DORTMUND UNIVERSITY OF APPLIED SCIENCES AND ARTS

FACULTY OF INFORMATIK

REPORT OF TRENDS OF ARTIFICIAL INTELLIGENCE IN BUSINESS INFORMATICS

SENTIMENT ANALYSIS

VAN DAO NGUYEN

Student ID: 7221965

QUYEN HO

Student ID: 7221978

Major: Master of Embedded Systems Engineering

Advisor: Prof. Dr. Sebastian Bab

Dortmund, 22nd January 2026

CONTENTS

I. Introduction

1. Overall

In this project, a sentiment analysis model is implemented to analyze the sentiments of the comments with data collected from the film trailers produced by Marvel. The sentiments consist of negative, neutral and positive classes. A pipeline of natural language preprocessing including tokenization and stop words is applied and there are two procedures including SVM machine learning model and lexicon-based analysis that are used to analyze sentiment. These two methods are used to predict the sentiment based on the dataset and the results of these methods will be compared to each other.

2. Support vector machine (SVM)

Support vector machine is a supervised machine learning algorithm that is used to classify object, sentiment, … and solve regression tasks. The objective of this algorithm is to find the best boundary separating different classes in dataset [4]. The boundary is also known as a hyperplane that is represented by the function (1):

$$wx + b = 0 \quad (1)$$

This is used for linear classification. W is the normal vector that is perpendicular to hyperplane (define the orientation of the hyperplane), b is the bias term representing how far the hyperplane is from the origin and x is the input vector. To obtain the best hyperplane, the algorithm must find the greatest margin, that is a distance from the nearest data point to hyperplane as shown in function (2). In order to increase distance margin, it is necessary to decrease ||w||.

$$d = \frac{wx + b}{||w||} \quad (2)$$

In addition, SVM can determine the hyperplane by polynomial function and radial basis function (>2D) when data is nonlinear and cannot be separated into classes by a straight line in 2D. Because the SVM uses hyperplane to classify objects, this machine learning algorithm is suitable for binary classification. However, more than 2 classes also can be separated by SVM because this algorithm contains two common strategies when dealing with multi-class classification such as one-vs-one (OvO) and one-vs-rest (OvR) shown in Fig. 1 and Fig. 2.

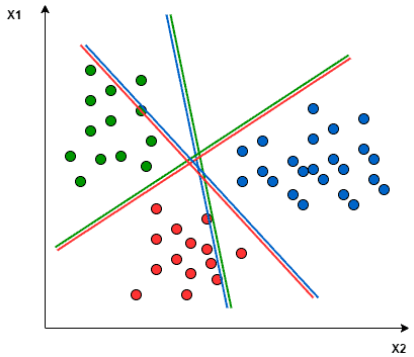

Fig. 1 One-vs-One Strategy [1]

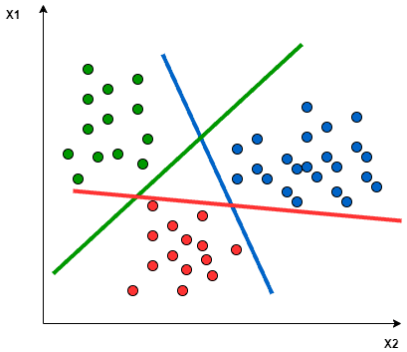

Fig. 2 One-vs-Rest Strategy [1]

For one-vs-one strategy [5], the binary SVM classifiers are applied to each pair of classes during the training phase. For example, if there are 3 classes of sentiments such as negative,

neutral and positive comment, three binary classifiers (negative-vs-neutral, negative-vs-positive and neutral-vs-positive) will be trained. Then, each binary classifier will vote for one class during prediction phase. For one-vs-rest strategy [5], a binary classifier will be trained for each class during training phase that means each class having it own binary classifier to distinguish the current class from the rest. During prediction phase, each classifier provides a score for the respective class.

3. Lexicon-based sentiment analysis

A lexicon model is one of the tools used in natural language processing (NLP). In contrast to the supervised SVM model, the lexicon-based approach is an unsupervised analysis method, as it does not require the training phase with labeled data to determine the sentiment of a text [2]. This technique uses a dictionary of lexical words which are associated with a particular score and computes the overall sentiment of a text by aggregating these scores.

In this project, the lexicon-based method is used to compare with the supervised SVM model and to evaluate the performance of machine learning for sentiment analysis of Marvel trailer comments.

