

BaseLine for Project

1.Updated Problem Formulation

In an era dominated by information, staying informed is crucial, but navigating through vast amounts of news can be overwhelming. Introducing SentiNEWS, a revolutionary application designed to transform the way we consume news – making it efficient, accessible, and insightful.

We are basically creating an application where we can give you the best result based on user query by searching all through the web using the Google own information retrieval engine using SerpAPI. After searching through the results we take the top 5 most relevant results and use our summarizing technique to summarize the article by separating out Title and other text from the article and tokenizing each word to detect the news article Polarity and hence sentiments.

Our application is robust as the user query is searched directly through SerpAPI thus giving accurate results, works for any language making it multilingual, gives summary for easy and quick reading, gives pictorial demonstration of the news using DALL.E making the news attractive. Thus making it practical for day to day usage for a person who wants to remain updated with news without investing a lot of time.

2.Literature Review

In 2017, [Tarun B. Mirani](#) and [Sreela Sasi](#) from Gannon University, in their paper: *Two-level Text Summarization from Online News Sources with Sentiment Analysis*, employed a two-level text summarization approach with sentiment analysis to extract important sentences from online news articles covering various topics such as politics, sports, health, science, and movie reviews. The study utilized an extractive summarization method to generate summaries from two to three news articles for each topic. The extraction-based approach involved identifying important sentences and arranging them in order of their importance. Sentiment analysis was then performed to understand variations in news articles from different sources, determining positive, negative, or neutral opinions. Additionally, they used the ROUGE metric to evaluate the performance of the summarization. The study concluded that the positive polarity in the text processing indicated a public inclination towards government policy. The research utilized Python language and various packages for tasks such as fetching URLs, pre-processing news articles, and evaluating the performance of the summarization using the ROUGE metric. Overall, they presented a novel approach to text summarization and sentiment analysis, contributing to the advancement of automatic text summarization and sentiment analysis from online news sources.

Link: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8076735>

Shortcomings in this paper: The paper lacks exploration or discussion on the implementation of multilingual support. It overlooks the integration of keyword-based image generation.

In 2013, [Alexandra Balahur](#), [Ralf Steinberger](#), [Mijail Kabadjov](#), [Vanni Zavarella](#), [Erik van der Goot](#), [Matina Halkia](#), and [Bruno Pouliquen](#), in their paper: *Sentiment Analysis in the News*, focused on sentiment analysis in news articles, particularly in English language news. They identified the differences between sentiment analysis in highly subjective text types like product reviews and news articles, emphasizing the need to clearly define the target of any sentiment expressed and to restrict the analysis to the immediate context of the target. The authors conducted experiments to test the relative suitability of various sentiment dictionaries and attempted to separate positive or negative opinion from good or bad news. They also discussed the impact of category-defining word lists on sentiment analysis and the challenge of interpreting sentiment that is not expressed lexically in news texts. The authors outlined their experiments and evaluations, including the impact of using different lexicons and word windows, as well as the errors encountered in sentiment classification, such as misclassifying sentences as neutral due to lack of explicit sentiment words, misclassifying sentiment due to irony, and problems related to co-reference. The paper concluded by outlining future work, including evaluating the impact of using negation and valence shifters, extending the lexica for additional languages, and assessing methods to compare opinion trends across sources and time.

Link: <https://arxiv.org/ftp/arxiv/papers/1309/1309.6202.pdf>

Shortcomings in this paper: The paper also lacks discussion on multilingual support implementation along with integration of keyword-based image generation.

In 2009, [Mijail Kabadjov](#), [Josef Steinberger](#), [Bruno Pouliquen](#), [Ralf Steinberger](#), and [Massimo Poesio](#) presented a paper on "*Multilingual Statistical News Summarisation: Preliminary Experiments with English*." The paper aimed to develop a generic approach for summarizing multilingual news clusters, particularly those produced by the Europe Media Monitor (EMM) system. The authors highlighted the challenges of developing a multilingual news summarizer within the EMM system and emphasized the need for multilingual capabilities, which have been less addressed in the text summarization literature. The paper introduced the use of Latent Semantic Analysis (LSA) and multilingual entity disambiguation to enhance the summarization of news clusters, aiming to provide succinct and comprehensive summaries for decision-makers within the European Union who utilize the EMM system. The authors discussed various approaches to text summarization, including shallow linguistic analysis, machine learning, and more sophisticated approaches, emphasizing the multi-faceted nature of text summarization research. They also presented preliminary experiments with the TAC 2008 data, demonstrating promising improvements over a summarization system ranked in the top 20% at the TAC 2008

competition. The paper concluded by outlining future steps, including incorporating intra-document co-reference resolution and expanding the range of normalized entity information to enhance the summarization process.

Link:

https://www.researchgate.net/publication/221155821_Multilingual_Statistical_News_Summarisation_Preliminary_Experiments_with_English

Shortcomings in this paper: Though this paper implements multilingual support, it is only limited to major European languages such as English, French, German, Spanish, Italian, and others. It's coverage may not extend to Indian languages. Again this paper also lacks integration of keyword-based image generation.

