Project Veritas

Project Proposal

Problem-Solved

People are constantly inundated with information positioned on an extreme-end of the spectrum for a given topic. This makes it difficult for people to become informed without having preconceived notions. Over time, this type of feedback loop results in echo chambers of isolated groups, largely to the detriment of society.

Goal

Project Veritas aims to fix this by presenting users with information that is clearly labeled. Users submitting a query will be presented with information that is marked *positive*, *negative* or *neutral*.

The intent is to obtain text from each of these three categories, such that the user is presented with information from varying perspectives, clearly marked. Done correctly, the goal is then achieved; a user clicking through to a given article will be forewarned if an article is likely to be polarizing.

User Demographics

The intended user is eventually any and every person able to interact with a computer or smartphone. There is a concern for results that might be sensitive to certain age groups.

Data

Data will be in the form of text content pulled from news sources. Ideally these sources will present a wide spectrum of opinion and vantage points, as this would offer the greatest selection for the user.

Possible APIs

- Twitter
- NewsAPI
- Natural Language Processing to classify results

Overview

There will be a search bar that a user can use to search topics. Ideally, there will also be a few suggestions of currently-trending topics that the user can click on to immediately query. Depending on the query time of the data pipeline, these presented categories might be “pre-fetched” to save time.

The user’s query will be received by the backend, where it will be URL-encoded and potentially have additional flags added to achieve the desired results. The query will then be sent to one or more information-source APIs to fetch results.

The returned results will then be fed into a NLP Sentiment Analysis model to gauge the sentiment of the text, after which results will be distributed into different “buckets.”

Once processed, the results that populate the page will be clearly marked *positive*, *negative* or *neutral*.

Users will be able to search without creating an account. If they do create an account and are logged in, additional features may be available, such as previous queries, favorite articles, or shareable links. Passwords will be stored as bcrypt hashes.

In regards to the database schema, there will be a User and Article model to start (a many-to-many relationship).

Stretch Goals

- If an article is too large to be returned directly via the news API, follow the link to the full article (for either sentiment analysis, or display on-site).
- Use a major cloud provider service (Azure, AWS, GCP) to gauge sentiment.
    - *Better*: Train and incorporate my own model.
- Add the ability for users to upvote, favorite and share links to articles.

Concerns

- NLP sentiment-analysis may not work well with shorter text often found in tweets. Additional flags may need to be used to pre-filter results.
- How to handle media (images, videos, urls) that are included in the article or tweet.
- Scope/breadth - I have not seen a service like this, so it will not be a simple implementation of a known concept.