Research and Development of an Application for Sentiment Analysis of Social Media Comments

NGUYỄN NGỌC HẢI
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student493@ictu.edu.vn

***Abstract*—This thesis studies sentiment classification of Vietnamese social-media comments. It describes collecting and preprocessing text, assigning labels, extracting language features, and comparing machine-learning and deep-learning classifiers. The reported comparison includes SVM, CNN, LSTM, and a combined CNN-LSTM model. The source table gives accuracy values of 82.2%, 84.9%, 84.4%, and 84.6% respectively, making CNN the highest in that comparison. The report also illustrates positive, neutral, and negative comment predictions and notes that the sentiment-classification model still needs optimization. The work is a documented experiment on text comments, without evidence of a deployed social-media service.**

***Index Terms*—sentiment analysis, machine learning, deep learning, natural language processing, social media, recurrent neural network**

I. Introduction

Comments on social networks contain opinions expressed in informal Vietnamese, including abbreviations and spelling variation. The thesis frames sentiment classification as a natural-language-processing task that maps such text into sentiment categories. It reviews machine learning, deep learning, and recurrent neural networks before developing an applied comparison. The study aims to recognize emotional orientation in user comments and to examine how several model families perform on the data described in the report [1]. Earlier ICTU student work on related designs is included here for context [5]. Related published studies provide context for the subject and methods [6], [7], [8], [9], [10]. These contextual works were selected for this English summary; they are not asserted to appear in the original thesis bibliography.

II. Methods

The documented pipeline collects social-media comments, cleans special characters and other noisy text, normalizes and labels the material, and extracts language features. The report then trains or compares SVM, CNN, LSTM, and CNN-LSTM classifiers for sentiment prediction. Its result table reports precision, recall, F1 score, and accuracy for each method. Examples of processed Vietnamese comments and predicted sentiment labels illustrate the applied classification. The summary relies on the source's reported comparison rather than inferring unreported software libraries or deployment details [1].

Supervisor: Nguyễn Tuấn Anh, Thai Nguyen University of Information and Communication Technology (anhnt@ictu.edu.vn).

III. Results

The source reports accuracy of 82.2% for SVM, 84.9% for CNN, 84.4% for LSTM, and 84.6% for CNN-LSTM. CNN is the highest on the reported accuracy measure. The table also gives precision, recall, and F1 scores for the same methods, and the report presents examples of classified Vietnamese comments. Its interpretation notes difficulties distinguishing some sentiment categories, particularly neutral and positive comments. These figures describe the experiment in the thesis and do not by themselves establish performance on unrelated social-media data [1].

IV. Discussion

The comparison gives a concrete basis for judging the tested approaches: CNN led in reported accuracy, while the combined CNN-LSTM model followed closely. The report's examples and confusion discussion indicate that class boundaries remain imperfect. Informal language and overlap between neutral and positive expressions are plausible sources of error in the documented examples. The author concludes that the algorithm still requires optimization, so the result should be treated as a student experiment rather than a production-ready classifier [1].

V. Conclusion

The thesis develops a Vietnamese comment-sentiment experiment and compares SVM, CNN, LSTM, and CNN-LSTM methods. Its reported results place CNN highest in accuracy at 84.9%, with the other three methods also evaluated in the source table. The report demonstrates classification examples but acknowledges remaining errors and the need to optimize the algorithm. The documented contribution is the preprocessing and model-comparison workflow together with measured results for the thesis dataset. Further testing and refinement would be needed before broader use [1].

References

[1] NGUYỄN NGỌC HẢI, "Research and Development of an Application for Sentiment Analysis of Social Media Comments," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2024.
[2] Dr. Adrian Rosebrock, "Deep Learning for Computer Vision with Python", PyImageSearch (2017)
[3] Delip Rao và Brian McMahan, "Natural Language Processing with PyTorch", Manning Publications (2019)

[4] Aurélien Géron, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O’Reilly Media (2019)
[5] ICTU Student 197, ”Application of Genetic Algorithm in Text Classification of Vietnamese Language,” internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2010.
[6] Nhan Cach Dang et al., ”Sentiment Analysis Based on Deep Learning: A Comparative Study,” Electronics, 2020. doi:10.3390/electronics9030483.
[7] Md Jakaria Islam et al., ”Sentiment analysis of social media text using natural language processing,” World Journal of Advanced Research and Reviews, 2023. doi:10.30574/wjarr.2023.18.3.1161.
[8] Kartika Fithriasari et al., ”Deep Learning for Social Media Sentiment Analysis,” MATEMATIKA, 2020. doi:10.11113/matematika.v36.n2.1226.
[9] Qurat Tul et al., ”Sentiment Analysis Using Deep Learning Techniques: A Review,” International Journal of Advanced Computer Science and Applications, 2017. doi:10.14569/ijacsa.2017.080657.
[10] Praphula Kumar Jain et al., ”A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents,” ACM Transactions on Asian and Low-Resource Language Information Processing, 2021. doi:10.1145/3457206.