e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

UNVEILING SOCIAL SENTIMENTS: A COMPREHENSIVE ANALYSIS OF REDDIT POSTS

Shreyas K*1

*1Student, Department Of Computer Science And Engineering, SJB Institute Of Technology, India.

DOI : https://www.doi.org/10.56726/IRJMETS65909

ABSTRACT

Social media has changed the way people communicate, as individuals and organizations can share their opinions, thoughts, and experiences in real-time. This large reservoir of user-generated content provides a valuable resource for sentiment analysis, an important endeavor in understanding public opinion and sentiment trends. This paper explores various methodologies and techniques for analyzing sentiments in social media data, with challenges posed by the informal, noisy, and context-dependent nature of social media language. The natural language processing techniques used include text preprocessing, feature extraction, and sentiment classification algorithms such as Support Vector Machines (SVM), Naive Bayes, and advanced deep learning models like Recurrent Neural Networks (RNNs) and Transformer-based architectures. Real-world datasets are used on platforms such as Twitter, Facebook, and Instagram to classify sentiments in the categories positive, negative, and neutral.

The key findings reveal the superiority of deep learning models in capturing nuanced sentiment expressions and call for domain adaptation techniques to adapt to language variations across platforms. Further, the paper discusses the practical applications of sentiment analysis in brand perception management, market analysis, and crisis detection. Through experimental validation and comparative analysis, this research advances state-of-the-art sentiment analysis methodologies tailored to the distinctive characteristics of social media data. The proposed framework not only enhances decision-making for stakeholders but also offers valuable insights into public sentiment trends and behaviors across online platforms.

Keywords: Sentiment Analysis, Social Media, NLP, Deep Learning, SVM, Naive Bayes, RNNs, Transformers, Public Sentiment, Market Analysis.

I. INTRODUCTION

Social media has changed the way people express opinions, experiences, and knowledge. Among them, Reddit has become a strong platform for diversified discussions and communities. The sentiment expressed in Reddit posts gives important insights into the public attitude towards different topics, products, and events. Sentiment analysis, therefore, is the process of extracting and analyzing those sentiments to understand user perspectives and trends.

This project looks into the specific case of content extraction from Reddit posts to present meaningful sentiments. Through the determination of issues of unstructured text, contextual variability, and casual language specific to Reddit, the development of such a robust sentiment analysis tool could be achieved. The insights acquired would be a crucial input in informed decision making by businesses, researchers, and the communities themselves concerning this platform in general.

The objective of this project is to develop an automated system that employs advanced NLP techniques to analyze and quantify the sentiments expressed in Reddit posts. The system will categorize the sentiments into one of three classes: positive, negative, or neutral. This system will enable businesses and organizations to understand public opinion, refine their engagement strategies, and optimize their presence on Reddit.

Challenging issues persist with sentiment analysis on Reddit, beginning with volumes and velocities in the millions across posts and comments. Manual parsing of content to extract insights within such vast sums is not practically feasible, giving an added thrust towards automation. Complexity further arises linguistically due to significant usage of colloquial vocabulary, abbreviations, emojis, and other idiomatic expressions as a norm across Reddit. Sentiments in discussions on Reddit can also be context-dependent, and the system has to interpret and categorize sentiments correctly, even in ambiguous situations. Another major challenge is sarcasm and irony. These expressions are common on Reddit and can drastically alter the sentiment of a post,

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

making accurate analysis difficult. All other noise and spams existing in the data of Reddit needs to be removed so that proper analysis can focus on meaningful things.

Several important components fall under the scope of this project. Data collection will specifically focus on Reddit, gathering posts and comments from subreddits that may reflect public sentiment. The project will apply data preprocessing techniques to deal with noise, spam, and other irrelevant content typical to Reddit in order to make the analysis meaningful. It will also focus on developing and training sentiment classification models that can categorize the posts on Reddit as positive, negative, or neutral, but always giving special care to sarcasm, irony, and contextual variations. In addition to sentiment classification, the system will include visualization and reporting tools to allow businesses to easily interpret trends in sentiment and to generate reports with actionable insights. Finally, trend analysis will be employed to forecast future sentiment trends, helping businesses anticipate shifts in public opinion and adapt their strategies accordingly.

