PROJECT PROPOSAL

Machine Learning & Data Mining 20222

ARTIFICIAL NEURAL NETWORK FOR SENTIMENT ANALYSIS

GROUP 10

Mai Anh Đức

Nguyễn Bá Hải

Đinh Ngọc Khuê

Trần Phúc Thành

Nguyễn Văn Trình

CONTENTS

I. INTRODUCTION

Sentiment analysis is a technique used to determine the emotional tone of a piece of writing. It is often used to analyze product reviews. Artificial neural networks (ANNs) are a type of machine learning algorithm that can be used for sentiment analysis. ANNs are modeled after the structure of the human brain and can be trained to recognize patterns in data.

A. Background and Motivation

- Online product reviews play a crucial role in influencing consumer purchasing decisions.
- Sentiment analysis of product reviews using artificial neural networks is that it can help companies understand how their products are being received by customers.
- By analyzing product reviews, companies can identify areas where their products are doing well and areas where they need improvement.

B. Problem Statement and Objectives

- The goal of this project is to develop an Artificial Neural Network (ANN) model for sentiment analysis of product reviews.
- The model will classify reviews as positive, negative, or neutral, enabling businesses to understand customer sentiments.

C. Significance of Sentiment Analysis

- Businesses can utilize sentiment analysis to improve product development, marketing strategies, and customer satisfaction.
- It can help companies improve their products and services by identifying areas where they need improvement and also help to identify trends in customer sentiment over time and make adjustments to their products and services accordingly.

II. LITERATURE REVIEW

A. Overview of Sentiment Analysis:

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that involves extracting and analyzing subjective information from text. It aims to determine the sentiment or

emotional tone expressed in a piece of text, whether it is positive, negative, or neutral.

Sentiment analysis has various applications across different domains, including:

- Social Media Analysis: Sentiment analysis is widely used to analyze social media data, such as tweets, comments, and reviews, to understand public opinion, track brand sentiment, or identify trends.
- Customer Feedback Analysis: Sentiment analysis helps businesses analyze customer reviews, surveys, and feedback to gain insights into customer satisfaction, identify areas for improvement, and make data-driven decisions.
- Market Research: Sentiment analysis can be used to analyze market trends, consumer preferences, and reactions to products or services, enabling companies to understand the market landscape and make informed marketing strategies.
- Reputation Management: Sentiment analysis assists in monitoring and managing the online reputation of individuals, companies, or brands by identifying positive or negative sentiments in online content.

B. Existing Approaches and Techniques:

Traditionally, sentiment analysis has been approached using various machine learning techniques, including:

- Rule-Based Methods: These methods involve defining a set of rules or patterns to identify sentiment-bearing words or phrases and assign sentiment labels based on predefined rules. However, they often lack the flexibility to capture complex patterns and require manual rule creation.
- Machine Learning Methods: These methods use labeled training data to train a machine learning model, which can then classify sentiment in new, unseen text. Traditional machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees have been commonly used for sentiment analysis.

C. Role of Artificial Neural Networks (ANN) in Sentiment Analysis:

Artificial Neural Networks (ANN) have gained significant attention and have become a popular choice for sentiment analysis due to their ability to capture complex patterns in textual data. ANN models, such as Recurrent

Neural Networks (RNN) or more advanced models like Long Short-Term Memory (LSTM) or Transformer-based architectures, can effectively process sequential data and learn contextual dependencies.

ANNs have the advantage of automatically learning feature representations from the data, which eliminates the need for manual feature engineering. They can capture the nuances of sentiment in text, including contextual information, word order, and dependencies between words, leading to improved sentiment classification performance.

D. Limitations and Challenges in Sentiment Analysis:

Sentiment analysis still faces several challenges and limitations, including:

- Handling Negation and Sarcasm: Negation and sarcasm can significantly affect the sentiment conveyed in text. Understanding negation cues and sarcasm detection are ongoing research challenges in sentiment analysis.

- Context-Dependent Sentiments: Sentiment analysis needs to consider the context in which the text is expressed, as the sentiment can change based on the topic, domain, or cultural factors. Contextual understanding is crucial for accurate sentiment classification.

