University of Hertfordshire
School of Engineering and Computer Science

MSc Computer Science (Data Science and Analytics)

Module: MSc CS Project

Detailed Project Proposal

Project title – A Study on Sentiment Analysis Techniques: Investigating Algorithms and Vectorization Methods

Sanjana Hombal
21054419
Level 7
Academic Year 2022-24

1.0 Project Introduction

Understanding and interpreting human emotions portrayed in textual data has become critical in an era where digital communication is the norm. The goal of sentiment analysis, a branch of natural language processing (NLP), is to computationally assess and determine the text's emotional content. Sentiment analysis is essential for a wide range of applications in many different industries, from evaluating customer feedback for businesses to monitoring public opinion on social media platforms.

Sentiment analysis algorithms have advanced, yet reliably and quickly extracting sentiment from text remains a difficult task. Automated sentiment analysis systems have substantial challenges due to the intricacy of human language, which encompasses nuances, sarcasm, and cultural background. Furthermore, the creation of reliable algorithms that can effectively handle large-scale analysis is required due to the enormous volume of textual data that is produced every day across numerous platforms.

The growing significance of sentiment analysis in the current digital environment served as inspiration for this study. As a keen observer of technology development and its effects on society, I have found it fascinating that sentiment analysis has the ability to glean insightful information from textual data.

The project aims to identify optimal combinations of machine learning algorithms and text vectorization techniques that yield the highest performance in sentiment analysis tasks. By systematically evaluating different algorithm-vectorization pairs, the goal is to determine which combinations are most effective for accurately classifying sentiment in textual data.

2.0 – Background

Jain, S.K. and Singh, P in [1] provide a systematic survey of sentiment analysis and opinion mining, focusing on classifying opinions into positive, neutral, and negative categories. It discusses the importance of sentiment analysis in decision-making and challenges like language limitations and dealing with sarcastic text. Various techniques include lexicon-based, machine learning-based, and hybrid approaches. Previous works cover methodologies like rule-based systems, sentiment classification on social media data, and sentiment analysis on regional languages. It also addresses issues like fake reviews and sentiment analysis performance.

Haisal A. D et al., in [2] found that TF-IDF feature vector representations generally outperform Word2vec and Doc2vec in book review sentiment classification. Combining TF-IDF with Word2vec led to improved sentiment classification, showing better performance across various evaluation metrics like classification accuracy, precision, recall, and F1-score. This combined approach of TF-IDF and Word2vec was recommended for future studies based on its effectiveness in representing word-level information combined with contextual information

Chaturvedi, S et al., [3] This paper focuses on sentiment analysis using machine learning techniques to classify reviews based on sentiments. Various algorithms like

Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and Random Forest were employed to predict sentiment accuracy. Among these classifiers, Support Vector Machine showed the highest accuracy of 94.08% in sentiment analysis of e-commerce reviews. The study aimed to automatically recognize and evaluate texts for monitoring public opinion and customer preferences. Sentiment analysis plays a crucial role in understanding and analysing communication through transcription, with machine learning techniques being popular due to their simplicity and accuracy in recent years.

Willianto, T et al., [4] focuses on sentiment analysis of e-commerce product reviews using TF-IDF and Backward Elimination. Data collection involves web scraping, data labelling, and pre-processing steps. Feature extraction with TF-IDF and Backward Elimination feature selection improve classification performance. The dataset is split for training and testing, with machine learning algorithms like SVM, Naive Bayes, Decision Tree, K-NN, and Random Forest used for classification. The best accuracy achieved is 85.97% with SVM and feature selection, showing improved performance across all classifiers used in the research

3.0 – Research Questions

How do different text vectorization techniques, such as Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) impact the performance of machine learning algorithms in sentiment analysis tasks?

How do different evaluation metrics, such as accuracy, precision, recall, and F1-score, vary across algorithm-vectorization combinations, and which combination provide the best sentiment analysis performance?

4.0 Project Aim

The project's main goal is to investigate combinations of algorithms and vectorizations in order to determine the best methods for sentiment analysis, the research compares and assesses several combinations of text vectorization and machine learning algorithms.

Comparing various text vectorization methods and how it impacts the machine learning algorithms' ability to execute sentiment analysis tasks. This entails investigating and evaluating the efficacy of different vectorization techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), and Bag-of-Words (BoW), in extracting sentiment information from textual data.