This project is expected to result in an automated sentiment analysis system specifically for Reddit, capable of classifying sentiment accurately and providing valuable insights. This will enable businesses and organizations to understand public perception expressed on Reddit and engage more effectively with Reddit communities. Businesses will be able to identify opportunities, address issues, and make informed decisions by analyzing sentiment trends on their social media strategy. Overall, this tool will be the most important source for businesses managing their online presence and reputation on Reddit.

II. LITERATURE SURVEY

This section reviews related works in sentiment analysis, highlighting their approaches and limitations. While most tools focus on binary classification (positive/negative sentiment), few address the unique dynamics of platforms like Reddit.

Adobe Social Analytics

Adobe Social Analytics evaluates the influence of social media on businesses by correlating user conversations with metrics like revenue and brand value. It uses natural language processing (NLP) to classify sentiments.

Brandwatch Sentiment Analysis

Developed in the UK, this tool classifies sentiments as positive, negative, or neutral. It is commercially available and caters to businesses looking to monitor their social presence.

Sentiment140

Designed by Stanford graduates, Sentiment140 analyzes sentiments on Twitter, supporting English and Spanish. It categorizes tweets into positive, negative, or neutral sentiments.

Social Mention

A real-time analysis tool for monitoring user opinions on social media platforms. It tracks specific brands, products, or topics within a user-defined timeframe.

TweetFeel

Focused on real-time Twitter data, TweetFeel employs machine learning-based sentiment analysis to classify tweets into positive and negative categories, offering quick insights into public sentiment.

Semantic Orientation through Gloss Classification

This research-oriented approach utilizes quantitative analysis of term definitions from online dictionaries for semi-supervised sentiment classification.

Adverb-Adjective Combinations (AACs)

This method measures sentiment strength by analyzing linguistic combinations of adverbs and adjectives. It provides a more nuanced understanding of sentiment expressions in text.

Our System's Uniqueness

Our project distinguishes itself by focusing exclusively on Reddit posts, a platform known for its diverse and context-rich discussions. In addition to sentiment analysis, our system provides deeper insights through trend analysis, forecasting, and sentiment-based profiling, tailored to Reddit's unique data landscape.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

III. SYSTEM REQUIREMENTS

Software Requirements

The development environment requires Windows 10 or later, macOS 10.13 or later, or Linux distributions with X11. The suggested development environments are PyCharm, Visual Studio Code (VS Code), or Jupyter Notebook / JupyterLab. The project will be developed using Python 3.7 or later. Libraries like pandas and numpy will be used to make it easier to manipulate and analyze the data. For machine learning tasks, it will use scikit-learn, TensorFlow, or PyTorch. NLP tasks will use NLTK, spaCy, and TextBlob. Matplotlib, seaborn, and Plotly will be used for data visualization. Web scraping will be done using BeautifulSoup, Scrapy, and Tweepy. Libraries for HTTP requests will be used, such as requests, and web frameworks like Flask or Django. Relational databases like SQLite or PostgreSQL will be used for data storage and management.

Hardware Requirements

Intel Core i3 or equivalent AMD processor is the minimum. An Intel Core i5/i7 or equivalent AMD processor is recommended for the best performance. The system should have at least 4 GB of RAM, while 8 GB or more is preferable. For graphics, integrated graphics (Intel HD Graphics 4000 or equivalent) is the minimum requirement, but a dedicated graphics card, such as the NVIDIA GeForce GTX 1050 or AMD Radeon RX 560 or higher, is recommended for better performance in data processing and model training. Storage requirements include at least 500 MB of free disk space for the development environment and dependencies, with an SSD being recommended for faster performance. The display resolution should be at least 1366x768, though a 1920x1080 resolution or higher is preferred. For peripherals, a display resolution of 1366x768 is the minimum, with 1920x1080 or higher being recommended.

