SOTICS 2023

The Thirteenth International Conference on Social Media Technologies, Communication, and Informatics

ISBN: 978-1-68558-103-9

November 13th – 17th, 2023

Valencia, Spain

SOTICS 2023 Editors

Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain

SOTICS 2023

Forward

The Thirteenth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS 2023), held on November 13 - 17, 2023 in Valencia, Spain, was an event on social eco-informatics, bridging different social and informatics concepts by considering digital domains, social metrics, social applications, services, and challenges.

The systems comprising human and information features form a complex mix of social sciences and informatics concepts embraced by the so-called social eco-systems. These are interdisciplinary approaches on social phenomena supported by advanced informatics solutions. It is quit intriguing that the impact on society is little studied despite a few experiments. Recently, also Google was labeled as a company that does not contribute to brain development by instantly showing the response for a query. This is in contrast to the fact that it has been proven that not showing the definitive answer directly facilitates a learning process better. Also, studies show that e-book reading takes more times than reading a printed one. Digital libraries and deep web offer a vast spectrum of information. Large scale digital library and access-free digital libraries, as well as social networks and tools constitute challenges in terms of accessibility, trust, privacy, and user satisfaction. The current questions concern the trade-off, where our actions must focus, and how to increase the accessibility to eSocial resources.

We take here the opportunity to warmly thank all the members of the SOTICS 2023 technical program committee, as well as all of the reviewers. We also kindly thank all the authors who dedicated much of their time and effort to contribute to SOTICS 2023.

We also gratefully thank the members of the SOTICS 2023 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that SOTICS 2023 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of social eco-informatics. . We also hope that Valencia provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

SOTICS 2023 Steering Committee

Elina Michopoulou, University of Derby, UK

SOTICS 2023 Publicity Chair

Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain

SOTICS 2023

Committee

SOTICS 2023 Steering Committee

Elina Michopoulou, University of Derby, UK

SOTICS 2023 Publicity Chair

Lorena Parra Boronat, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Jose Miguel Jimenez, Universitat Politecnica de Valencia, Spain

SOTICS 2023 Technical Program Committee

Lavanya Addepalli, Universitat Politecnica de Valencia, Spain
Md Shoaib Ahmed, Brain Station 23, Dhaka, Bangladesh
Millicent Akotam Agangiba, University of Mines and Technology, Tarkwa, Ghana
William Akotam Agangiba, University of Mines and Technology, Tarkwa, Ghana
Federico Martín Alconada Verzini, Universidad Nacional de La Plata, Argentina
Samer Al-Khateeb, Creighton University, US
Musfique Anwar, Jahangirnagar University, Bangladesh
Hassan Atifi, Troyes University of Technology (UTT), France
Faical Azouaou, École supérieure en Sciences et Technologies de l'Informatique et du Numérique, Algeria
Jiyang Bai, Florida State University, USA
Qanita Bani Baker, Jordan University of Science and Technology, Jordan
Asif Ali Banka, IUST Kashmir, India
Grigorios N. Beligiannis, University of Patras, Greece
Hernane Borges de Barros Pereira,Centro Universitário Senai Cimatec, Brazil
Christos Bouras, University of Patras, Greece
James Braman, Community College of Baltimore County, USA
Miguel Carvalho, INESC-ID Lisboa, Portugal / Coordinating Center for Communications and Information and Innovation Technologies, Regional Government of the Azores
K C Chan, University of Southern Queensland, Australia
Luisa Fernanda Chaparro Sierra, Tecnologico de Monterrey, Mexico
Dickson K.W. Chiu, University of Hong Kong, Hong Kong
Joshua Chukwuere, North-West University (NWU), South Africa
Subhasis Dasgupta, San Diego Supercomputer Center | University of California San Diego, USA
Vasily Desnitsky, SPIIRAS, Russia
Arianna D'Ulizia,National research council of Italy - IRPPS Research, Italy
Ritam Dutta, Surendra Institute of Engineering & Management | Maulana Abul Kalam Azad University of Technology, West Bengal, India
Luis Enrique Sánchez Crespo, Universidad de Castilla-La Mancha, Spain

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Stock Price Prediction Based on Investor Sentiment Using BERT and Transformer Models

Chien-Cheng Lee and ANISH SAH
Department of Electrical Engineering, Yuan Ze University
Taoyuan, Taiwan
e-mail: cclee@saturn.yzu.edu.tw, anishkb009@gmail.com

***Abstract*—This paper investigates the impact of investor sentiment on the stock market by predicting stock closing prices and future trends in stock returns. Our study involves gathering abundant investor messages from three social media platforms: Stocktwits, Yahoo Finance, and Reddit. To gauge investor sentiment from the collected messages, we employ Bidirectional Encoder Representations from Transformers (BERT), a transformer-based pre-trained language model. We present a novel application of a Transformer-based model for stock trend prediction. This model architecture leverages the self-attention mechanism to capture the interdependence of stock data, facilitating accurate forecasting of stock trends. By integrating investor sentiment with stock prices and inputting this combined information into the transformer model, we predict the performance of APPLE and SPY stocks datasets. Our experimental results reveal that the transformer model exhibits strong performance regardless of whether sentiment features are included. Moreover, incorporating sentiment does enhance the forecasting accuracy for both stock closing prices and future trends in stock returns.**

Keywords-Bert; Transformer; StockTwits; Stock Returns.

I. INTRODUCTION

In recent years, there has been a growing trend in utilizing text mining technology to automatically extract and analyze substantial amounts of textual data. By employing Natural Language Processing (NLP) techniques, these opinions are summarized, and their applications extend to various domains, including market forecasting [1]. Prior research has explored the financial implications of investor sentiment using conventional NLP techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) [2]. Conventional methods like TF-IDF have limitations in effectively capturing the overall sentiment of complete sentences due to their reliance on lexical frequency analysis. Fortunately, the advancements in NLP have led to the emergence of Bidirectional Encoder Representations from Transformers (BERT), a groundbreaking technology for language modeling [3]. BERT offers robust semantic representation and has demonstrated remarkable performance across various natural language understanding tasks. Its capabilities address the challenges faced by traditional approaches and present new opportunities for more accurate sentiment analysis in financial applications.

However, BERT models are pretrained on diverse text sources like Wikipedia and BookCorpus datasets, which differ considerably from the language used in stock market and economic narratives. This can result in challenges when directly applying these pretrained models for investor sentiment prediction, particularly in interpreting technical terms specific to the financial domain. To address this issue, specialized stock market BERT models are necessary.

The Transformer model, initially introduced by Vaswani et al. [4] in 2017, has brought about a revolutionary change in the realm of deep learning, especially in NLP tasks. However, its significance goes beyond NLP, as researchers have increasingly acknowledged its potential for addressing diverse problems, including stock prediction [5]. The strength of the Transformer model lies in its ability to capture long-range dependencies and learn complex patterns in sequential data. These attributes make it particularly suitable for modeling stock price time series, which are characterized by complex dynamics and influenced by various factors.

In this study, we investigate the influence of investor sentiment on the stock market. To achieve this, we gather investor messages from social media platforms and employ them to pretrain a specialized stock market BERT model for predicting investor sentiment. Subsequently, we conduct experiments by combining investor sentiment with Transformer models to predict stock closing prices and forecast future trends in stock returns. The experimental results reveal that the transformer model exhibits strong performance in both predictions. Furthermore, the incorporation of sentiment features significantly enhances the accuracy of the predictions.

The rest of this paper is organized as follows. Section II reviews related work. Section III describes dataset collection, language model training, sentiment prediction, and stock price and trend prediction. Section IV provides an evaluation of our approach by analyzing the obtained results and presenting comparisons with other methods. Finally, Section V offers some concluding remarks.

II. RELATED WORK

Lexicon-based methods rely on calculating the sentiment score of a vocabulary based on word frequency in optimistic and pessimistic texts. For instance, Oliveira et al. [2] used TF-IDF to determine sentiment scores for the vocabulary. However, the creation of lexicons poses challenges and limitations. One of the primary concerns is that general lexicons may not be well-suited for sentiment analysis in the stock market domain due to certain terms, such as "undervalued," having opposite sentiment interpretations in financial contexts.

 ISBN: 978-1-68558-103-9

In contrast, language model-based methods leverage models like BERT, which have emerged as robust language representation models pretrained on extensive amounts of unlabeled text. Howard and Ruder [7] have proposed various fine-tuning approaches for BERT, including universal language model fine-tuning (ULMFiT).

After conducting sentiment analysis, multiple studies have explored the impact of investor sentiment on the stock market. Kim et al. [8] utilized statistical methods and Naïve Bayes classification to determine investor sentiment from Yahoo Finance messages, evaluating its predictive power for stock returns. Meanwhile, Renault [6] derived investor sentiment from Stocktwits messages and demonstrated its efficacy in forecasting using linear regression models.

