<Title of the Project>

# Abstract

Founded in 2004, Yelp is a local-search service powered by crowd-sourced review forum. The company trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses. Often, a review describes various dimensions about a business and the experience of the user with respect to those dimensions.

Our main research question will focus on: What are the characteristics of leaders among users? or what characteristics make these user-generated reviews credible to other consumers?

To do so the aims of the project will be to analyze reviews, find criteria for leaders and respond to the following questions related to those leaders:

* Do reviews from leaders tend to be positive or negative?
* What are the most common words they used?
* Are reviews specific to one category of business?
* What are the most complaints about businesses? (Topic Modelling of the reviews) Can we classify them concerning the business categories?
* Are the categories of trending businesses different from top reviewed businesses?
* Are reviews influenced by user's friends? (relationship between users' friends and review patterns)

I will use a subset of the Yelp Challenge Round 12 dataset and apply text classification and sentiment analysis to respond to the questions above. All this will be done with R and Spark.

# Introduction

First, provide the context of the problem and then state the problem (your main research question). Second, write briefly that what are you proposing to solve this problem (don’t write details of the solution here). (You can use part of your abstract here)

Sentiment analysis of customer reviews has a crucial impact on a business's development strategy. Despite the fact that a repository of reviews evolves over time, sentiment analysis often relies on offline solutions where training data is collected before the model is built. If we want to avoid retraining the entire model from time to time, incremental learning becomes the best alternative solution for this task. In this work, we present a variant of online random forests to perform sentiment analysis on customers' reviews. Our model is able to achieve accuracy similar to offline methods and comparable to other online models.

# Literature Review

Write summary of the related papers that you reviewed here. Write the summary in your own words—don’t use the technical jargon from the paper that you don’t understand. Keep this section short—a short paragraph or few sentences about each paper you reviewed should be sufficient.

# Dataset

Give the description of the dataset that you are using along with the individual attributes you will or will not use in your analysis. Also mention the source of the dataset (where did you get it from). In case the data is curated and created by you please explain the details. Descriptive statistics of the attributes and datasets can also be provided here.

# Approach

Create a block diagram for the steps of your approach to clearly provide an overview. For example, if you first scrapped twitter, second applied NLP techniques to extract keywords, third labelled the tweets as positive and negative using a set of keywords, and fourth build a classifier, then you should create a box for each of the steps with arrows connecting one step to the next one. A sample block diagram is shown below.

Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done. For example, in the above case you would create subheadings for each of the steps.

## Step 1: <Name of the step>

Write details of the step 1. If there is any source code that you’d like to share then provide the link of the Github.

## Step 2: <Name of the step>

Write details of the step 2. If there is any source code that you’d like to share then provide the link of the Github.

………………………………………………………………………………………………………………………..

………………………………………………………………………………………………………………………..

## Step N: <Name of the step>

Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.