IV. SYSTEM DESIGN

The design of the sentiment analysis system is aimed at collecting, preprocessing, analyzing, and deploying sentiment analysis models on social media data. The system runs in two main phases: Data Collection and Preprocessing, and Sentiment Analysis and System Deployment.

The following is the overall design of the flowchart for sentiment analysis of social media content.

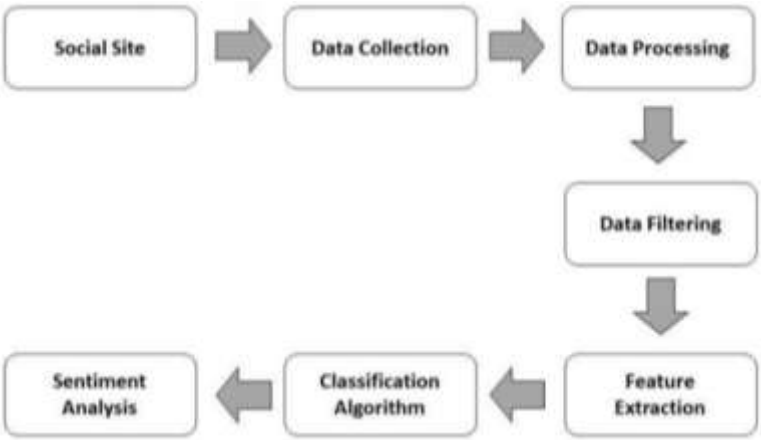

Fig 4.1: System Flowchart

Step 1: Data Collection

The system collects data from social media platforms like Twitter, Facebook, and Instagram. Data is gathered using APIs, such as the Twitter API, or through web scraping techniques using tools like BeautifulSoup and Scrapy. The raw data collected includes posts, comments, and tweets from these platforms.

Step 2: Data Cleaning

Once the data has been collected, cleaning is done on this data to remove noises such as special characters and URLs so that only the relevant textual data is retained. It uses the assistance of tools like regular expressions (Regex) and custom cleaning scripts in order to process this step to clean out text data.

Step 3: Tokenization

In the tokenization phase, the cleaned text is split into individual words or tokens. To achieve this, NLP tools

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

like NLTK or spaCy break the text down into more manageable units that can be used for further analysis.

Step 4: Normalization

Normalization includes all text being written in lower case and the stop words, which are words such as "the", "is", etc. that don't contribute much to sentiment analysis. This process is also performed with the help of tools such as NLTK and spaCy to make sure the text data is uniform and prepared for analysis.

V. SYSTEM DEPLOYMENT

Step 1: Sentiment Analysis

The process of sentiment analysis uses both lexicon-based approaches and machine learning algorithms to classify the sentiment of the data. Lexicon-based approaches use predefined sentiment dictionaries, such as VADER, to determine sentiment scores. Machine learning models, including supervised learning algorithms like Support Vector Machines (SVM) and Naive Bayes, are also used. More complex sentiment analysis tasks are performed using advanced deep learning architectures such as LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers). Tools such as scikit-learn, TensorFlow, and PyTorch are used to perform these analyses, which result in sentiment scores or labels such as positive, negative, or neutral.

Step 2: Model Deployment

Deploy these models for actual real-time usage after completing sentiment analysis. Scalability and response are ensured as the system has been designed through a microservices architecture. Tools used to deploy models as microservices include Flask, Django, Docker, and Kubernetes, which, in turn provide users with the real-time service of sentiment analysis.

Step 3: Compliance and Ethical Considerations

Measures are taken for compliance with the data privacy law, such as the General Data Protection Regulation, to ensure that the system will be used ethically. Besides, strategies against bias are followed to ensure a fair and non-biased operation of the sentiment analysis system, thus not only protecting data subjects but also businesses.

VI. IMPLEMENTATION AND TESTING

The two major components used for the implementation of sentiment analysis on Reddit comments were the Jupyter Notebook script and the Flask web application. The Jupyter Notebook was in charge of the entire process of sentiment analysis, which included fetching Reddit comments with the PRAW library, analyzing them with the VADER sentiment analysis tool, and training a logistic regression model to classify the sentiment.