- Addressing Class Imbalance: Imbalanced datasets, where one sentiment class dominates over others, can bias the model towards the majority class. Techniques like oversampling, undersampling, or using advanced loss functions can help address class imbalance issues.

- Improving Model Generalization: Sentiment analysis models should generalize well to unseen data and different domains. Transfer learning, domain adaptation, or utilizing pre-trained language models can improve model generalization and performance.

III. METHODOLOGY

A. Dataset Selection and Preprocessing

1. Dataset Selection

- Choose a publicly available product review dataset, such as the Shopee platform Customer Reviews dataset, that is relevant to the project objectives and covers a diverse range of products.

2. *Dataset Preprocessing*

- Perform data preprocessing to ensure the dataset is clean and suitable for analysis.
- Remove any irrelevant information or noise from the dataset, such as HTML tags, special characters, or punctuation marks.
- Tokenize the text data by splitting it into individual words or subword units for further processing.
- Normalize the text by converting it to lowercase, removing stopwords, and applying stemming or lemmatization techniques if necessary.

B. Feature Extraction and Representation

1. *Feature Extraction Techniques*

- Explore different feature extraction techniques to represent the preprocessed text data effectively.
- Consider techniques such as the bag-of-words model, TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings like Word2Vec or GloVe.
- Evaluate the advantages and limitations of each technique based on the project requirements and the characteristics of the dataset.

2. *Numerical Representation*

- Convert the preprocessed text data into numerical representations suitable for input to the ANN model.
- For the bag-of-words model, create a matrix where each row corresponds to a review and each column represents a unique word, with the matrix elements indicating word frequencies or TF-IDF scores.
- For word embeddings, map each word or subword unit to a high-dimensional vector representation to capture semantic relationships.

C. Designing Artificial Neural Network Architecture

1. *ANN Architecture Design*

- Develop a feed-forward ANN architecture that is capable of effectively addressing the sentiment classification task.
- Design the architecture with an input layer, one or more hidden layers, and an output layer.
- Experiment with different configurations of the hidden layers, considering the depth (number of layers) and width (number of nodes) to find the optimal

architecture for the given dataset.

2. *Activation Functions and Layer Configurations*

- Select appropriate activation functions for the hidden layers, such as ReLU (Rectified Linear Unit), sigmoid, or tanh, based on the nature of the problem and the activation function's ability to capture non-linear relationships.
- Choose a suitable activation function for the output layer, such as softmax, to produce probability distributions over the sentiment classes.

D. Model Training and Evaluation

1. *Dataset Split*

- Split the preprocessed dataset into training and testing sets, maintaining the class distribution to account for potential class imbalance.

2. *Model Training*

- Train the ANN model using backpropagation and gradient descent optimization techniques.
- Define a suitable loss function, such as categorical cross-entropy, to measure the model's performance during training.

3. *Hyperparameter Tuning*

- Fine-tune the hyperparameters of the model, including learning rate, batch size, and number of epochs, to achieve optimal performance.

4. *Model Evaluation*

- Evaluate the trained ANN model using various metrics, such as accuracy, precision, recall, and F1-score, to assess its performance on the testing set.
- Analyze the model's strengths and weaknesses, and consider potential improvements or adjustments to enhance its performance.

IV. EXPECTED RESULTS AND DISCUSSION

A. Expected Results

1. *Improving Accuracy:*

- ANN-based models have shown promising results in sentiment analysis tasks, often achieving high accuracy rates. With proper training and tuning, you can expect the ANN model to accurately classify sentiment in text data.

2. *Learning Complex Patterns:*

- ANNs have the ability to learn complex patterns from the data, including subtle linguistic cues and context. This enables the model to capture more nuanced sentiment information and improve the overall accuracy of sentiment analysis.

3. *Generalization:*

- An effective ANN model should be able to generalize well to unseen data, meaning it can accurately classify sentiments in new and previously unseen text samples. This allows the model to be applicable in the real-world without training a dataset.

