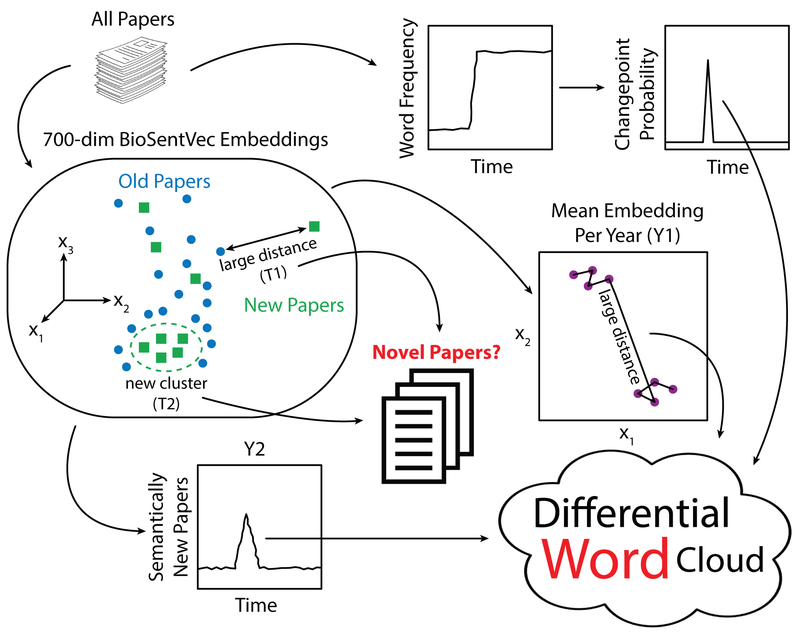
Stanford University

Identifying Novel Research

What if it were possible to identify papers that strayed from the mainstream? While many of these might end up as blind alleys, a subset of these might turn out to be harbingers of innovative, influential, or impactful directions in research. A few of the potential approaches to identify outliers, first-to-report, or first-in-field papers are topic modeling, clustering, trend analysis, citation network analysis, and machine learning approaches for predicting high impact papers. We present a set of strategies from changepoint analysis and text embedding to address two questions. Which are the papers in a research area that are substantially different from previous work? Which papers are part of a related cluster that is substantially different from previous work? We use infectious diseases from two different time scales to illustrate the approach - COVID-19 over a temporal resolution of weeks, and leprosy, considered a neglected tropical disease by the World Health Organization. This poster is based on the eponymous paper in the Proceedings of the Pacific Symposium on Biocomputing 26:107-118 (2021).

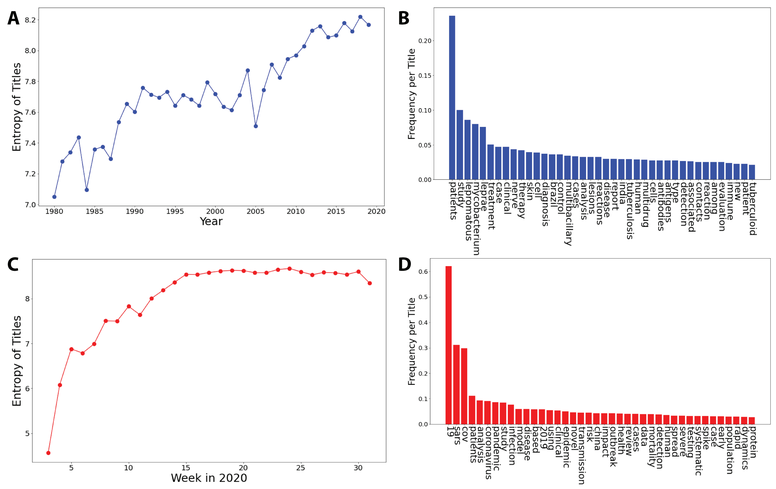
Methodology

We use the following approaches to detect potentially novel papers or subtopics in one temporal window (or subcollection) with respect to another. For instance, one may compare papers in 2020 with all preceding years (novel compared to entire research legacy). Alternatively, one may compare papers published in 2020 with those published in 2019 (a change in direction of research compared to recent past). We employ 4 different strategies: T1, T2, Y1 and Y2.