The first step in the implementation was setting up the necessary libraries and resources. We imported libraries such as praw for interacting with Reddit, nltk for natural language processing, and sklearn for machine learning functionalities. The VADER lexicon and NLTK stopwords were downloaded to assist in sentiment analysis and text preprocessing. Reddit API credentials were configured using praw to fetch data from a specified post.

After fetching the data from Reddit, the VADER's SentimentIntensityAnalyzer tool was used for analyzing the comments' sentiment according to the polarity score. Predefined thresholds of these scores grouped them into three categories: 'Positive', 'Negative', and 'Neutral' sentiments. For preprocessing the comments, all comments were converted to their lowercase versions and visualizations are created for a better understanding of the distribution of sentiments. Matplotlib and seaborn libraries were used to plot the sentiment distribution, word count distribution, and scatter plots for sentiment versus word count.

Second, a logistic regression model was learned to classify comments as positive, negative, or neutral. To convert the comment text data to numerical format, TF-IDF vectorizer is used. Data set is then split into the training and the testing subsets for evaluation of model with precision, recall, and F1-score.

Testing the model was conducted by testing it on new, unseen Reddit comments to ensure that the trained logistic regression model would be able to classify sentiments. Further, the visualizations for sentiment analysis include pie charts, count plots, and word count distributions, while word clouds are generated for positive and negative comments to show words that are often used with an emotional tone.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

The Flask web application was created for a friendly interface to sentiment analysis. Users could simply input the URL of a Reddit post, and the application fetches comments, performs sentiment analysis, and provides visualizations in real time. The application further allows the users to download the results as a CSV file so that there is graphical and textual insight into the sentiment of the Reddit post.

The testing of the Flask application is performed for verifying its capability of passing different Reddit URLs, extracting valid post IDs, and making correct sentiment analysis and displaying the visualizations without any glitches. Error-handling mechanisms are put into place so that the app may quit gracefully in case of invalid input or any problems fetching data from Reddit.

In conclusion, the Jupyter Notebook and Flask application were successfully implemented to perform sentiment analysis on Reddit comments. The results were validated through extensive testing, with the logistic regression model showing high accuracy in classifying sentiment. The Flask application offered a dynamic and user-friendly platform for users to interact with the sentiment analysis system, making it a practical tool for analyzing social media sentiment.

VII. RESULTS AND SNAPSHOTS

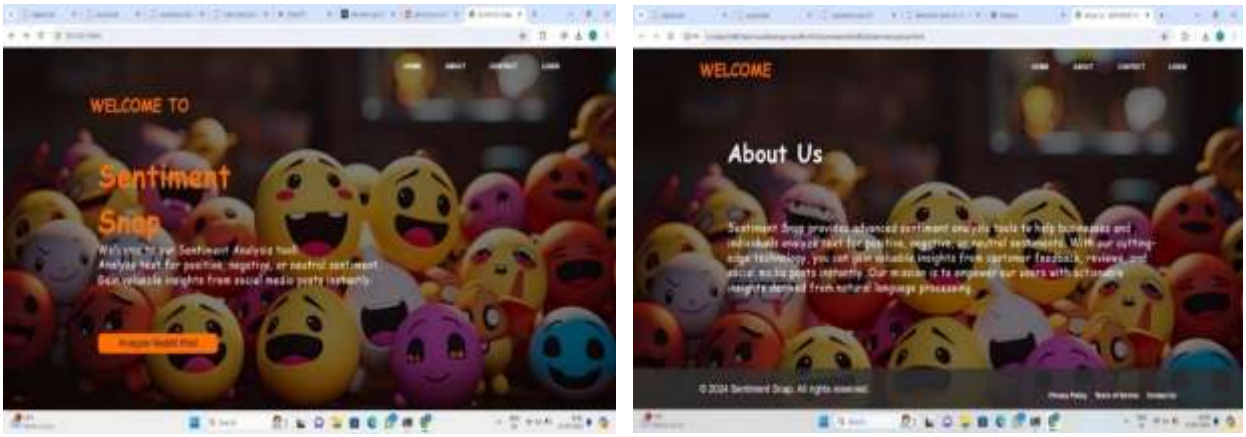

Fig 7.1: Home Page & About us Page

The homepage of the sentiment analysis tool is very intuitive because it will enable users to analyze the sentiment of text extracted from Reddit posts. Simply by inputting or selecting a post on Reddit, the input can easily clarify whether the post's sentiment is positive, negative, or neutral. Navigation from the homepage is also smooth, and users can always return to the analysis page for further interactions with other posts.