B. Discussion

1. *Neural Network Architecture:*

- The choice of neural network architecture, such as LSTM, GRU, or CNN, can impact the performance of the sentiment analysis model.
- Experimentation with different architectures and hyperparameters can help determine the optimal configuration for your specific dataset and sentiment analysis task.

2. *Data Preprocessing:*

- Proper preprocessing of the text data is crucial for sentiment analysis.
- Techniques like removing stop words, handling spelling variations, and dealing with negations can improve the model's performance. It is important to consider the specific characteristics of the dataset and adjust preprocessing steps accordingly.

3. *Training Data Size:*

- The size and quality of the training dataset play a significant role in the performance of the ANN model.
- A larger and diverse training dataset can help the model learn a wide range of sentiment patterns and improve its generalization capability.

V. PROJECT TIMELINE

A. Week 1:

1. *Dataset Acquisition and Preprocessing*

- Identify a suitable product review dataset.
- Download and preprocess the dataset by removing noise, tokenizing, and normalizing the text.

2. *Feature Extraction and Representation*

- Explore feature extraction techniques like bag-of-words or word embeddings.
- Convert the preprocessed text data into numerical representations suitable for ANN input.

B. Week 2:

1. *ANN Architecture Design and Implementation*

- Design a simple feed-forward ANN architecture with an input layer, one or more hidden layers, and an output layer.
- Experiment with different activation functions and layer configurations to find an optimal setup.

2. *Model Training and Evaluation*

- Split the dataset into training and testing sets.
- Train the ANN model using backpropagation and gradient descent optimization.
- Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.

C. Week 3:

1. *Results Analysis and Comparison*

- Present and interpret the performance of the trained ANN model.
- Compare the model's performance with baseline models like Naive Bayes or Support Vector Machines.
- Discuss the strengths and weaknesses of the ANN model in sentiment analysis.

2. *Discussion on Model Performance and Limitations*

- Identify potential limitations of the proposed ANN model and discuss possible enhancements or mitigations.

D. Week 4:

1. *Report Writing and Finalization*

- Summarize the project objectives, methodology, and findings in a comprehensive report.
- Include an introduction, literature review, methodology, results, discussion, and conclusion sections.
- Revise and proofread the report to ensure clarity and coherence.

2. *Presentation Preparation*

- Create a presentation slide deck summarizing the project's key aspects,

findings, and implications.
- Practice presenting the project to effectively communicate the research process and outcomes.

VI. CONCLUSION

A. Summary of the Proposed Project:

The project aimed to develop an Artificial Neural Network (ANN) model for sentiment analysis of product reviews. By analyzing customer sentiments expressed in product reviews, businesses can gain valuable insights into customer opinions and attitudes, which can inform product development, marketing strategies, and overall customer satisfaction. The project involved the acquisition and preprocessing of a product review dataset, followed by feature extraction and representation. A simple feed-forward ANN architecture was designed and trained using backpropagation and gradient descent optimization. The model's performance was evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score. The project also included a comparative analysis with baseline models to assess the effectiveness of the ANN model in sentiment analysis.

B. Expected Outcomes and Deliverables:

The expected outcomes of the project include:

- A preprocessed product review dataset suitable for sentiment analysis.
- Numerical representations of the text data obtained through feature extraction techniques.
- A trained ANN model capable of classifying product reviews into positive, negative, or neutral sentiments.
- Performance evaluation metrics such as accuracy, precision, recall, and F1-score to assess the model's effectiveness.
- A comparative analysis of the ANN model with baseline models to understand its advantages and limitations in sentiment analysis.

VII. REFERENCES

1. Chen, S. (2019). *Sentiment Analysis with Neural Networks: A Beginner's Guide*.

Medium. [Online].
2. Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.
3. Li, H., He, J., & Wu, X. (2018). *Deep Learning for Sentiment Analysis: A Survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.
4. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning Word Vectors for Sentiment Analysis*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), 142-150.
5. Wang, S., & Manning, C. D. (2012). *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification*. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL-Short), 90-94.