Deep learning techniques have also gained substantial traction in stock market research. For instance, Zhong and Enke [9] developed a neural network model that employs economic-related features to predict daily stock market return directions. Wang et al. [10] explored the use of a Transformer model for predicting stock market indices. By incorporating a self-attention mechanism to capture complex relationships in stock market data, the model exhibits higher accuracy than traditional forecasting methods. Additionally, Zhang et al. [11] introduced the Transformer Encoder-based Attention Network (TEANet) framework. This approach effectively captures temporal dependencies and facilitates accurate analysis of financial data. The application of these deep learning techniques marks significant progress in stock market prediction and analysis.

III. MATERIALS AND METHODS

A. Investor Message Dataset Collection

We designed a Python web scraping program to gather investor messages from three prominent social media platforms: Stocktwits, Yahoo Finance, and Reddit (specifically, the World News and News communities). Our data collection from Stocktwits spanned from July 2009 to December 2021. To ensure data integrity, we eliminated duplicate messages and messages containing solely URL addresses and emojis. Ultimately, we retained approximately 34 million messages. Among these, around 13 million were marked with sentiment labels: 11 million were labeled as bullish, 2 million as bearish, while the remaining 21 million were unmarked.

For Yahoo Finance, we collected data from the years 2016 to 2019, which amounted to approximately 1.4 million messages. The Reddit data was gathered during the period from July 2020 to November 2021, amounting to around 60,000 messages. The limited amount of data is due to the difficulty of capturing data, most of which is unlabeled.

B. Stock Market BERT Language Model

For our language model, we selected the BERT-Based pre-trained model provided by Google, which leverages general domain text and two advanced training techniques: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In this study, we took the pre-trained BERT model and performed additional pre-training, specifically tailored to the stock market domain. We accomplished this by utilizing unlabeled data from Stocktwits, Yahoo Finance, and Reddit, employing the MLM approach. This further pre-training of the model allows it to better understand the intricacies and nuances of the stock market language, enhancing its effectiveness in predicting investor sentiment for stock-related tasks.

C. Sentiment Predict Model

MLM further pre-trained model is used to build investor sentiment predictor by fine-tuning the target task classifier with binary output, bullish and bearish. The sentiment predictor uses the BertForSequenceClassification model implemented by the Hugging face library, with a single linear layer added on top for classification. In this study, the accuracy of the sentiment classification model on our validation dataset was 89%. Examples of sentiment prediction results are shown in Table I.

TABLE I. EXAMPLES OF SENTIMENT PREDICTION RESULTS FOR STOCKTWITS

Examples message	Label	Prediction
$SPY 2nd break of the trendline	0	0
$aapl liquidation of options friday hit this stock. a massive winning move!	1	1
$aapl rising corporate debt for stock buybacks. financial engineering vs innovation.	0	0
$FB Careful here. Speculation is overwhelming as shown in this chart.	1	1

After constructing the investor sentiment predictor, we can utilize it to predict whether each unlabeled message exhibits bullish or bearish sentiment. Subsequently, we calculate the daily sentiment index (S_t) using the following formula (1):

$$S_t = \frac{O_t - P_t}{N_t} \quad (1)$$

where O_t is the number of bullish messages on day t, P_t is the number of bearish messages on day t, and N_t is the total number of messages on day t. In this way, the daily sentiment index S_t ranges from -1 to 1. A negative value indicates prevailing bearish sentiment, a positive value indicates prevailing bullish sentiment.

D. Transforme Learning Models for the Impact of Investor Sentiment

In this study, we use the Transformer model to predict stock trends and evaluate its effectiveness in capturing the intricate dependencies and patterns within our stock dataset. The Transformer encoder combines self-attention and feed-forward layers, enabling it to efficiently capture the relationship between tokens in a sequence. By leveraging these mechanisms, our model can learn and exploit the dependencies among different stock data elements.

The subsequent layers of our model consist of global average pooling, dense layers, and softmax activation, collectively responsible for generating the classification output. Global average pooling is used to condense the sequence into a fixed-length representation for further analysis and decision-making. Dense layers introduce non-

 ISBN: 978-1-68558-103-9

linearity and higher-level representations, that enhance the model's ability to understand complex patterns. Finally, a softmax activation function predicts stock price returns by assigning probabilities to different classes.

During our experiments, we explored various parameter settings. The best-performing configuration utilizes one Transformer encoder block, one multi-head attention layer, and two convolutional layers used as a feed-forward network. To prevent overfitting and improve generalization, we add dropout after the multi-head attention layer. The training process of our model involves using the Adam optimizer and the sparse categorical cross-entropy loss function. The model's performance was assessed using the sparse categorical accuracy metric, measuring the accuracy of the predicted stock returns.

IV. Experimental Results

To study the impact of investor sentiment on the stock market, we conducted two experiments: predicting stock closing prices and predicting future trends in stock returns. The experiments were performed on a computer with an NVIDIA Geforce GTX 1080Ti GPU card with 32 GB of memory. We used Tensorflow to implement the models. For computational reason, we only investigated Apple (ticker: AAPL) and S&P 500 ETF (ticker: SPY). Stock price data, including daily opening, highest, lowest, adjusted closing prices, and closing prices, was downloaded from July 2010 to November 2021 on Yahoo Finance. Investor messages from Stocktwits over the same time period for these two stocks were also collected. The total number of sentiment messages for AAPL and SPY were 530,099 and 1,823,709, respectively.

In our analysis, we designated the data from July 2010 to December 2019 as the training data, while the data after that period was assigned as the test data. As part of the normalization process, we scaled each individual data column, particularly the stock price data, to fall within the range of [0, 1]. After prediction, we restore the output to its original range using the stored normalization parameter.

A. Prediction Results of Closing Prices

To demonstrate the impact of sentiment, we create two feature sets: one comprising both sentiment index and stock price data (daily opening, highest, lowest, and closing prices), and the other containing only stock price data. We utilize a 25-day data window to forecast one day ahead, specifically predicting the data for the 26th day. Subsequently, the feature set is input into the Transformer models to make predictions on stock closing prices.

In order to evaluate the prediction performance, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-square (R^2) are used as evaluation criteria in this study. Their formulas are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y_i}| \quad (2)$$

$$RMSE = \sqrt{\frac{i}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

Here, n is the total number of samples; y_i and $\hat{y}_i$ represent the true value and predicted value of the test set, respectively. The smaller the *MAE* value, the better the prediction. Likewise, the smaller the *RMSE* value, the better the prediction. That is, the closer the values of *MAE* and *RMSE* are to 0, the smaller the error between the predicted value and the true value. In R^2, the mean of the true values of the test set is not directly represented. Instead, R^2 measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (predictors) in the model. The value range of R^2 is (0, 1). The closer R^2 is to 1, the better the model fits the data.

Table II shows the closing price prediction results of the test data (including MAE, RMSE and R2 standards) of two stocks compared with Long Short-Term Memory (LSTM). The inclusion of sentiment features in the predictions results in better performance compared to predictions without sentiment features. Furthermore, the presence of sentiment features leads to an improvement in the accuracy of closing price prediction. Figure 1 illustrates the superior closing price prediction curves of the Transformer model for two stocks when using 25-day data with sentiment features. Our predicted closing price curves closely align with the true price curves, demonstrating the model's remarkable accuracy in capturing the stock price trends.

B. Prediction Results of Future Trends in Stock Returns

Most investors are concerned with future trends in stock returns rather than the actual price value. Hence, we adopt accuracy as a measure to assess the prediction performance of future trends in stock returns. The stock return is determined based on whether the closing price is up or down, and any changes in the closing price will consequently alter the stock return. The stock return $R_{t,i}$ of stock i on date t is calculated as follows:

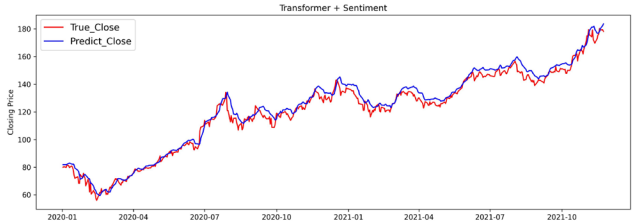
$$R_{t,i} = \frac{Close_i(t) - Close_i(t-1)}{Close_i(t-1)} \times 100 \quad (5)$$

where $Close_i(t)$ is the closing price of stock i on date t. A positive return indicates that the closing price today is higher than the closing price of the previous day. In classification problems, accuracy is a commonly used evaluation metric. To convert the evaluation of stock return regression into a binary classification, we define the labels for future trends as follows: a label of 1 represents an expected increase in stock return, while a label of 0 indicates a decrease or no change in stock return. By adopting this approach, we can assess the model's performance in predicting the direction of future stock returns, which is more intuitive for investors seeking to understand potential trends.