The about page is an introduction to the tool called Sentiment Snap, designed to offer enhanced sentiment analysis capabilities. Using one of the latest and greatest natural language processing technologies, Sentiment Snap assists users in understanding and interpreting the sentiment embedded in text data. The page further continues by elaborating on the mission of the company: to provide actionable insights to assist users in drawing meaningful conclusions from textual data, thereby enhancing decision-making processes across various domains.

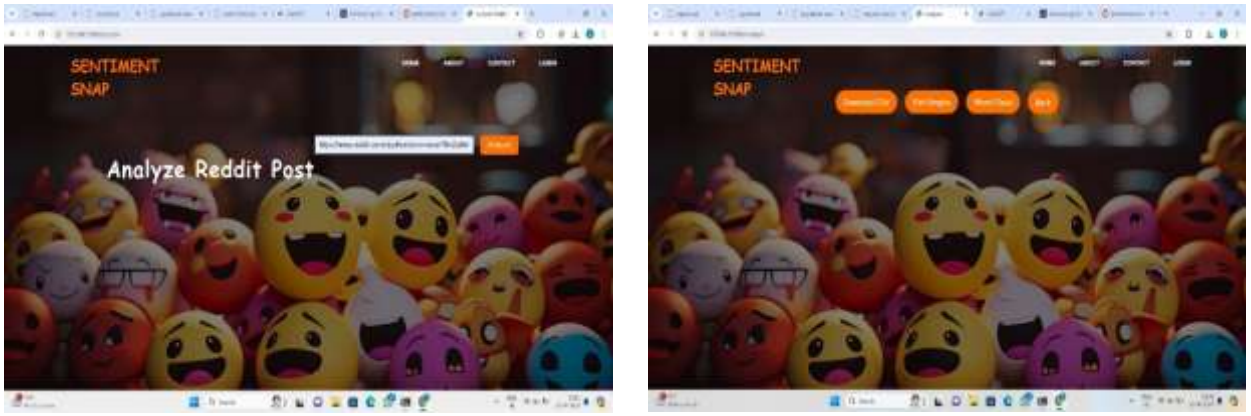

Fig 7.2: Url pasting Page & Analysis Page

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:01/January-2025 Impact Factor- 8.187 www.irjmets.com

The URL pasting file page of the sentiment analysis tool introduces the user to Sentiment Snap, a web application constructed to read out the sentiment of any Reddit posts. At this page, a user is requested to input the URL of any Reddit post they would want analyzed. Users then click the "Analyze" button after pasting their URL for the receipt of the analysis of the sentiment.

The analysis page provides a set of tools that assist in providing an in-depth exploration of the data being analyzed. It has export buttons for viewing the results as CSV, drawing visual graphs of the results, and displaying word clouds in order to access the sentiment data in different ways. These are integral parts of comprehensive sentiment analysis from the application itself, enabling a user to draw conclusions visually, as well as in detail, from the analysis.