In 2015, [Ubale Swati](#), [Chilekar Pranali](#), and [Sonkamble Pragati](#) conducted a study on sentiment analysis of news articles using a machine learning approach. The study aimed to determine the overall sentiment expressed in news articles. The authors highlighted the importance of sentiment analysis in tracking a company's behavior over time and its application in social media monitoring to gauge consumer sentiment towards a brand. The study also delved into the challenges of fine-grained sentiment analysis, particularly in the context of news articles, and the techniques and algorithms used to improve accuracy. The authors discussed the framework and techniques employed for sentiment analysis, including crawling and extraction, data preprocessing, feature extraction, sentiment identification, scoring, and classifier training. They also referenced relevant works and algorithms, positioning sentiment analysis as a valuable tool for extracting insights from news articles and social media to aid decision-making processes for businesses and organizations. They also referenced works that addressed semantic parsing and fine-grained sentiment analysis, highlighting the challenges posed by the variety of ways in which opinions can be expressed in news articles. Overall, the study provided a comprehensive overview of sentiment analysis of news articles using a machine learning approach, emphasizing its significance in various domains and the reliability and efficiency of the proposed methodology.

Link: https://www.iraj.in/journal/journal_file/journal_pdf/12-127-1430132488114-116.pdf

Shortcomings in this paper: The paper lacks summary conciseness alongwith implementation of multilingual support. It also lacks keyword based image generation.

How is our implementation better than others?

-> None of the above papers used any API to fetch news but we are using SerpAPI which is based on Google's news search engine to get the best news pages available on the internet. -> We are

also generating keyword based images so that it could quickly summarize and gain the attention of the users by providing a more comprehensive user experience.

-> Multilingual support is one of the features which we want to incorporate in our implementation. Given the global nature of news and the diversity of languages in online content, the absence of multilingual support limits the applicability and scalability of the proposed approach to non-English news sources.

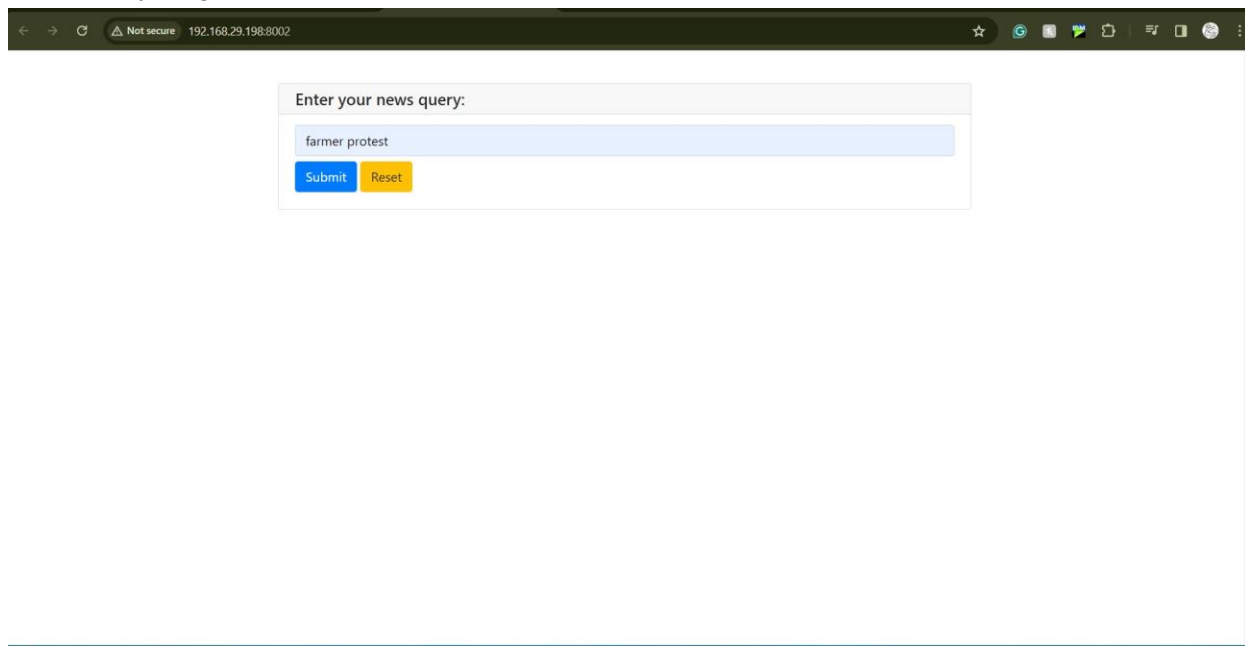
3. Prototype/Code

Code for the Prototype:

Provided in baseline.ipynb

Screenshots of the early implementation of SentiNews App:

