# Linguistic Analysis of Research Publications' Abstracts

Aman Kumar [1] and Prithvi Poddar[2]

[1]Department of Mathematics, IISER-B
[2]Department of Electrical Engg and Computer Science, IISER-B

5 November 2020

## 1 Brief overview

In this project, we propose to conduct a linguistic analysis on the abstracts of research publications from the following 6 fields:

1. Computer Science
2. Physics
3. Mathematics
4. Statistics
5. Quantitative Biology
6. Quantitative Finance

The aim of this study will be to identify the linguistic features and patterns that might be the best choice as parameters, for classifying the publications, into the classes as mentioned above. After having completed the analysis, if time permits, we can test our study by running a classifier on the parameters of our finding and use our collected data about these parameters, to examine how well our classifier performs in classifying unseen publications.

## 2 Plan of action

### 2.1 Dataset

We will use the dataset provided here: Janatahack-Independence Day 2020 ML Hackathon. This dataset consists of a list of 20972 research publication titles and their abstracts, along with a tag of which class they belong to.

### 2.2 Steps in analysis

**Vocabulary:** Firstly we compute the vocabulary of the different classes of the research papers and perform suitable statistical analysis on the data we get and also study the lexical diversity among different fields.

**Seuencing of POS tags:** Next we perform POS tagging on the abstracts of the papers and model the sequence of the tags (preferably using n-gram models) and study the results of the same

**Word embedding:** We will perform word embedding on the abstracts and try to use the data for clustering and regression, which will be helpful even for classification tasks.

**Parsing and lexical analysis:** Finally we perform parsing on the data and do a lexical analysis to study any sort of sentimental pattern (or any other feature) that presents itself in different groups of research papers.

**Extras:** We are open to suggestions from our mentors regarding additional studies to be performed on the data, which they might feel will provide even more interesting insights on the data.

## 3 Outcome

We expect to be able to provide a report with a comprehensive study of the linguistic features that might help us distinguish the research papers for different topics. We are hopeful that we will get some interesting results.