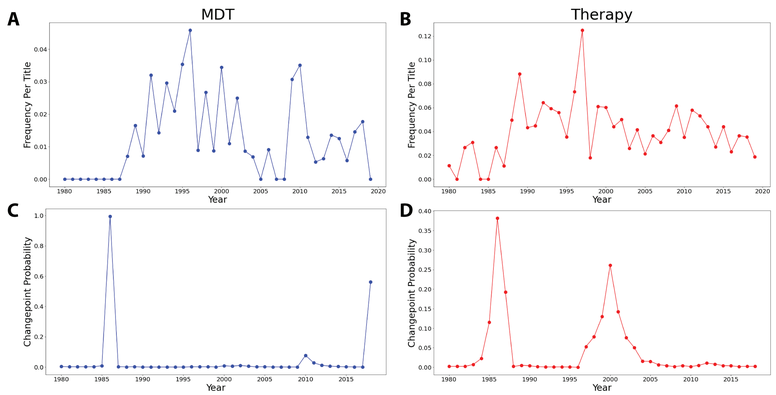


Changepoint Analysis

We use two different approaches to detect changes over time in the focus of research papers on a particular topic. The first approach consists of using changepoint analysis to detect changes in the frequency of words within titles or abstracts. The second approach consists of embedding titles in vector space and using the distance between titles as an approximation of the corresponding semantic difference. To illustrate these approaches, we have chosen to focus on a pair of contrasting infectious diseases, COVID-19 and leprosy. COVID-19 has a short history as a pandemic affecting millions of people, while leprosy is one of the oldest human diseases. While much progress has been made, and effective treatment is available when diagnosed early, around 200,000 new cases continue to be reported each year.



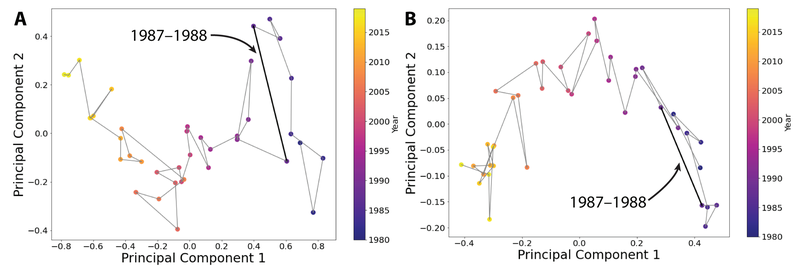
The corresponding changepoint analysis highlights points in time when there is a high probability of a significant change in the probability of the occurrence of a term in a title. Changepoint analysis finds identical changepoints for “MDT” and “Therapy,” indicating a surge of literature mentioning multidrug therapy for the treatment of leprosy. In fact, this corresponds to the period of excitement when (dapsone+rifampin+clofazimine) was recommended by WHO in the 1980s as curative treatment.



Semantic Space Embeddings

A low-dimensional projection of strategy Y1 is shown below. The long path between 1987 and 1988 is in alignment with previous results (changepoint analysis, Strategy T2) in this paper regarding the literature on multidrug therapy in leprosy. While the title might be the most succinct ‘sentence’ representation of the topic of a paper, a rhetorical or terse title may fail to convey the essence of a paper.

We therefore attempted to embed entire abstracts as an alternative version of strategy Y1. While this shows a possibly less noisy path from year to year, the variance (range of values on axes) decreases so that the semantic ‘hops’ from year to year become smaller. Results from these strategies could be used to calibrate and determine the thresholds for measures of novelty (projected path lengths or significant proportion of novel papers) to be indicative of novelty in the recent past.



We also estimated the proportion of titles in a given year that are located far away from any title in the preceding time period. A large change in the proportion of such titles in a given year may suggest a new and growing area of research.