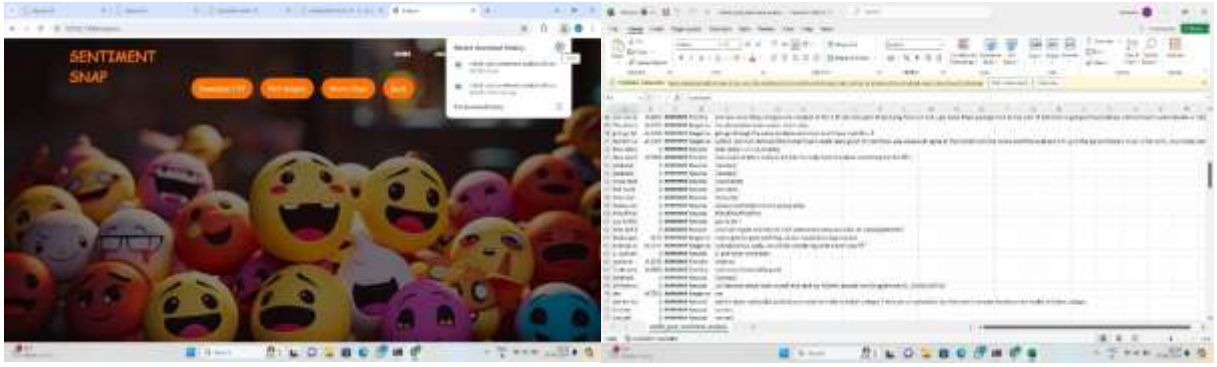

Fig 7.3: Downloading csv file page & Csv file

The downloading CSV file page contains a recent download history panel where users can view the sentiment analysis CSV files they have downloaded for Reddit posts. This panel gives an easy overview of the files, making it easier for users to track their past analyses and access them quickly when needed.

The CSV file opened in an Excel sheet provides a detailed view of the comments on the Reddit post along with their respective polarity scores. Each comment is analyzed and assigned a polarity value, which is then categorized as positive, negative, or neutral based on the sentiment expressed. This structure allows users to assess the sentiment of individual comments within the post and gain a more granular understanding of the overall sentiment distribution.

The sentimental analysis graph file comprises visual representations for the results obtained from the analysis of the post on Reddit. It contains the pie graph: a clear interpretation of the proportion of the entire sentiment distribution positive, negative, and neutral. And a bar graph is also exhibited, which gives comparative views of the various comments' strength of sentiment across different comments. The word count distribution further adds to the analysis by showing the frequency of key terms used within the post, reflecting the prominent topics and sentiments expressed.

The word cloud gives a visual impression of the positive and negative sentiments drawn from the Reddit post. It is a word cloud that shows which words are more frequently used, the larger the words, the more frequent they appear. It will distinguish between the positive and negative words, making it easy and intuitive for the users to grasp the key themes and sentiments expressed in the post.

Fig 7.4: Sentimental analysis graph & Wod colud

VIII. CONCLUSION

Sentiment analysis is a critical role in understanding customer opinions, especially in the business world where knowing customer thoughts on products, services, and brands can guide decision-making, trend analysis, and competitor identification. Sentiment Analysis for Social Media extracts sentiments from social media data, records them along with user information, and applies data mining techniques for product profiling and forecasting.

The challenge was to determine whether a sentence is positive, negative, or neutral, which was initially addressed using SentiWordNet, a lexical data source. However, this method faced issues with context, short terms, and ambiguous phrases like "not good" or "not bad." Moreover, the project had to account for multilingual comments, emotional symbols, slang, and noisy human language, complicating sentiment analysis but ultimately providing valuable insights for businesses.

IX. REFERENCES

[1] David Osimo and Francesco Mureddu, "Research challenge on Opinion Mining and Sentiment Analysis."

[2] Maura Conway, Lisa McInerney, Neil O'Hare, Alan F. Smeaton, Adam Bermingham, "Combining Social Network Analysis and Sentiment to Explore the Potential for Online Radicalisation," Centre for Sensor Web Technologies and School of Law and Government.

[3] Adobe® SocialAnalytics, powered by Omniture®.

[4] Brandwatch. [Online]. Available at: http://www.brandwatch.com/

[5] Sentiment140. [Online]. Available at: http://www.sentiment140.com

[6] Fabrizio S. Andrea E., "Determining the Semantic Orientation of Terms through," October 31–November 5, 2005.

[7] Carmine C., Diego R. Farah B., "Sentiment Analysis: Adjectives and Adverbs are better," ICWSM Boulder, CO USA, 2006.

[8] Lucas C., "Sentiment Analysis: A Multimodal Approach," Department of Computing, Imperial College London, September 2011.