II. Data

A dataset consisting of 1,003 user comments was collected from various Marvel trailers published on YouTube. The comments were gathered directly from the comment sections of official trailer videos. The dataset includes comments from trailers of the following Marvel movies:

- Iron Man 3 -- Official Trailer UK Marvel | HD
- Marvel Studios' Avengers: Infinity War - Official Trailer
- Marvel Studios’ Shang-Chi and the Legend of the Ten Rings | Official Trailer
- Marvel Studios' Thor: Love and Thunder | Official Trailer
- Marvel Studios' The Marvels | Official Trailer
- Thor: Ragnarok Teaser Trailer [HD] + "Thor: Ragnarok" Official Trailer
- The Fantastic Four: First Steps | Final Trailer
- Marvel Studios’ Ant-Man and The Wasp: Quantumania - Official Trailer
- Marvel Studios’ Eternals | Final Trailer
- Captain America: Brave New World | Official Trailer
- Marvel Studios' Doctor Strange in the Multiverse of Madness | Official Trailer
- Marvel Studios' Black Widow | Official Trailer
- Official Trailer | The Falcon and the Winter Soldier
- Marvel Studios' The Falcon and The Winter Soldier | Final Trailer

Each comment was manually labeled as positive, negative, or neutral. The dataset is used for both training and testing of the supervised SVM model, as well as for evaluating the lexicon-based sentiment analysis approach. Collection of comments also contains the balanced distribution of three classes to ensure a fair analysis for both methods.

III. Methodology

1. Overall pipeline

In this project, the main process as shown in Fig. 3 is followed. Firstly, data is collected from comments on Youtube. When raw data is available (text), the raw data will be preprocessed for supporting the training model to produce a good result with high accuracy. Because the input type of the machine learning model is vector, the step of vectorizing must be executed after preprocessing step.

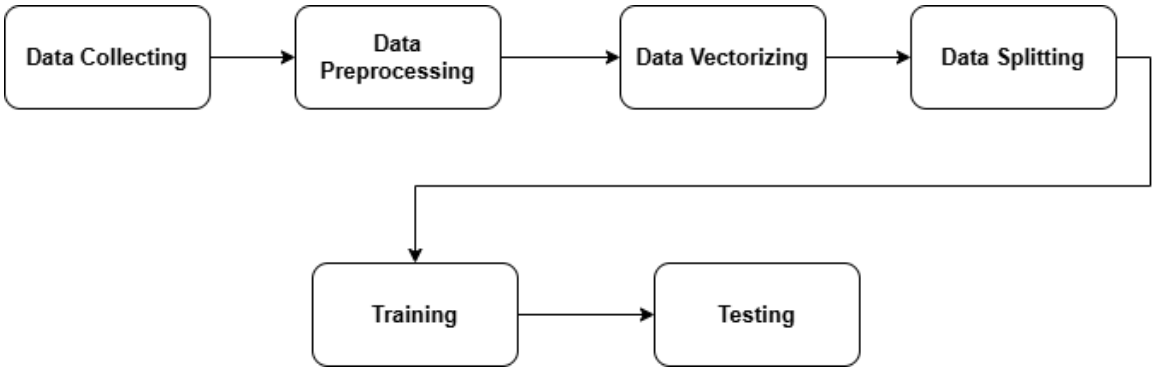

Fig. 3 Main Pipeline

Before training model with the processed data, the data is separated into two parts because the model requires a part of data for training and a part of data for testing. When everything required is ready, the model is trained and tested in the training and testing phase.

2. Data processing and training pipeline

The Fig. 4 illustrates the process of data preprocessing, and which method is used for processing the data.