The project aims to contribute to the broader body of knowledge in sentiment analysis by advancing understanding of the relationship between text vectorization techniques, machine learning algorithms, and sentiment classification performance. By conducting rigorous experiments and analysis, the objective is to generate valuable insights.

5.0 – Description of Idea

Drawing upon a systematic survey on sentiment analysis, the project aims to leverage machine learning algorithms, specifically Support Vector Machine (SVM) and Naive Bayes, in conjunction with text vectorization techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW). The inspiration for this approach stems from a comprehensive review of existing literature, including studies such as "Sentiment Analysis on Business Data using Machine Learning" and "Sentiment Analysis on E-commerce Product Review using Machine Learning and Combination of TF-IDF and Backward Elimination."

Based on findings by Haisal A. D et al., the combination of TF-IDF with word2vec yielded superior performance in sentiment classification of book reviews. Additionally, as noted by Usha in their paper, the Bag of Words technique, which assigns fixed integer IDs to word occurrences, was utilized. Hence, TF-IDF and Bag of Words technique will be adopted for the vectorization methods due to its effectiveness, ability to represent text data in a format suitable for machine learning algorithms.

From the findings of paper [1] and [2] I decided to make use of supervised machine learning algorithms such as Support Vector Machines, Naïve Bayes as it allows for the classification of sentiments based on labelled data, providing better results for sentiment analysis. This approach involves training the model on a labelled training set and validating it on a test set. Classifiers like Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest, was evaluated for sentiment analysis of product reviews, with Support Vector Machine showing the highest accuracy of 93.93%. The Naïve Bayes had accuracy of 93.4%, algorithm performed well in classifying sentiments in the review’s dataset, demonstrating its effectiveness in sentiment analysis tasks

6.0 – Project Methodology

I will concentrate on gathering textual data for this research from online forums and social media platforms because they frequently have a lot of user-generated content that is good for sentiment analysis. I will specifically target social media sites like Reddit, Twitter, and online forums that discuss relevant topics (such product reviews and movie discussions).

I will do extensive research using academic repositories, data repositories like Kaggle and UCI Machine Learning Repository, and pertinent research publications in the field of sentiment analysis to find appropriate datasets for sentiment analysis. I'll look for datasets that satisfy the following requirements: -

1. Big enough to offer adequate information for assessment and training.
2. Labelled or annotated with sentiment labels (neutral, negative, and positive).
3. Diverse in content and representative of different domains or topics.

Phases of the investigative work:

- I will design the architecture of the sentiment analysis system, outlining the components and their interactions.
- The system will include modules for data collection, preprocessing, text vectorization, algorithm implementation, parameter tuning, evaluation, and result analysis.
- Implement machine learning algorithms for sentiment analysis, including Support Vector Machine (SVM) and Naive Bayes classifiers.
- Configure the algorithms with appropriate parameters and settings, considering factors such as kernel functions for SVM and smoothing parameters for Naive Bayes
- Experiment with different vectorization methods, including TF-IDF and Bag-of-Words, to assess their impact on sentiment analysis performance.
- I will be adjusting the algorithm parameters as necessary, leveraging parameter tuning techniques like grid search or random search.
- Evaluate the trained models on the test set to assess their generalization performance.
- Measure metrics such as accuracy, precision, recall, and F1-score to quantify the performance of each algorithm. To obtain highest possible levels of accuracy for the two selected algorithms (Support Vector Machines, Naïve Bayes) with the purpose of comparing them.
- Compare the performance of different algorithms and text vectorization techniques to identify the most effective combinations for sentiment analysis.
- Compare the results of this project with the state-of-the-art research that has been done in this field.

7.0 – Bibliography

[1]. Jain, S.K. and Singh, P. (2018) ‘Systematic survey on sentiment analysis’, 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) [Preprint]. doi:10.1109/icsccc.2018.8703370.

[2]. Abubakar, H.D. and Umar, M. (2022) ‘Sentiment classification: Review of text vectorization methods: Bag of words, TF-IDF, word2vec and doc2vec’, SLU Journal of Science and Technology, 4(1 & 2), pp. 27–33. doi:10.56471/slujst.v4i.266.

[3]. Chaturvedi, S., Mishra, V. and Mishra, N. (2017) ‘Sentiment analysis using machine learning for Business Intelligence’, 2017 IEEE International Conference on

Power, Control, Signals and Instrumentation Engineering (ICPCSI)[Preprint]. doi:10.1109/icpcsi.2017.8392100.