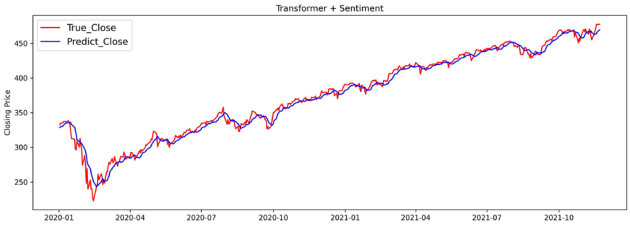
First, we calculate the daily stock return, denoted as $R_{t,i}$, and integrate it into the existing stock price data. This updated dataset includes the daily open, high, low, close prices, along with the corresponding stock returns. Next, we employ Transformer models to build stock return prediction models, allowing us to evaluate and compare their performance. By

 ISBN: 978-1-68558-103-9

incorporating stock returns, we aim to enhance the accuracy and effectiveness of our predictions, leading to valuable insights for investment decision-making.



(a)



(b)

Figure 1. Closing price prediction curves of the Transformer model for stocks using 25-day data with sentiment features. (a) AAPL, (b) SPY

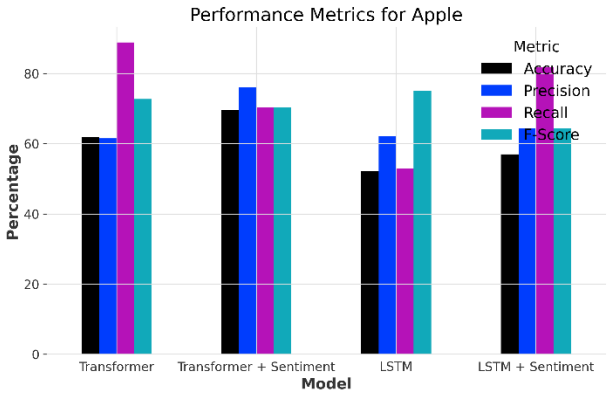
TABLE II. PREDICTION RESULTS OF CLOSING PRICES

Stock	Models	25- Days Data		
		MAE	*RMSE*	R^2
AAPL	Transformer	0.0301	0.0404	0.9878
	Transformer + Sentiment	0.0279	0.0374	0.9882
	LSTM	0.0372	0.0483	0.9829
	LSTM+ Sentiment	0.03298	0.0429	0.9823
SPY	Transformer	0.0176	0.0264	0.9885
	Transformer + Sentiment	0.0169	0.0217	0.9855
	LSTM	0.0271	0.0328	0.9825
	LSTM + Sentiment	0.0194	0.0230	0.9837

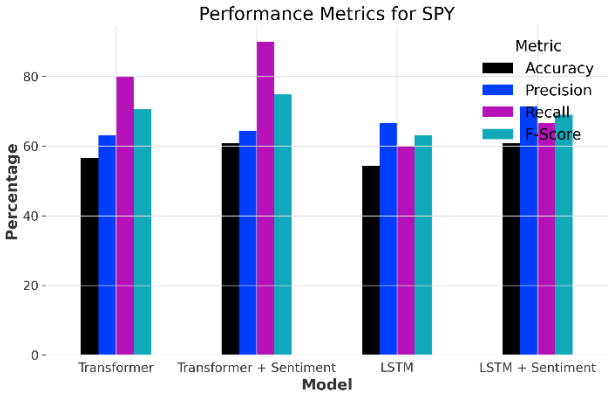
TABLE III. PREDICTION METRICS OF FUTURE TRENDS IN STOCK RETURNS (IN %)

Model	*Stock*	*Accuracy*	*Precision*	*Recall*	*F-Score*
Transformer	Apple	61.85	61.53	88.88	72.72
	SPY	56.52	63.15	80.00	70.58
Transformer + Sentiment	Apple	69.56	76.00	70.37	70.37
	SPY	60.86	64.28	90.00	75.00
LSTM	Apple	52.18	62.06	52.94	75.00
	SPY	54.34	66.67	60.00	63.15
LSTM + Sentiment	Apple	57.00	64.28	81.81	64.28
	SPY	60.86	71.42	66.66	68.96

To evaluate the influence of sentiment, we create two sets of features: one with sentiment and the other without sentiment. Table III and Figure 2 show the performance metrics compared to LSTM for predicting future trends in stock returns using 25 days of data. incorporating sentiment as a feature significantly improves the accuracy of these predictions. Notably, AAPL stock demonstrates the highest accuracy, achieving an impressive 70%. These findings highlight the valuable impact of sentiment in enhancing the precision of stock return predictions, particularly for the AAPL stock.



(a)



(b)

Figure 2. Visualized performance metrics for the stock data of (a) AAPL and (b) SPY.

V. CONCLUSIONS

In this paper, we employ BERT for sentiment classification in the stock market, focusing on individual stocks. Following this, we conduct a series of experiments to investigate the impact of investor sentiment on the stock markets. By harnessing the capabilities of the powerful Transformer model and optimizing its parameters, our experimental results reveal that incorporating sentiment information can potentially offer substantial benefits in enhancing the accuracy and effectiveness of stock market predictions. The findings underscore the value of sentiment analysis in better understanding market dynamics and making more informed investment decisions.

REFERENCES

[1] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, “Text mining for market prediction: A systematic review,”

Expert Systems with Applications, vol.41, no. 16, pp. 7653-7670, 2014.

[2] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," Decision Support Systems, vol. 85, pp. 62-73, 2016.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[4] A. Vaswani, et al., "Attention is all you need," The 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, 2017.

[5] T. Muhammad, et al., "Transformer-Based Deep Learning Model for Stock Price Prediction: A Case Study on Bangladesh Stock Market," ArXiv, 2022. abs/2208.08300.

[6] T. Renault, "Intraday online investor sentiment and return patterns in the US stock market," *Journal of Banking & Finance*, vol. 84, pp. 25-40, 2017.

[7] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv preprint arXiv:1801.06146, 2018.

[8] S. H. Kim and D. Kim, "Investor sentiment from internet message postings and the predictability of stock returns," *Journal of Economic Behavior & Organization*, vol. 107, pp. 708-729, 2014.

[9] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," Financial Innovation, vol. 5, no. 1, pp. 1-20, 2019.

[10] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep Transformer model," *Expert Systems with Applications*, vol. 208, 118128, 2022.

[11] Q. Zhang, et al., "Transformer-based attention network for stock movement prediction," *Expert Systems with Applications*, vol. 202, 15, 2022.

 ISBN: 978-1-68558-103-9

From Unstructured Data to Digital Twins: From Tweets to Structured Knowledge

Sergej Schultenkämper[#1], Frederik S. Bäumer[#2], Yeong Su Lee[*3], Michaela Geierhos[*4]
[#]*Bielefeld University of Applied Sciences and Arts*
Interaktion 1, 33619 Bielefeld, Germany
[1]sergej.schultenkaemper@hsbi.de
[2]frederik.baeumer@hsbi.de
[*]*University of the Bundeswehr Munich*
Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
[3]yeongsu.lee@unibw.de
[4]michaela.geierhos@unibw.de

***Abstract*—This paper focuses on extracting relevant information from unstructured data, specifically analyzing text shared by users on Twitter. The goal is to build a comprehensive knowledge graph by extracting implicit personal information from tweets, including interests, activities, events, family, health, relationships, and professional information. The extracted information is used to instantiate a Digital Twin and develop a personalized alert system to protect users from threats, such as social engineering or doxing. The paper evaluates the effectiveness of state-of-the-art large language models, such as GPT-4, for extracting relevant triples from tweets. The study also explores the notion of Digital Twins in the context of cyber threats and presents related work in information extraction. The approach includes data collection, multi-label classification, relational triple extraction, and evaluation of the results. The dataset used is from Twitter, and the study analyzes the challenges posed by user-generated data. The results show the accuracy of the extracted triples and the personal characteristics that can be identified from tweets for the development of the Digital Twin. The results contribute to the ADRIAN research project, which focuses on machine learning-based methods for detecting potential threats to people's privacy.**

***Index Terms*—*Digital Twin*; *Data Privacy*; *Semantic Triple*.**

I. Introduction

Through almost every activity on the Web, users leave footprints, both active and passive [1]. This includes obvious information, such as images, text, and video that users knowingly upload, as well as information that is transmitted without user intervention, such as the device's IP address or user agent. Moreover, information hidden in text and images that are unknowingly posted is difficult for users to keep track of. This has already been demonstrated in an impressive and media-effective way, e.g., by the automatic identification of vacation announcements and the extraction of hidden Global Positioning System (GPS) image data from Twitter (aka "X"), which could be used, for example, to scout vacant properties for burglaries [2] or to reveal the running routes of soldiers on secret army bases, whose publication on sports portals revealed the exact location of the military installations [3]. Even small amounts of information, when combined with other information, can be a threat, as research has shown [4].