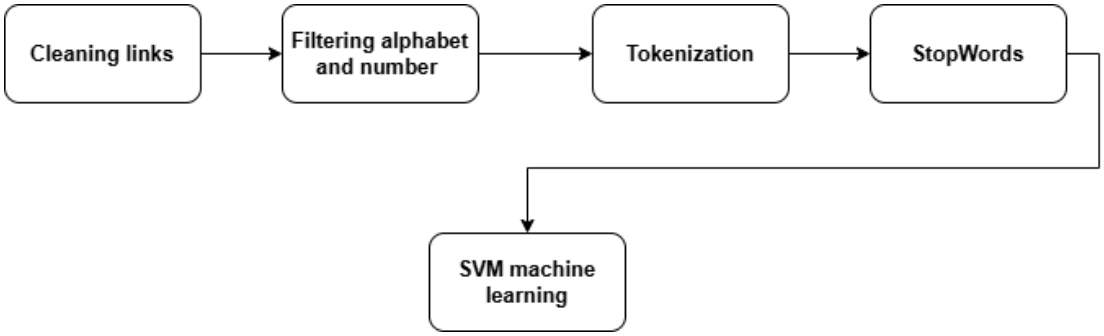

Fig. 4 Detailed Pipeline

Firstly, the text comments are checked with the link filter. This filter will remove links contained in the text comments. Then, a filter for checking alphabetic characters and numbers is created, this filter will delete all characters that are not number and alphabetic character. Next, the method of tokenization is used to tokenize words in a text comment. These tokens will then be checked by the method of stop words. This method will compare the tokens with a defined dictionary that contains unnecessary words, if the tokens match the words in the dictionary, these tokens will be removed from the text comment.

3. Sentiment analysis with VADER lexicon

The Valence Aware Dictionary and Sentiment Reasoner (VADER) is chosen for lexicon-based approach. VADER is a lexicon and rule-based model for sentiment analysis and specifically designed for microblog-like contexts such as social media text, movie reviews, and product reviews [3]. Hence, VADER is suitable for analyzing short and informal text, such as YouTube comments. Marvel trailer comments often contain slang, emojis, capitalization for emphasis, and informal expressions, which are handled by VADER's rule-based. VADER is freely available and has been used in sentiment analysis research due to its robustness and strong performance on social media datasets.

VADER contains over 9000 lexical features which can be a word or an emotion. Each feature is rated from –4 (negative) to 4 (positive) with 0 for neutral, and VADER then calculates the overall score for a text that is examined.

The process of lexicon approach is like the SVM approach, and it has even fewer steps (Fig. 5). The analysis is performed immediately after data preprocessing, which is the same as described above, and the result of lexicon-based model can be evaluated afterwards.

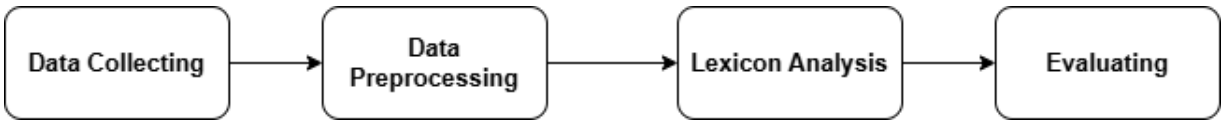

Fig. 5 Process for lexicon analysis approach

IV. Experiments

1. Support vector machine (SVM)

Because the input type of the SVM is vectors and the data type used in this project is text (comments and sentiments), the data must be converted to vector or digit value by using vectorizing before fetching into SVM for training. Firstly, the target labels must be converted to digit.

```
positives = df[df["sentiment"] == "positive"] # group all the positive
comments
negatives = df[df["sentiment"] == "negative"] # group all the negative
comments
neutrals = df[df["sentiment"] == "neutral"]   # group all the neutral comments
negatives["sentiment"] = -1 # convert text label to digit lable -1 for
negative
positives["sentiment"] = 1  # convert text label to digit lable 1 for positive
neutrals["sentiment"] = 0   # convert text label to digit lable 0 for neutral
```

There are 3 classes in this project including negative, neutral and positive. The codes above show the converting of sentiment from negative to -1, neutral to 0 and positive to 1. Secondly, the reviews are converted to digit by using TfidfVectorizer().

```
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(preprocessed_text)
y = np.asarray(data["sentiment"]) # label
```

After the input is ready for training, the dataset is divided into training data and testing data by using train_test_split(). In this project, the parameter test_size of train_test_split() is set to 0.2 for dividing 80 percent of datasets used in training and 20 percent of datasets used in testing. Because this project is a multi-class classification project and the number of classes is relatively small (3 classes), one-vs-one strategy of SVM is applied for training and creating hyperplanes by setting the parameter decision_function_shape = 'ovo' in the API function SVC(decision_function_shape='ovo').fit(X_train, y_train).

