# Capstone Project Proposal

Name: Alex Ann Smith & Daniel Puig

## Objective

The objective of this project is to develop a web application that allows users to analyze the sentiment expressed in comments on social media posts and videos. Users will be able to input a link to a YouTube video, YouTube channel, Instagram post, or TikTok, and the application will retrieve a relevant sample of comments, perform sentiment analysis on them, and present the results in an interactive dashboard. This tool will provide insights into public opinion and reactions to online content, which can be valuable for content creators, marketers, and researchers.

## Impact

This project has the potential to provide a valuable tool for understanding public sentiment on social media. By automating the process of collecting and analyzing comments, it can:

* Help creators understand how their audience is reacting to their work and identify areas for improvement.
* Enable marketers to gauge public opinion on products, brands, and campaigns.
* Provide researchers with a tool to study online discourse and trends in social media sentiment.
* Offer insights into the dynamics of online communities and the factors that influence public opinion.

## Dataset(s)

The primary data for this project will be comments extracted from social media platforms. The following sources will be used:

* **YouTube:**
  + Data Source: YouTube Data API v3 (<https://developers.google.com/youtube/v3>)
  + Data Types: Text comments, timestamps, potentially user information.
  + Reliability: The YouTube Data API is the official and most reliable source for accessing YouTube data.
* **Instagram:**
  + Data Source: instaloader (<https://instaloader.github.io/>)
  + Data Types: Text comments, timestamps, potentially user information.
  + Reliability: instaloader uses an unofficial method and thus the reliability is less than an official API would provide. The package also may break if Instagram changes its format. Unfortunately Instagram does not provide an official API alternative, but we’ll try this out!
* **TikTok:**
  + Data Source: TikTokApi (<https://github.com/davidteather/TikTok-Api>)
  + Data Types: Text comments, timestamps, potentially user information.
  + Reliability: TikTokApi uses an unofficial method and thus the reliability is less than an official API would provide. The package also may break if TikTok changes its format. Unfortunately TikTok does not provide an official API alternative, but we’ll try this out!

## Approach

We’ll start out by creating a script that can pull comments from a specified video using YouTube’s Data API for us to be able to create a sample dataset to begin experimenting with the sentiment analysis.

Once we have a test set of data, there are a couple of options as to which NLP packages will be used in our sentiment analysis:

* **NLTK (Natural Language Toolkit)**
  + This NLP package appears to be reliable and somewhat widely used. It provides many options for the kind of sentiment analysis that is performed, such as semantic reasoning, which is useful for us since the structure and word usage found in comments can sometimes be unintuitive.
* **Textblob**
  + This NLP package is built on top of NLTK, so it shares many of the same capabilities, but it is geared to being a more beginner friendly NLP package. Since this package is quite similar to NLTK, and apparently easier to use, we will probably try this one out first. If we run into any issues that indicate that a different package will be needed, we will switch.

After determining which NLP package we will use, our next step will be to ensure that it suits our purposes well, i.e., it is able to accurately analyze the comments and draw out key conclusions. Comments are not the same as bodies of words and sentences that one might find in a book or newspaper, so we expect that the sentiment analysis may need some adjustments.

While we continue to explore the NLP packages and fine-tune the sentiment analysis we will also be conducting testing of the Instagram and TikTok 3-rd party sources to see if we can use them to reliably gather information for our application.

Once we have determined that our NLP package is a reliable method for sentiment analysis, we will then need to determine what our dashboard will look like and what kinds of information we want to display. Then, we will need to write our code for the dashboard, ensuring that the processed data is accurately transferred to be displayed. This will probably take a couple of iterations to make sure we are displaying what will be most helpful for the user, as well as get rid of any bugs that may pop up.

## Timeline

Feb 03 - Complete a working script for collecting YouTube comments and generate a sample dataset

Feb 10 - Start evaluating NLTK and Textblob for sentiment analysis on our dataset

Feb 17 - Choose a path forward for our sentiment analysis approach and explore TikTok/Instagram tools

Feb 24 - Create a mock-up of final dashboard and finalizing what analysis/data we want to surface for users

Mar 03 - Start development of application

Mar 10 - Spring break

Mar 17 - Continue application development, building core features and integrating YouTube API

Mar 24 - Test and refine core application features, have initial implementation of sentiment analysis

Mar 31 - Make final decision on feasibility for Instagram and TikTok data; Continue refinement of application

Apr 07 - Have a live preview of the website, reach out for users to test

Apr 14 - Begin preparing poster and video presentations; finalizing application features and acting on user feedback

Apr 21 - (mock presentations) continue debugging and refinement, finalize poster and video presentations

Apr 28 - (mock presentations) Final touches on presentation.

Apr 29 - (Poster presentation)

Apr 30 - (Video presentation due)

## Possible Issues

* **API Restrictions and Rate Limiting:** The YouTube Data API has quota limits, and unofficial methods for Instagram and TikTok may be subject to rate limiting or blocking. We will mitigate this by implementing caching, respecting API guidelines, and potentially allowing users to specify a limited number of comments to retrieve.
* **Sentiment Analysis Accuracy:** Sentiment analysis can be inaccurate, especially with nuanced language, sarcasm, and context-dependent comments. We will address this by carefully evaluating different sentiment analysis libraries and potentially exploring techniques to improve accuracy for social media text.
* **Unofficial Methods Breaking:** The third party packages may break at any time due to changes in the respective social media sites. We will mitigate this risk by prioritizing Youtube in our development as it has an officially supported API.
* **Limited Experience with Web App Deployment:** While we're enthusiastic about creating a functional web application, we have limited experience with the deployment and hosting of an application of this scale. We will attempt to mitigate this by dedicating time to researching and learning more about best practices with web app development.