In this paper, we focus on texts shared by users on Twitter to extract relevant information from unstructured, noisy data. According to [5], over 500 million tweets are shared every day for various reasons, such as expressing personal views, sharing news, or discussing current events. Data analysis can help extract both implicit and explicit personal information [6]. For this reason, this study focuses on analyzing tweets regarding *personal interests*, *activities*, *events*, *family*, *health*, *relationships*, and *professional information*. The goal is to construct a robust knowledge graph by retrieving information from unstructured sources. The instantiation of the Digital Twin (DT) is based on this graph. In this context, we plan to develop a personalized early warning system that will be centered around the DT. In case of excessive disclosure of sensitive information, this system can notify users in Online Social Networks (OSNs). Providing early warnings of potential threats, such as social engineering or doxing, can mitigate risks to users. To address the challenge of information extraction from tweets, we investigate the suitability of current state-of-the-art Large Language Models (LLMs), such as Generative Pretrained Transformer 4 (GPT-4), for extracting significant triples in the form of subject, predicate, and object. The purpose of this study is to evaluate the effectiveness of LLMs for information extraction. The objective is to identify significant triples from unstructured and minimally informative Twitter data of varying lengths [7].

All of these considerations are part of the research project ADRIAN, which stands for "Authority-Dependent Risk Identification and Analysis in online Networks". For this purpose, we discuss related work in Section II and describe the dataset and our approach in Section III. Subsequently, we present our modeling strategies and results in Section IV. Finally, we discuss our findings (Section V) and draw conclusions in Section VI.

II. Related Work

In this section, we discuss the notion of DTs in the context of cyber threats and the related work on information extraction.

 ISBN: 978-1-68558-103-9

A. Digital Twins in the Context of Cyber Threats

The term DT is ambiguous and finds application in various research and practical domains, including medicine and computer science [3] [8] [9]. The advancement of artificial intelligence has broadened the scope of this term, and in a broader sense, DTs can be defined as computer-based models or (physical and/or virtual) machines that simulate, emulate, mirror, or act as a "twin" of a real-world entity, which could be an object, a process or a human [8]. In our context, we identify three distinct levels of integration for DTs, according to [8]: (a) *Digital Model*, (b) *Digital Shadow*, and (c) *Digital Twin*. A digital model is a basic virtual representation of a physical object or system, without any automatic exchange of information between the virtual and physical worlds. Any changes to the physical object must be manually updated in the digital model. In the future, a digital shadow builds on this concept by enabling a unidirectional, automatic flow of information from the physical to the virtual world. Sensors capture information from the physical entity and transmit signals to the virtual model. Finally, a complete DT exists when bidirectional communication is established between the virtual and physical environments, facilitating the automatic exchange of information. This allows the DT to accurately reflect the real-time state and evolution of its physical counterpart. However, when considering socio-technical systems, the dynamics change as these systems include both human and machine components.

Therefore, it becomes relevant to explore the notion of a *Human DT* [10]. Despite its growing importance, a standardized definition or understanding of this concept has not yet been achieved [9] [11] . The digital information available about individuals is often referred to as the *Digital Footprint* or *Digital Representation*, with these terms often used interchangeably. These terms refer to the data left behind by users on the Internet, often unknowingly, without identifying or linking to a specific individual. The concepts of Digital Footprint, Digital Shadow, and Digital Twin can be distinguished on the basis of several aspects: Identifiability, active or passive data collection, individualized or aggregated evaluation, real-time or delayed analysis, decision-making authority, and comprehensive representation [10]. The Human DT aims to store and analyze relevant characteristics of an individual in a given situation. This may include demographic or physiological data, skill or activity profiles, or health status [9] [10].

In the ADRIAN project, we define a DT as a digital representation of a real person, instantiated using publicly available information from the web [3] [12] . It is important to note that a DT can never fully capture the complexity of a real person, but rather reproduces specific characteristics that, either alone or in combination with other characteristics, may pose a threat to the individual [4].

B. Relational Triple Extraction

Relation Extraction (RE) is the identification of pairs of entities and their relations, expressed as (HEAD, RELATION, TAIL), from unstructured text. Traditional approaches to RE are divided into two separate tasks: Named Entity Recognition (NER) and Relationship Classification (RC) [13] [14]. However, this approach is prone to the problem of error propagation. Therefore, recent studies have aimed to overcome this problem by exploring common models for extracting relational triples in an end-to-end manner [15]. For example, one approach is the text generation technique, which treats a triple as a series of tokens, and uses the encoder-decoder architecture to generate triple components, similar to machine translation [16] [17].

Recent studies on text generation have highlighted In-Context Learning (ICL) as an important feature of GPT-3.5, GPT-4, and other transformer-based models [18] [19]. Unlike traditional machine learning models that require explicit and task-specific training datasets, models with ICL capabilities can learn and adapt to new tasks by using the context provided during inference. For example, GPT-3.5 or GPT-4 achieve ICL by providing a set of contextual examples at the beginning of a prompt. These examples help guide the model's responses. As a result of this capability, these models can perform a wide range of tasks without the need for task-specific fine-tuning, resulting in highly flexible and adaptable models.

Xu et al. (2023) [19] explore the use of LLMs for few-shot RE. The paper focuses on the approach of ICL, which involves designing prompts of varying complexity to help LLMs understand the task of RE. Two types of prompts are used: (1) The text prompt contains only the essential elements for requirements engineering. (2) The instruction prompt includes a task-related instruction that describes the requirements engineering task as well as the essential elements. The results indicate that instructions and schemes in ICL are crucial for RE with LLMs. In general, the instruct prompt model outperformed the text prompt model. According to the study, the inclusion of task-related information, such as instructions or schemes, is essential for effective ICL with LLMs.

Wan et al. (2023) [18] present GPT-RE, an innovative method for Relational Extraction (RE) that utilizes the ICL abilities of GPT-3. The method consists of two primary components. The first component is entity-aware demonstration retrieval, which reconstructs the context by incorporating information about the entity pair that is critical for RE. The second component is inferential demonstration, which enhances the demonstrations with inferences derived from the ground truth relationship labels. This facilitates GPT-3 to gain better understanding of the demonstrations and enhance its performance. The study's results indicate that GPT-RE surpasses fine-tuning on three datasets, implying that GPT-3 possesses the potential to perform outstandingly when the retriever has prior task knowledge. It is observed that the quality of demonstrations holds greater significance than their quantity. Furthermore, demonstrations enriched with reasoning present consistent improvement across all k-shot settings, implying that GPT-3's reasoning ability could be successfully unlocked by employing reasoning based on ground truth relational labels, thus enhancing ICL. The proposed GPT-RE

 ISBN: 978-1-68558-103-9

method attains the highest scores in the SemEval 2010 and SciERC datasets, exhibiting its efficacy in RE.

III. Approach and Dataset

This section presents our approach (Fig. 1), which includes (1) data collection, (2) multi-label classification, (3) relational triple extraction, and (4) evaluation of the extraction. Selecting an appropriate data source is the first step, and we chose Twitter as our data source. Twitter contains a significant amount of structured and unstructured data, making it a suitable candidate for instantiating DTs and performing threat analysis as part of the ADRIAN project. Twitter's Application Programming Interface (API) provides extensive and easily accessible data, adding to its attractiveness as a data source. Tweets, constrained by character limits, present a significant challenge. Furthermore, hashtags, user mentions, URLs, and emoticons in tweets generate substantial noise, making it difficult to extract reliable information for meaningful insights [7] [20]. In our data collection, we randomly selected 300 users from our database to determine the frequency of tweets. Of these selected users, 246 had posted tweets, with an average frequency of 3,532 tweets (cf. Table I).

