Sentiment Analysis

A Major Project Synopsis Submitted to

Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal
Towards Partial Fulfillment for the Award of

Bachelor of Technology
(Computer Science and Engineering)

Under the Supervision of
Prof. Shraddha Sharma

Submitted By
Gourav Sharma (0827CS201083)
Harsh Dad (0827CS201089)
Harsh Jaiswal (0827CS201091)
Ishani Pandey (0827CS201103)

Department of Computer Science and Engineering
Acropolis Institute of Technology & Research, Indore
July-Dec 2023

1. Abstract

Analyze user-generated content, such as comments or posts, to determine the sentiment of the content. This can help social networks identify negative or harmful content and take action to remove it.

2. Introduction of the Project

Sentiment analysis, also known as opinion mining, is the process of extracting subjective information from text data and determining the sentiment expressed within it. This project aims to develop an efficient sentiment analysis model that can analyze large volumes of textual data and classify it into positive, negative, or neutral sentiments. The project utilizes advanced natural language processing (NLP) techniques and machine learning algorithms to achieve accurate sentiment classification.

3. Objective

- Develop a comprehensive dataset for sentiment analysis, encompassing diverse domains and sentiment categories.
- Preprocess the textual data by removing noise, handling punctuation, and converting it into a suitable format for analysis.
- Explore various feature extraction techniques to represent the text data effectively.
- Implement and train different machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and Recurrent Neural Networks (RNN), for sentiment classification.
- Evaluate the performance of the model's using metrics like accuracy, precision, recall, and F1-score.
- Fine-tune the models and optimize their hyperparameters to improve their performance.
- Deploy the best-performing sentiment analysis model as a web application for real-time sentiment analysis of user-generated text.

4. Scope

The scope of sentiment analysis includes social media monitoring, customer feedback analysis, market research, public opinion analysis, brand monitoring and reputation management, financial analysis, Voice of the Customer (VoC) analysis, product analytics, healthcare and patient feedback analysis, and fraud detection and risk assessment. It helps businesses understand public opinion, track brand sentiment, improve products and services, make data-driven decisions, manage brand reputation, predict market trends, enhance customer experience, analyze patient feedback, and detect fraudulent activities. The scope of sentiment analysis is vast and encompasses various domains and applications. Here are some key areas where sentiment analysis finds significant value:

- **Social Media Monitoring:** Sentiment analysis is extensively used to monitor and analyze sentiments expressed on social media platforms like Twitter, Facebook, and Instagram. It helps in understanding public opinion, tracking brand sentiment, identifying trends, and managing online reputation.
- **Voice of the Customer (VoC) Analysis**: Sentiment analysis is an integral part of Voice of the Customer programs, where it helps extract insights from customer feedback across multiple channels. It enables companies to understand customer sentiments, identify areas of improvement, and enhance customer experience

- **Fraud Detection and Risk Assessment**: Sentiment analysis can be used to identify potentially fraudulent activities and assess risks by analyzing text data related to insurance claims, financial transactions, or customer interactions. It helps in early detection of suspicious activities and mitigating risks.

5. Study of Existing System

- **Problems Addressed**

  IBM Watson Natural Language Understanding: IBM Watson offers the Natural Language Understanding (NLU) service, which provides sentiment analysis as one of its features. NLU utilizes machine learning models to analyze sentiment in text data. It is available as a cloud-based service and provides APIs for easy integration into applications.

- **Advantages**
  - Multilingual Support.
  - Cloud-Based Service.
  - Natural Language Understanding Capabilities.
- **Disadvantages**
  - Cost
  - Dependency on IBM Infrastructure
  - Language and Cultural Bias
- **Gaps Identified**
  - Contextual Understanding.
  - Domain-Specific Sentiment.
- **Reference link**
  - https://www.ibm.com/products/natural-language-understanding

- **Problems Addressed**

  Aylien: Aylien offers sentiment analysis software that provides text classification and sentiment analysis capabilities. It includes features like document-level sentiment analysis, entity-level sentiment, and aspectbased sentiment analysis. Aylien's sentiment analysis software can be used through their API or integrated into custom applications.

