Data Mining Project Proposal

Team :

1. Prathik Bharath Jain
2. Akarsh Reddy Tatimakula

Introduction :

In the digital age, a company's online presence on platforms like Twitter, YouTube, and Tumblr can make or break its success. Our project aims to simplify this process by developing a system that, with just a company name, collects data from multiple social media platforms via API's and scrapes the data, processes it using Natural Language Processing (NLP), and provides valuable insights, including sentiment analysis and trending topics. This innovative solution will empower businesses to monitor and respond to online conversations effectively, adapt their strategies in real-time, and enhance their online reputation management capabilities, ultimately providing a competitive edge in the digital landscape.

Related work :

Analysis of online presence is a crucial aspect for any organization to make itself marketable. There are some related researches done for instance :-
"Online reputation measurement of companies based on user-generated content in online social networks" by Hossein Shad Manaman, Shahram Jamali and Abolfazl AleAhmad. https://www.sciencedirect.com/science/article/pii/S074756321530073X

2)"Evaluating Online Reputation Monitoring Systems" by Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke & Damiano Spina.https://link.springer.com/chapter/10.1007/978-3-642-40802-1_31

These research papers provide adequate research motivation for Natural language processing additionally we will mine data from API's.

Proposed work :

The project will involve the development of a data collection and analysis system capable of gathering relevant data from various social media platforms, including Twitter, YouTube, and Tumblr and scrape data from few open sources. The collected data will consist of text-based content, such as tweets, comments, and posts, associated with the provided company name. To extract meaningful insights from this data, we will leverage Natural Language Processing (NLP) techniques, including sentiment analysis and topic modeling. Sentiment analysis will allow us to determine the overall sentiment (positive, negative, or neutral) of the content related to the company, providing an understanding of public perception. Topic modeling will help identify trending topics and discussions surrounding the company, enabling businesses to stay informed about relevant conversations in real-time.

We will start by focusing on a select group of companies to develop and refine the system's capabilities. Once we have a robust and effective solution in place, our goal is to expand its applicability, allowing it to seamlessly accommodate any company. This phased approach

will ensure that we deliver a tailored and reliable system for our initial users, while also preparing it for broader adoption across diverse business entities

Evaluation :

The project's success will be assessed using a multifaceted approach. We will measure the accuracy of sentiment analysis and topic modeling through quantitative metrics like precision and recall. Additionally, the system's effectiveness in identifying trending topics and its real-world impact on businesses will be key evaluation criteria. Along with this the visualizations created can be considered in evaluating the system.

Milestones :

1.Data collection : Data Collection(API access) , Preprocessing,Data Storage (*October)*
2. Data cleaning : Removing Stop words , Tokenization. (*October November)*
3. Training NLP model : Semantic Analysis (*November)*
4. Test and evaluation : Evaluating the model (*November - Early Decembe*r)
5. Data Analysis and Visualization . (November - *Early December)*