TABLE I
Descriptive Statistics of the Twitter Dataset

Dataset Feature	Count
No. of Users	246
Avg. Tweets/User	3,532
Median Tweets/User	546
Min. Tweets/User	1
Max. Tweets/User	80,689
Total Tweets	869,069
Top Languages	EN, DE, FR, ES, TR
No. of Reply Tweets	274,504
No. with Attachments	106,997
No. with Geolocation	43,138
No. of Retweets	236,553

It should be noted that a user can have a very large number of tweets, with our approach it is not possible to process such a large number of tweets because the extraction with GPT-4 consumes immense cost, and therefore in this work we limit ourselves to a number of 5,000 tweets for the extraction of the triples. Therefore, we use pre-filtering to identify relevant tweets at the initial stage. For this purpose, we chose the OffMyChest dataset [21], since it is one of the few available datasets with personal information. The dataset classification varies from 'little personal information', which includes basic author information, such as *age*, *occupation*, or *location*, to information about *family*, *interests*, or *hobbies*. Conversely, 'much personal information' consists of sensitive data that has the potential to expose individuals, such as their *health conditions*, *physical attributes*, *behavior*, or *personal experiences*, as explicitly stated in the record description. However, the dataset description does not distinguish between these two categories of information. So, based on our requirements, we annotated the information with the following labels: *Event*, *Family*, *Health*, *Interests*, *Personal Information*, *Relationships*, and *Work/School*, as shown in Table III. We use the annotated dataset to train the Cross-lingual Language Model - Robustly Optimized BERT Pretraining Approach (XLM-RoBERTa) [22] for multi-label classification. After pre-filtering with this model, 50 tweets from 100 users are randomly selected for relational triple extraction. To structure the triples, we use properties from Schema.org [23], as shown in Table II.

TABLE II
Defined Categories and Associated Properties

Category	Properties
Event	s:attendee
Family	s:children, s:parent, s:sibling, s:spouse
Health	s:diagnosis, s:drug, s:healthCondition
Interests	s:interests
Personal Information	s:birthDate, s:birthPlace, s:email, s:gender, s:location, s:nationality
Relationships	s:colleague, s:knows
Work/School	s:alumniOf, s:jobTitle, s:workLocation, s:worksFor

All of the entities used are associated with Schema.org's Person class, and their structure enhances the clarity and interpretability of the data. Furthermore, the targeted triples can be directly integrated into our current applications. Thus, the DT can be instantiated directly. The generated dataset can be used to fine-tune available open-source models, such as LLaMa-2 [24] for triple extraction.

IV. Modeling and Results

The modeling phase consists of two main tasks: (1) multi-label classification and (2) few-shot prompt relational triple extraction. The first task involves annotating data from the OffMyChest dataset and then training an XLM-RoBERTa model based on the annotated data. The second task focuses on relational triple extraction, with core work including prompt engineering with examples for the few-shot approach and defining an output scheme.

A. Multi-Label Classification

We use XLM-RoBERTa, a multilingual language model, to train the multi-label classification model. This model uses the Masked Language Model (MLM) technique, like Bidirectional Transformers for Language Understanding (BERT) [25], but was trained on monolingual content from 100 different languages [22]. Our approach can be used in multiple languages due to the multilingual capabilities of XLM-RoBERTa and GPT-4. Our model is trained and evaluated on 1,803 annotated sentences from the OffMyChest dataset. The dataset was partitioned for training (80 %, 1,442 sentences) and evaluation (20 %, 361 sentences). The Weights & Biases library [26] is used to facilitate hyperparameter optimization and to monitor the progress of the experiment. Optimal performance was achieved on XLM-RoBERTa using a learning rate of 5e-5, over five training epochs, and with a training batch size of 16. The evaluation results are shown in Table III.

 ISBN: 978-1-68558-103-9

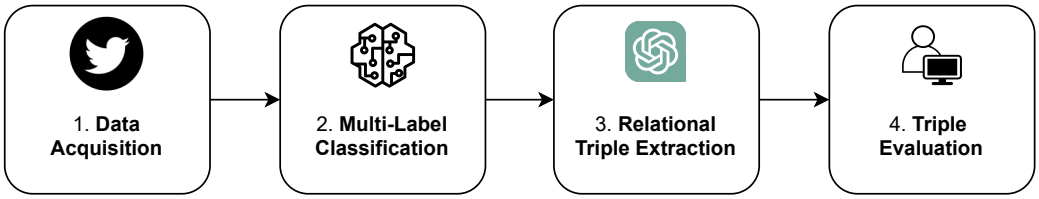


Fig. 1. Proposed approach for relational triple extraction

TABLE III
XLM-RoBERTa Classification Results

Class	Precision	Recall	F_1-score	Support
Event	0.7736	0.8542	0.8119	48
Family	0.8505	0.8349	0.8426	109
Health	0.7969	0.7846	0.7907	65
Interests	0.8333	0.6604	0.7368	53
Personal Information	0.7093	0.7871	0.7462	155
Relationships	0.7795	0.9000	0.8354	110
Work/School	0.7963	0.8600	0.8269	50

The results show the performance of the XLM-RoBERTa model across different classes. The highest precision score is observed in the "*Family*" class, while "*Relationships*" yields the highest recall. Regarding the F_1-score, the "*Family*" class presents the highest value. On the contrary, the "*Interests*" class has the lowest recall and F_1-score, while the "Personal Information" class has the lowest precision.

B. Few-Shot-Prompt Relational Triple Extraction

The prompt is designed to use the standard message and function call feature to create a scheme with appropriate descriptions. The prompt specifies that subject-predicate-object triples must be extracted from the provided tweets. Each tweet contains the author's name and the message content. It is important to note that the subject does not have to be the author of the tweet; other subjects should be identified and extracted. As recent studies have shown positive results with ICL [18] [19], we will provide examples in the form of triples.

Designing an output scheme involves using the function call feature of GPT-4. The function takes a scheme as input and generates a corresponding JavaScript Object Notation (JSON) object as its output. This JSON object can then be used to interact with external APIs or databases. The scheme is designed to extract subject-predicate-object triples from the input data. The subject refers to the entity that performs an action or the entity about which a statement is made. The predicate describes either the action performed by the subject or the state of the subject; a predefined set of values defines this attribute (cf. Table II). The object represents either the entity that is the subject of the action performed by the subject or the entity that is in some way involved in the action expressed by the predicate. Table IV illustrates an instance of the expected output of the function, where each row corresponds to a triple.

TABLE IV
Example output for the defined scheme

Subject	Predicate	Object
John	s:worksFor	Microsoft
Mary	s:location	New York

The next step is to validate the results of the GPT-4 model. For this task, we use the author, the tweets, and the extracted triples. To validate the results, we use LabelStudio [27] to annotate the correctness of the extracted subject, predicate, and object. A total of 1,288 triples were extracted from 5,000 tweets using GPT-4. The evaluation score is calculated based on the accuracy of each extracted property (cf. Table V).

TABLE V
Results for the GPT-4 Triple Extraction

Predicate Class	Subject	Predicate	Object	Support
s:alumniOf	1.0000	0.6897	0.8966	30
s:attendee	0.9760	0.9162	0.9401	167
s:birthDate	1.0000	0.8235	0.8824	17
s:birthPlace	1.0000	0.5000	1.0000	2
s:children	1.0000	1.0000	1.0000	9
s:colleague	1.0000	0.9605	0.9605	76
s:diagnosis	1.0000	0.6923	0.9231	14
s:email	1.0000	0.6667	1.0000	3
s:healthCondition	0.9091	0.8636	0.9545	22
s:interests	0.9942	0.9796	0.9650	352
s:jobTitle	0.9800	0.7800	0.8400	101
s:knows	1.0000	0.9643	1.0000	29
s:location	0.9908	0.9495	0.9312	219
s:nationality	1.0000	1.0000	0.8333	6
s:parent	0.9516	0.9032	0.9355	63
s:sibling	1.0000	0.8696	0.9565	23
s:spouse	1.0000	0.9000	0.9000	10
s:workLocation	1.0000	0.8889	0.8889	18
s:worksFor	0.9840	0.8560	0.8960	127
Micro Avg.	0.9866	0.9142	0.9331	Σ 1,288
Macro Avg.	0.9887	0.8528	0.9318	Σ 1,288
Weighted Avg.	0.9867	0.9142	0.9332	Σ 1,288

The high accuracy for the subject should be taken with caution, as it mostly concerns the tweeter. The model performs well in extracting predicates and objects from the given tweets. The model achieves average accuracies of 0.9142 (micro), 0.8528 (macro), and 0.9142 (weighted) for predicates, and 0.9331 (micro), 0.9318 (macro), and 0.9332 (weighted) for objects. Predicates with low counts, such as "s:birthPlace", "s:children", and "s:nationality", are underrepresented and require more data for a reliable measure of model performance.