2. Lexicon-based analysis

After preprocessing, each comment is analyzed using the VADER sentiment analyzer. For each comment, VADER outputs four polarity scores: positive, negative, neutral, and a compound score. The compound score, which is from −1 to +1, indicates the overall sentiment of the text. Based on a threshold value, comments are classified into negative, neutral, or positive. In this experiment, a comment is positive if the compound score is above 0.1, negative if it is below −0.1, and neutral otherwise. The code below shows exactly the condition for classification, threshold is defined as 0.

```
# analyze the sentiment for review
analyzer = SentimentIntensityAnalyzer()
scores = analyzer.polarity_scores(review)
# get aggregate scores and final sentiment
agg_score = scores['compound']
final_sentiment = 'neutral'
if agg_score > threshold+0.1:
    final_sentiment = 'positive'
elif agg_score < threshold-0.1:
    final_sentiment = 'negative'
```

V. Results

1. Support vector machine (SVM)

The performance of SVM is evaluated on a test set consisting of 201 comments. The evaluation metrics are precision, recall, F1-score and overall accuracy. As shown in Fig. 6, the SVM model achieves an accuracy of 72%, which shows strong performance for classifying sentiments on Marvel trailer comments.

	precision	recall	f1-score	support
-1	0.85	0.72	0.78	72
0	0.57	0.68	0.62	63
1	0.75	0.74	0.75	66
accuracy			0.72	201
macro avg	0.73	0.72	0.72	201
weighted avg	0.73	0.72	0.72	201

Fig. 6 Classification report of support vector machine model

The negative (-1) class has a precision of 0.85, a recall of 0.72, and F1-score of 0.78. This result indicates an effective classification of negative comments.

The detection of positive (1) comments is also reliable with a balanced performance, 0.75 precision, 0.74 recall, and 0.75 F1-score.

However, neutral (0) comments are not classified as well as negative and positive comments with precision, recall, and F1-score below 0.7.

In general, the SVM model demonstrates robust and balanced performance across all three sentiment classes.

2. Lexicon-Based Sentiment Analysis

The lexicon-based sentiment analysis using the VADER model is applied to the full dataset of 1,003 comments and its performance is shown in Fig. 7. The model achieves an overall accuracy of 55% demonstrating moderate performance, which is lower than the SVM approach.

	precision	recall	f1-score	support
negative	0.62	0.57	0.60	341
neutral	0.51	0.29	0.37	332
positive	0.52	0.78	0.63	330
accuracy			0.55	1003
macro avg	0.55	0.55	0.53	1003
weighted avg	0.55	0.55	0.53	1003

Fig. 7 Classification report of lexicon-based VADER model

The negative class achieves a precision of 0.62, recall of 0.57, and an F1-score of 0.60, showing that negative comments are moderately identified using lexicon-based rules.

The positive class achieves a high recall of 0.78, which shows that most positive comments are successfully detected. However, the lower precision of 0.52 indicates that some non-positive comments are wrongly classified as positive.

The neutral class has the weakest performance, with a recall of only 0.28 and an F1-score of 0.37, indicating difficulty in distinguishing neutral comments from positive or negative ones.

3. Discussion

The comparison between the supervised SVM model and the lexicon-based approach highlights the advantages of machine learning for sentiment analysis. The supervised SVM model outperforms the lexicon-based method which shows an improvement in overall accuracy and evaluation metrics, particularly in neutral comments.

While the lexicon-based approach achieves moderate accuracy without training, the SVM model improves classification performance for all sentiment categories through the training process with labeled data.

Two models are evaluated on different dataset sizes, as the SVM model is tested on a subset of the labeled data, whereas the lexicon-based approach is applied to the full dataset. Despite that, the results clearly demonstrate that supervised learning is more effective for sentiment analysis of Marvel trailer comments.

VI. Conclusion

In conclusion, the two methods of sentiment analyzing such as SVM model and lexicon-based sentiment analysis can afford to decide which sentiment is conveyed by the input comments. However, the accuracies of SVM model and lexicon-based method are just 72% and 55% respectively, which are not high enough for applications in reality. To achieve maximum performance, it is essential to collect more data with high quality, build a strong pipeline for preprocessing data in the future work.

REFERENCES

[1] baeldung, “Multiclass Classification Using Support Vector Machines,” https://www.baeldung.com/. Feb 28th 2025. Accessed: Jan 6th 2026. [Online]. Available: https://www.baeldung.com/cs/svm-multiclass-classification

[2] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.

[3] C. Hutto and E. Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”, ICWSM, vol. 8, no. 1, pp. 216-225, May 2014.

[4] geeksforgeeks, “Support Vector Machine (SVM) Algorithm,” https://www.geeksforgeeks.org/. Accessed: Jan 6th 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

[5] geeksforgeeks, “Multi-class classification using Support Vector Machines (SVM),” https://www.geeksforgeeks.org/. Jul 23rd 2025. Accessed: Jan 6th 2026. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/multi-class-classification-using-support-vector-machines-svm/