- **Advantages**
  - Industry-Specific Sentiment Analysis.
  - Emotion Analysis.
- **Disadvantages**
  - Language Support and Accuracy.
  - Interpretation of Neutral Sentiment.
- **Gaps Identified**
  - Handling of Complex Sentence Structures
- **Reference link**
  - https://aylien.com/product/news-api

- **Problems Addressed**

  Lexalytics: Lexalytics offers sentiment analysis software called Semantria, which provides advanced sentiment analysis capabilities. It includes features like entity recognition, theme extraction, and industryspecific sentiment analysis. Semantria can be deployed both in the cloud and on-premises.
- **Advantages**
  - Customization and Fine-Grained Analysis
  - specific domain or industry.
  - Robust Sentiment Analysis.
- **Disadvantages**
  - Performance and Scalability
  - Learning Curve for Integration
- **Gaps Identified**
  - Handling of Sarcasm and Irony.
  - Bias in Training Data
- **Reference link**
  - https://www.lexalytics.com/semantria/

6. Project Description

The project focuses on developing a sentiment analysis system that can accurately classify the sentiment expressed in textual data. Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that involves identifying and categorizing subjective information from text, determining whether the sentiment expressed is positive, negative, or neutral. The objective of this project is to build a robust sentiment analysis model that can handle large volumes of text data and provide valuable insights into sentiment patterns. The model will be trained using machine learning algorithms and advanced NLP techniques, leveraging annotated datasets for sentiment classification.

7. Methodology/Planning of the Project work

- **Data Collection**: Gather a diverse collection of textual data from various sources, such as social media platforms, product reviews, and news articles, to build a comprehensive sentiment analysis dataset.
- **Data Preprocessing**: Clean the collected data by removing irrelevant information, handling special characters, and converting text to lowercase. Apply techniques like tokenization, stop word removal, and stemming/lemmatization to further refine the data.
- **Feature Extraction**: Utilize techniques such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings (e.g., Word2Vec, GloVe) to represent the textual data as numerical features suitable for machine learning models.
- **Model Development**: Implement and train multiple machines learning models, including Naive Bayes, SVM, and RNNs, using the preprocessed data and extracted features. Tune the models' hyperparameters to achieve optimal performance.
- **Model Evaluation**: Assess the performance of the trained models using various evaluation metrics, comparing their accuracy, precision, recall, and F1-score. Perform cross-validation and statistical analysis to validate the results.

- **Model Optimization**: Fine-tune the best-performing models by adjusting hyperparameters, applying regularization techniques, and exploring ensemble methods to improve their sentiment classification accuracy.

8. Expected Outcome

- Development of a comprehensive sentiment analysis dataset suitable for training and evaluation purposes.
- Implementation and evaluation of various machine learning models for sentiment classification.
- Identification of the most effective feature extraction techniques for sentiment analysis.
- Optimization of the sentiment analysis models to achieve higher accuracy and performance.
- Deployment of a user-friendly web application for real-time sentiment analysis.

9. Resources and Limitations

Resources:

Textual Data: Sentiment analysis relies on a vast amount of textual data from diverse sources such as social media platforms, customer reviews, news articles, and surveys. Access to relevant and representative data is crucial for training accurate sentiment analysis models.

Machine Learning and NLP Techniques: Sentiment analysis utilizes various machine learning algorithms and natural language processing (NLP) techniques. Resources such as libraries, frameworks, and pre-trained models for tasks like tokenization, stemming, part-of-speech tagging, and sentiment lexicons are essential for developing effective sentiment analysis systems.

Limitations:

Ambiguity and Polysemy: Sentences or phrases with multiple meanings or ambiguous sentiments pose challenges for sentiment analysis. Contextual understanding and disambiguation are necessary to accurately interpret and classify sentiment in such cases.

Data Limitations: The availability of high-quality and labeled training data can be a limitation. Building large and diverse annotated datasets for sentiment analysis can be time-consuming and costly, impacting the performance and generalizability of sentiment analysis models.

10. Conclusion

Sentiment analysis plays a crucial role in understanding public opinion, customer feedback, and social media sentiment. This project aims to develop an efficient sentiment analysis system capable of accurately classifying textual data into positive, negative, or neutral sentiments. By leveraging advanced NLP techniques and machine learning algorithms, the project aims to provide valuable insights and actionable information to businesses, researchers, and decision-makers. The deployed web application will enable real-time sentiment analysis, empowering users to gauge public sentiment on various topics and make informed decisions based on the analyzed data.

11. References

[1] Priya Chakriswaran, Durai Raj Vincent, Kathiravan Srinivasan, Vishal Sharma, Chuan-Yu Chang, and Daniel Gutiérrez Reina. Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues. Applied Sciences, 9(24), 2019.

[2] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gul-cehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.

[3] Alexis Conneau, Holger Schwenk, Lo¨ıc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. 2016.

[4] Felipe G. Contratres, Solange N. Alves-Souza, Lucia Vilela Leite Filgueiras, and Luiz S. DeSouza. Sentiment analysis of social network data for cold-start relief in recommender systems. In Álvaro Rocha, Hojjat Adeli, Luís Paulo Reis, and Sandra Costanzo, editors, Trends and Advances in Information Systems and Technologies, pages 122–132, Cham, 2018. Springer International Publishing.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[6] effrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)s, page 1532–1543, 10 2014.