 ISBN: 978-1-68558-103-9

V. Discussion

The multi-label classification gave satisfactory results for all categories except "*Interests*". It is reasonable to hypothesize that the ambiguity associated with this particular class may be due to the broad and diverse nature of interests, covering a variety of areas, such as hobbies, occupations, artistic pursuits, sports, and other domains. The wide range may make it difficult to achieve accurate and consistent classifications. Intent classification could serve as an additional approach to pre-selecting tweets. Understanding the intent behind a tweet is as important as identifying its topic. Determining whether the user is self-disclosing, sharing information, or commenting on others is critical. Such insights can lead to a nuanced interpretation of the tweet's content and context, helping to develop more accurate classification strategies. Our method is effective in quickly extracting triples that can be transformed into a graphical representation (Fig. 2).

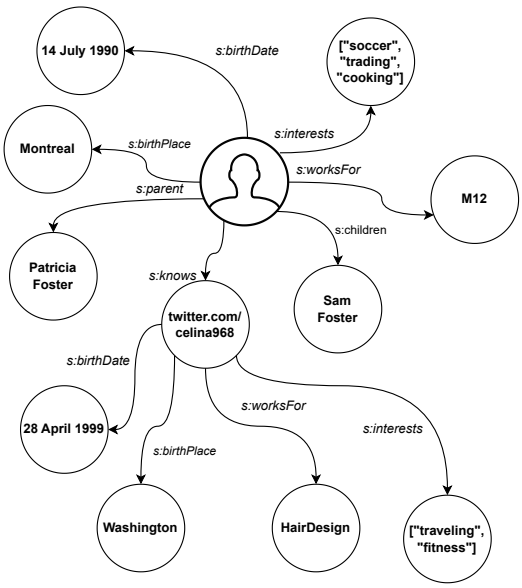


Fig. 2. An example of a knowledge graph constructed from extracted triples

The knowledge graph visually represents the relationships between extracted entities, providing a structured format that facilitates interpretation and analysis of the data. The nodes represent entities, while the edges represent the relationships between them, allowing the interconnectedness and hierarchical relationships of the extracted information to be explored.

However, the approach is not flawless due to the significant variance in the data, which requires normalization. Although the subject and predicate require minimal normalization, the object requires a different method due to the high variance during extraction. A common problem is the extraction of multiple individuals, usually seen with the target predicates "s:colleague" and "s:knows". Moreover, the existence of Twitter handles is a common problem that needs to be solved. Locations, represented by flags or abbreviations, are also extracted and need to be processed. Job titles sometimes include the company name, as in the case of "CTO of OpenAI". Although normalization in this particular case is straightforward, finding a solution for the following examples is much more challenging. It is necessary to analyze how to handle cases where there is no precise data, but statements like "*Father of 3 children*" or "*Photo of my dad and me ...*" that do not contain information about a specific person, but still have relevant information when combining the extracted information. This data should be considered as it may help to find additional family-related information in subsequent tweets or to enrich the dataset with information from the photo. For example, some information was labeled as "s:interest", but to better reflect its purpose, it could be categorized using additional predicates, such as "s:searchAction" or "s:skills". This would be useful if a user was searching for an employee or writing a tweet about specific skills. Furthermore, the "s:email" category seems too narrow in scope, and replacing it with "s:contactPoint" would allow for the inclusion of contact information from OSNs, such as LinkedIn links.

Closed LLMs, such as GPT-4 should be viewed critically in research, as they are not open, require payment, and lack transparency. In addition, they are not always robust, and their APIs are often unreliable and inaccessible. Therefore, open-source LLMs should be considered as an alternative. Open-source models offer transparency, flexibility, and community collaboration, allowing for review and improvement by a wider range of researchers. They are often more robust and adaptable, making them a preferable choice for research efforts. Reducing reliance on closed and expensive LLMs and encouraging the use of open-source alternatives is crucial to promoting openness, transparency, and progress in language modeling research. Of course, this requires that the necessary computing power (e.g., graphics cards) be made available. To be fair, this also comes at a cost, but it reduces dependency and increases transparency.

VI. Conclusion

Finally, this paper addressed the challenge of extracting relevant information from unstructured data, specifically tweets. The study demonstrated the potential of LLMs, such as GPT-4, to extract triples and build a comprehensive knowledge graph. By instantiating a DT and developing a personalized early warning system, the study aimed to protect users from potential threats arising from the disclosure of sensitive information. The results of the study demonstrate the effectiveness of the XLM-RoBERTa model in multi-label classification and provide insights into the personal characteristics expressed in tweets. In addition, the few-shot prompt relational triple extraction approach demonstrates the potential of GPT-4 to extract structured information from unstructured data. The designed prompts and output scheme enable the identification and representation of subject-predicate-object triples, contributing to the construction of a knowledge graph.

These modeling and extraction techniques lay the foundation for further advances in the field. With the availability

 ISBN: 978-1-68558-103-9

of powerful tools and frameworks, there is an opportunity to train custom LLMs tailored to specific domains or datasets. This opens up new possibilities for improved accuracy and domain-specific insights in information extraction tasks. Looking ahead, training custom LLMs has several advantages. First, it allows for better adaptation to specific domains and datasets, leading to improved extraction accuracy and relevance. Custom LLMs can be fine-tuned for specialized datasets, including those containing personal information, thereby improving the performance and applicability of the extraction process. Moreover, training custom LLMs enables greater control over privacy and data security. By using in-house models, organizations can ensure that sensitive information remains within their infrastructure, reducing the risk of data breaches or unauthorized access. This approach addresses the growing concern about privacy and the need to protect personal data in today's digital landscape. In addition, custom LLMs provide the opportunity for continuous learning and refinement. By training models on evolving datasets and incorporating feedback from user interactions, the accuracy and performance of the extraction process can be continually improved. This adaptability is essential to keep pace with emerging trends, language variations, and evolving threats in online platforms.

In summary, training custom LLMs brings several benefits, including improved extraction accuracy, enhanced privacy and data security, and the potential for continuous learning and refinement. These benefits open up new avenues for research and development in the field of information extraction, providing valuable insights and actionable intelligence to mitigate threats and protect users in online environments.

ACKNOWLEDGMENT

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

REFERENCES

[1] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris, "Tracing Cross Border Web Tracking," in *Proceedings of the Internet Measurement Conference 2018*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 329–342.

[2] M. B. Flinn, C. J. Teodorski, and K. L. Paullet, "Raising Awareness: An Examination of Embedded GPS Data in Images Posted to the Social Networking Site Twitter," *Issues in Information Systems*, vol. 11, no. 1, pp. 432–438, 2010.

[3] F. S. Bäumer, S. Denisov, Y. Su Lee, and M. Geierhos, "Towards Authority-Dependent Risk Identification and Analysis in Online Networks," in *Proceedings of the IST-190 Research Symposium (RSY) on AI, ML and BD for Hybrid Military Operations (AI4HMO)*, A. Halimi and E. Ayday, Eds., 2021.

[4] F. S. Bäumer, N. Grote, J. Kersting, and M. Geierhos, "Privacy Matters: Detecting Nocuous Patient Data Exposure in Online Physician Reviews," in *International Conference on Information and Software Technologies*. Springer, 2017, pp. 77–89.

[5] A. P. Gopi, R. N. S. Jyothi, V. L. Narayana, and K. S. Sandeep, "Classification of tweets data based on polarity using improved RBF kernel of SVM," *International Journal of Information Technology*, vol. 15, no. 2, pp. 965–980, January 2020.

[6] M. Mazza, G. Cola, and M. Tesconi, "Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter," *Online Social Networks and Media*, vol. 31, p. 100224, 2022.

[7] W. Ahmed, P. A. Bath, and G. Demartini, "Using Twitter as a data source: An overview of ethical, legal, and methodological challenges," *The ethics of online research*, vol. 2, pp. 79–107, 2017.

[8] B. R. Barricelli, E. Casiraghi, and D. Fogli, "A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications," *IEEE Access*, vol. 7, pp. 167 653–167 671, 2019.

[9] S. Schultenkämper and F. Bäumer, "Privacy Risks in German Patient Forums: A NER-based Approach to Enrich Digital Twins," in *Information and Software Technologies*. Cham: Springer International Publishing, 2023, In press.

[10] G. Engels, "Der digitale Fußabdruck, Schatten oder Zwilling von Maschinen und Menschen," *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, vol. 51, no. 3, pp. 363–370, August 2020.

[11] K. Feher, "Digital identity and the online self: Footprint strategies – An exploratory and comparative research study," *Journal of Information Science*, vol. 47, no. 2, pp. 192–205, 2019.

[12] F. S. Bäumer, J. Kersting, M. Orlikowski, and M. Geierhos, "Towards a Multi-Stage Approach to Detect Privacy Breaches in Physician Reviews." in *SEMANTICS Posters&Demos*, 2018.

[13] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction," *Journal of Machine Learning Research*, vol. 3, pp. 1083–1106, 2003.