[4]. Willianto, T. and Wibowo, A. (2020) 'Sentiment analysis on e-commerce product using machine learning and combination of TF-IDF and backward elimination', International Journal of Recent Technology and Engineering (IJRTE), 8(6), pp. 2862–2867. doi:10.35940/ijrte.f7889.038620.

[5]. Maharani, W. (2013) 'Microblogging sentiment analysis with lexical based and machine learning approaches', 2013 International Conference of Information and Communication Technology (ICoICT) [Preprint]. doi:10.1109/icoict.2013.6574616.

[6]. Zheng, J., Zheng, L. and Yang, L. (2019) 'Research and analysis in fine-grained sentiment of film reviews based on Deep Learning', Journal of Physics: Conference Series, 1237(2), p. 022152. doi:10.1088/1742-6596/1237/2/022152.

[7]. Woldemariam, Y. (2016) 'Sentiment Analysis in a cross-media analysis framework', 2016 IEEE International Conference on Big Data Analysis (ICBDA) [Preprint]. doi:10.1109/icbda.2016.7509790.

[8]. Sentiment Analysis, data vectorization - Dev, W. (2021) Sentiment analysis, Data Vectorization, Medium. Available at: https://medium.com/analytics-vidhya/sentiment-analysis-data-vectorization-52d3e711f054 (Accessed: 01 March 2024).

[9]. Intuitive Guide to Understanding GloVe Embeddings - Ganegedara, T. (2022) Light on math ML: Intuitive Guide to Understanding Glove embeddings, Medium. Available at: https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010 (Accessed: 28 February 2024).

[10]. A Step-by-Step Tutorial for Conducting Sentiment Analysis - Zijing Zhu, P. (2021) A step-by-step tutorial for conducting sentiment analysis, Medium. Available at: https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-a7190a444366 (Accessed: 25 February 2024).

MSc project

A Study on Sentiment Analysis Techniques: Investigating Algorithms and Vectorization Methods

Project start: Mon, 1-22-2024

Display week: 1

TASK	DEPENDENCIES	PROGRESS	START	END
Literature Review				
Define research objectives		100%	2-4-24	2-8-24
Search for relevent literature		90%	2-8-24	2-10-24
Review and sumarize		90%	2-11-24	2-17-24
Write Literature Review		50%	2-18-24	2-21-24
Dataset Acquisition				
Identify relevant datasets	Literature review	100%	2-22-24	2-24-24
Obtain access to datasets	Literature review	100%	2-25-24	2-29-24
Download datasets	Literature review	90%	3-1-24	3-4-24
Data Preprocessing				
Handle special characters	Dataset Acquistion	25%	3-4-24	3-5-24
Tokenize text	Dataset Acquistion	0%	3-6-24	3-8-24
Remove stop words	Dataset Acquistion	0%	3-9-24	3-10-24
Lowercase text	Dataset Acquistion	0%	3-11-24	3-12-24
Algorithm Implementation				
Implement SVM algorithm	Data preprocessing	0%	3-13-24	3-17-24
Implement Naive bayes algorithm	Data preprocessing	0%	3-18-24	3-22-24
Configure algorithm parameters	Data preprocessing	0%	3-23-24	3-27-24
Text Vectorization				
Apply TF-IDF vectorization	Algorithm Implementaion	0%	3-28-24	3-31-24
Apply Bag-of-Words vectorization	Algorithm Implementaion	0%	4-1-24	4-6-24
Model Training and Evaluation				
Train SVM Model	Text Vectorization	0%	4-7-24	4-13-24
Train Naive Bayes Model	Text Vectorization	0%	4-14-24	4-20-24
Evaluate Model Performance	Text Vectorization	0%	4-21-24	4-21-24
Parameter Tuning and Optimization				
Perform Grid Search for SVM	Model Training and Evaluation	0%	4-22-24	4-26-24
Perform Grid Search for Naive Bayes	Model Training and Evaluation	0%	4-27-24	5-1-24
Comparitive Analysis				
Compare SVM and Naive Bayes performance	Parameter Tuning and Optimization	0%	5-2-24	5-5-24
Analyse Results	Parameter Tuning and Optimization	0%	5-6-24	5-10-24
Documentaion and Reporting	Comparitive Analysis	0%	5-11-24	5-18-24