User Query Page

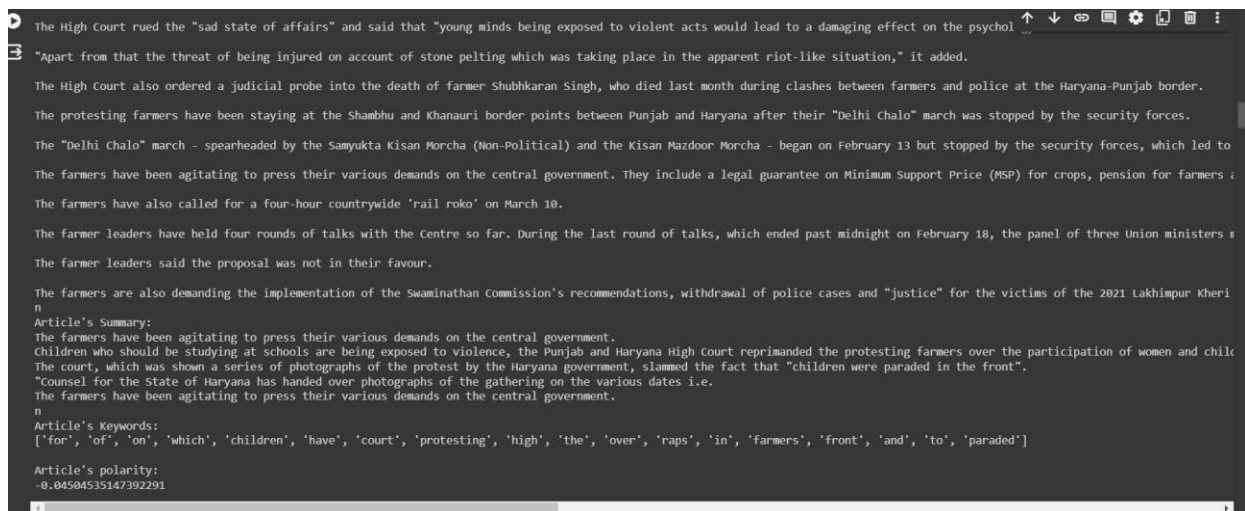
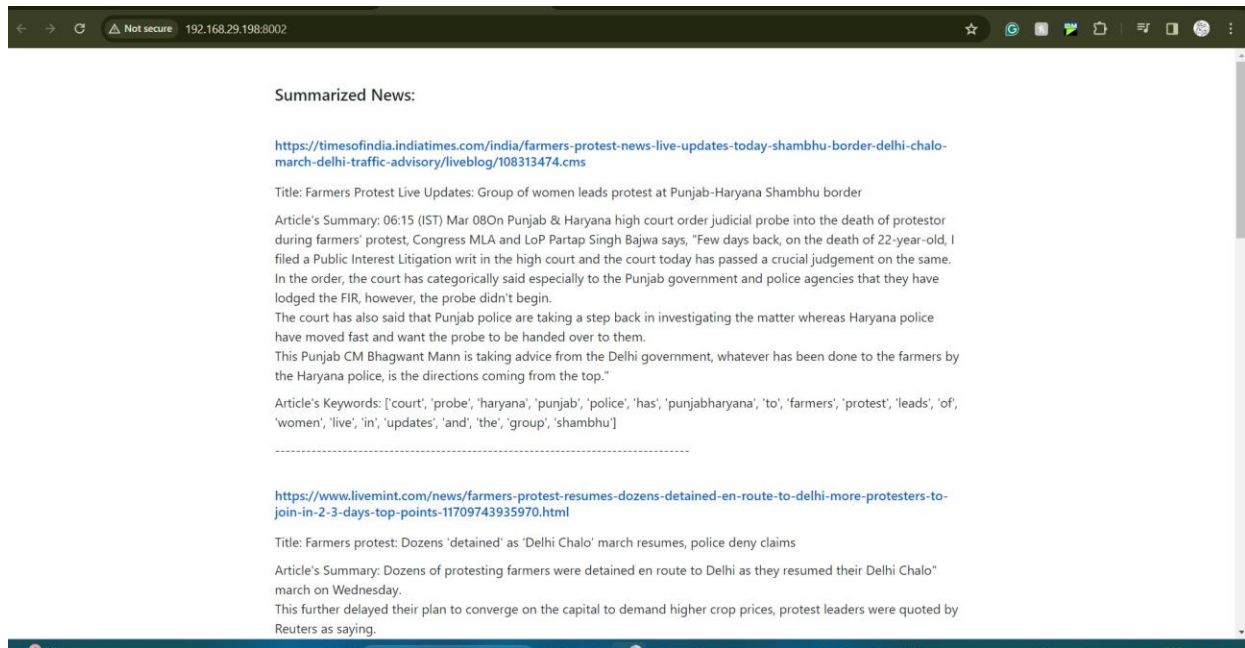


Enter your news query:

farmer protest

Submit Reset

Summary of top 5 results based on Query



About Screenshots:

We have implemented a webpage to take user query and based on serpAPI results we can provide a summary for top 5 results in another page using pywebio.

In another implementation we have found the polarity of the document based on the keywords present in them.

We are currently in process to integrate both these implementations, one more issue we are encountering is that multilingual support is only limited to provide summary, but more complex tasks such as polarity is not handled for the Hindi language , also the pictorial summarization of the news article is not ready yet. In the later stages of these assessments we will be able to implement all of these missing features.