[14] Y. S. Chan and D. Roth, "Exploiting syntactico-semantic structures for relation extraction," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 551–560.

[15] Y.-M. Shang, H. Huang, X. Sun, W. Wei, and X.-L. Mao, "Relational Triple Extraction: One Step is Enough," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, July 2022, pp. 4360–4366.

[16] X. Zeng, D. Zeng, S. He, K. Liu, and J. Zhao, "Extracting relational facts by an end-to-end neural model with copy mechanism," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 506–514.

[17] D. Zeng, H. Zhang, and Q. Liu, "Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 9507–9514.

[18] Z. Wan *et al.*, "GPT-RE: In-context Learning for Relation Extraction using Large Language Models," 2023.

[19] X. Xu, Y. Zhu, X. Wang, and N. Zhang, "How to Unleash the Power of Large Language Models for Few-shot Relation Extraction?" 2023.

[20] J. T. B. Jafar, "Information extraction from user generated noisy texts," Ph.D. dissertation, The Ohio State University, 2020.

[21] K. Jaidka, I. Singh, J. Lu, N. Chhaya, and L. Ungar, "A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure," in *Proceedings of the AAAI-20 Workshop on Affective Content Analysis*. New York, USA: AAAI, 2020, pp. 118–129.

[22] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451.

[23] Schema.org, "Schema.org - Structured Data for the Web," 2023, Available: https://schema.org, retrieved 2023/10/02.

[24] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," 2023.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: ACL, June 2019, pp. 4171–4186.

[26] L. Biewald, "Experiment Tracking with Weights and Biases," 2020, Available: https://www.wandb.com/, retrieved 2023/10/02.

[27] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020-2023, Available: https://github.com/heartexlabs/label-studio, retrieved: 2023/10/02.

 ISBN: 978-1-68558-103-9

Emoji Trends Between Abnormal Usage and Cultural Differences: A Case Study of Emojis

Ahmed Almohanadi
Tokyo Metropolitan University
Tokyo, Japan
email:ahmed-almahanadi@ed.tmu.ac.jp

Shohei Yokoyama
Tokyo Metropolitan University
National Institute of Informatics
The University of Tokyo
Tokyo, Japan
email:shohei@tmu.ac.jp

***Abstract*—Emojis can be used to understand people's emotions through non-lexical language around the world, and through different techniques, they can be used to study and understand people's trends and cultural traits and gauge abnormal activities or events. This paper analyzes a monthly data period consisting of multiple activities where unique trends and abnormal emojis are highly likely to occur. An established method has been created, and map projection was used to establish this relation. We found interesting cultural traits, special events, and coronavirus pandemic-related emojis projected on a clustered map, showing a strong relation between certain emojis and specific regions. All these are common emojis that can be seen daily and unique emojis that appear in special events.**

***Keywords*—Emoji; Twitter; Clustering; Geolocation Data; Data Analysis.**

I. Introduction

Emojis in today's online communication are very popular due to their brevity, conciseness, and ability to give the message an emotional coloring. Moreover, these special characters greatly facilitate global communication, allowing people to convey ideas and even complex narratives concisely and understandably.

As a result, emojis have become an integral part of almost any digital communication, from interpersonal communication to brands and companies communicating with their customers. Since emojis have become commonplace in modern communication, trends in their use have been repeatedly explored.

According to Kaur et al. [1], emojis can be used to portray emotional reactions among social media users who understand the emoji similarly. The images facilitate non-verbal communication cues on social media that enhance user exchange and interaction.

Using Emojis is a dynamic concept. The Emoji set in smartphones is constantly updated as a reaction to events in the world. Moreover, often, long-standing images take on new meanings, reflecting the ever-changing landscape of cultural, social, and technological development.

This evolution provides researchers with a unique opportunity to delve into the complex relationships between emojis and the societies that use them. In addition, emojis contribute to formulating and expressing public opinion regarding significant incidents and events.

Therefore, this research paper analyzes the emoji trends that are common today, as well as the features of their formation due to geographic and cultural differences between users. Analyzing the patterns and frequency of emoji usage across countries and regions, this study seeks to discover an exciting insight into the cultural nuances of digital expression. Since emojis are a cultural characteristic, this study suggests that different figures' usage patterns and meanings may differ from country to country.

Moreover, countries and regions may have localized symbols reflecting the interplay between regional cultures and digital language, providing a glimpse into the nature of online communication worldwide. An analysis of the geographical features of Emoji will help establish the universality of this form of communication and try to understand the global trends.

The prominence of emojis applies across many parts of the world, from the West to the East. According to Dyer and Kolic [2], the state of California in the United States has been found to have a high concentration and multiplicity of emoji usage, especially along the coastal area. The trend is connected to the high population in the state's central coastal cities, including San Diego, Los Angeles, Palo Alto, San Francisco, and others (Kejriwal et al., [3]). The high usage is also correlated to the linguistic and cultural diversity of the area, which promotes greater exchange and interaction.

The rest of this paper is organized as follows. Section 2 reports related work. Section 3 describes our proposed method. Section 4 presents the results and findings in three subsections. Section 5 concludes the paper and discusses future work.

II. Related Work

The peculiarities of using emojis have attracted considerable attention from researchers. However, due to geographical factors, the distinctive features of this type of communication have been identified relatively recently.

Moreover, most researchers in this field emphasize more than the influence of the region's geographical features but the cultural trends of the countries that determine the use of Emoji [4]. Assessing the influence of the country's cultural

 ISBN: 978-1-68558-103-9

characteristics on the use of Emoji, the researchers found variability in the interpretation of Emoji in different cultures.

Moreover, the fewer ways of communication between different countries and cultures, the more pronounced the difference in the use and meanings of emojis. For example, Guntuku et al. [4] found that there are significant differences in the interpretation and frequency of emoji usage between the East (China and Japan) and the West (United States and United Kingdom). Representatives of Eastern cultures often use text messages instead of using the corresponding figures.

In addition, since emojis were initially conceived as a way to color messages emotionally, their meaning in different cultures is influenced by general trends regarding the manifestation of emotions and feelings. Continuing to analyze the difference in emoji usage between Western and Eastern countries, Gao and VanderLaan [5] found that Westerners and Easterners appear to perceive emotions differently.

This difference in perception has been established even with basic emotions like happiness and sadness. Therefore, the researchers concluded that people might perceive the intent of the communication more accurately during exchanges with others of the same, compared with a different cultural background. This casts doubt on the claim that Emojis are a universal way of global communication.

On the other hand, using these figurines has demonstrated its ability to facilitate intercultural communication. Research has explored how these symbols bridge language barriers and facilitate interaction between different language communities, enriching global digital discourse.

In addition, the development of Emoji contributes to forming a diversified and inclusive world society. For example, the appearance of images with different skin colors or pairs in different combinations contributed to the unification of people since everyone can express themselves.

Despite the presence of regional differences, Kimura-Thollander and Kumar [6] note that the use of Emoji contributes to global cultural exchange and the formation of a common background among representatives of different communities.

Moreover, the participants in their survey came to the conclusion that it is necessary to create universal emojis that will have the same meaning in all countries of the world. Such an approach has the potential to simplify global communication but, at the same time, threatens to destroy global cultural diversification.

The use of emojis is constantly evolving and adapting to changing human needs. Therefore, many people attach particular meanings to individual images that are understandable only to them. Schouteten et al. note that in modern digital interpersonal communication, it is common for people to use emojis as a kind of shorthand.

At the same time, this form of communication may be meaningless for other people. Brands often use the same approach to create a specific image of the company in the consumer's perception. Distributing unique, localized emojis works the same way, with a group of people defined according to their nationality or where they live.

After collecting statistics on the most popular emojis in different countries, Cohen [7] determined that the red heart and the laughing face with tears are the most popular around the world. At the same time, the national flags of countries are in second place in popularity, which expresses their patriotic sentiments.

A rather remarkable figure is the image of two folded hands, which in some cultures is interpreted as a high-five, while in others, such as India, it expresses gratitude [8]. These patterns are due to distinctive cultural, linguistic, and national features that are characteristic of a particular region.

Finally, Emojis are a universal way to express public sentiment on social media. In addition, emoji usage trends may change in response to national situations or significant global events. Analyzing user pages on Twitter allows you to determine public opinion regarding certain global events, for example, as was the case with the COVID-19 pandemic. At the same time, due to the global nature of the incident, corona-related emojis are universal and have the same meaning regardless of region or country. In addition, there are unique emojis that do not have global popularity but are often used by certain groups of people. Examples of such images are sunflower, unicorn, rocket, artist palette, and top hat emojis [9].

These symbols may not be universally accepted, but their meaning in a particular community makes them invaluable for conveying subtle meanings. For example, they can be a manifestation of cultural self-identification, belonging to a particular subculture, and a form of self-expression. Unique emojis often have a distinct visual style that draws attention and catches the eye, which is why they are gaining popularity among social media users.

III. METHOD

In this study, the data collected for the research comprised using Twitter analysis tool for the period selected was one month between the 19th of November and the 18th of December of 2022; during this period, the Qatar World Cup 2022 was conducted, and Winter started in many areas of the world.

The gathered tweets were filtered with two criteria: first, they must include geolocation information, and second, the tweet must be classified as English according to the Twitter API dataset.

In the next step, the data was processed through multiple functions to be ready for final usage, including filtering from spam and duplication. The geolocation data was converted into radians to calculate the distance between points and establish density-based clusters.

For the abnormal analysis method, all emojis were extracted from the text each day of the month and used to calculate the Term Frequency–Inverse Document Frequency (TF-IDF) method, which is used to estimate the frequency and importance and calculate the z-score for each Emoji for the perspective day. The z-score threshold was set, and a search

 ISBN: 978-1-68558-103-9

for a high difference in this value was conducted to search for abnormality.

Lastly, the map clustering method used was Density-based spatial clustering of applications with noise (DBSCAN).

IV. Result

In this research paper, we have divided the results into three parts based on the clustering type and result category we have found.

A. Corona Related Emoji Map Clustering

In this experiment, random emojis were fed to the clustering map to find a unique emoji with a massive presence in one area besides the others. The corona-related Emoji, the mask face (😷), was populated within the US area, which matched a period where the coronavirus spread in another wave.

B. Culture Related Emoji Map Clustering

Another interesting Emoji that popped out was the skull emoji (💀). Its population was focused on African countries on the west coast of Africa. After studying the Emoji, reading the comments, and further research, it is inferred that the meaning of skull emoji has the same meaning as the Emoji with the Face with Tears of Joy Emoji (😂). It is inferred that dark humor in the Emoji type is used to indicate interest, awkwardness, and irony.

C. Abnormal Emoji Trend

In this experiment, TF-IDF was used to search for abnormal usage of all emojis; using a z-score table, the maximum threshold estimated for the period was 4, a very high value that resulted in many low spike points due to having many emojis being used extreme rare cases, most of them are random usage. Three interesting findings were made where an unexpected high point was estimated.

a) Nazar (amulet) emoji 🧿*:* This emoji had a peak on the 30th of 11 It was found mostly in America and India shown in Fig 1 It is a special day related to religion, marriage, and personal matters across all regions; many people are getting married on this day. The combination of religious Indian culture, with an interesting number of days, along with personal matters, made this Emoji spike.

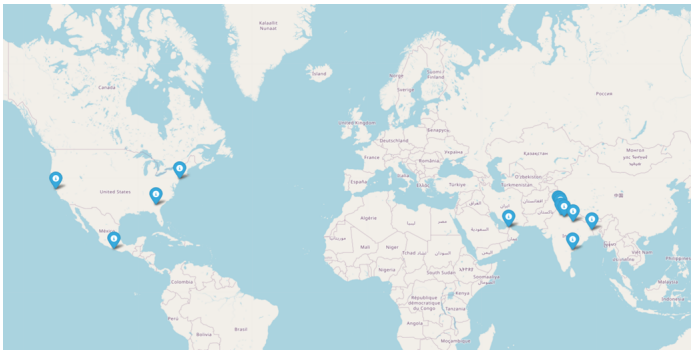


Fig. 1. Clusters mostly found in India and the United States

b) Goat emoji 🐐*:* This Emoji has been recently referred to by its acronym GOAT as Greatest Of All Time, which is typically used for superstars across all fields, including music, art, gaming, and sports. It spiked on the 14th of December, shown in Fig 2, because of the awaiting match between Argentina and Croatia, where Argentina won 3-0 with Lionel Messi opening the scoring and getting the MVP for that match. Messi has been called goat by fans across the world the GOAT, considering him the best player in the world, and he did win the 2022 Qatar World Cup with his team.

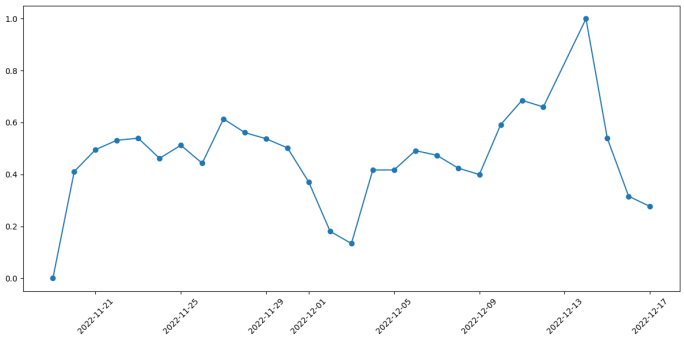


Fig. 2. Sentimental Values of the collected sample

c) Weather emojis 🌀 *and* 🌧️ *:* Both of the emojis have experienced a spike on the 14th of December. Going by the map clustering, the most significant spikes are in two areas: the east coast of the US and the second one is the UK and around the England channel areas. It was reported on the 14th of December that thunderstorms were spotted from New York on the east coast; however, in the UK, it was snowing beautifully, and it was one of the most beautiful days to share on social media. Therefore, the simultaneous events caused the spike in these emojis together.

V. Conclusion and Future Work

Clustering tweets with Geolocation and Emojis have yielded exciting results, from pandemic conditions and weather confirmation to unique traits and global trends. However, this study has been limited to many factors that might have reduced or hindered exciting discoveries.

The selected period was interesting for many reasons, and events happened all at the same time of that year. However, other periods may provide different results that can be interrupted differently.

The data sample was filtered using English language tweets where it is spoken globally. This restriction was necessary because local languages created massive clusters in their countries, causing an imbalance in the results. With that being said, it might yield interesting results when including different languages together, and it will definitely affect the result of the culture-related Emoji map.

An overall point to point out is that people with high-security alerts toward privacy tend to limit access to their location and prevent social media from accessing any related information. Such trivial denial of access can cause such kinds

 ISBN: 978-1-68558-103-9

of studies to lose their opportunity to add value to an important field that could change understanding in NLP studies.

We urge to spread awareness and ask the general people to allow providing location access and show the importance of the studies being done on such data and the current protection provided by the companies where location is not provided accurately for other parties but somewhat slightly adjusted for the provider's safety.

TF-IDF has generated reliable examples, but we were expecting more results; therefore, an investigation for a different abnormal spike detection should be prioritized. Detecting by number count or trend might require training a model to understand the different emojis and types of spikes required for the study.

DBSCAN has generated reliable clusters that could be analyzed and examined. Although other methods were considered, they were deemed inferior due to the outstanding performance of DBSCAN in these kinds of studies. However, this should be reconsidered as new methods are getting updated rapidly.

Overall, if added, automating the process within a long-term period can yield more interesting findings; the next phase will focus on automation and training models that can detect and provide the most interesting results in terms of clustered areas and abnormal changes, discovering unique traits unseen discoveries in the emoji language that is being used on a daily basis.

ACKNOWLEDGMENT

This research was conducted as part of "COVID-19 AI & Simulation Project" run by Mitsubishi Research Institute commissioned by Cabinet Secretariat, JAPAN

REFERENCES

[1] Kaur, S., Kaul, P., & Zadeh, P. M. (2020). Monitoring the dynamics of emotions during COVID-19 using twitter data. Procedia Computer Science, 177, 423–430.

[2] Dyer, J., & Kolic, B. (2020). Public risk perception and emotion on Twitter during the Covid-19 pandemic. Applied Network Science, 5(1).

[3] Kejriwal, M., Wang, Q., Li, H., & Wang, L. (2021). An empirical study of emoji usage on Twitter in linguistic and national contexts. Online Social Networks and Media, 24,

[4] S. C. Guntuku, M. Li, L. Tay, and L. H. Ungar, "Studying cultural differences in emoji usage across the east and the west," Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, no. 1, pp. 226-235, 2019. [Online]. Available: https://doi.org/10.1609/icwsm.v13i01.3224. [Accessed: the 16th of September, 2023].

[5] B. Gao and D. P. VanderLaan, "Cultural influences on perceptions of emotions depicted in emojis," Cyberpsychology, Behavior, and Social Networking, vol. 23, no. 8, pp. 567-570, 2020. [Online]. Available: https://doi.org/10.1089/cyber.2020.0024. [Accessed: the 16th of September, 2023].

[6] P. Kimura-Thollander and N. Kumar, "Examining the 'global' language of emojis: Designing for cultural representation," Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 495-508, 2019. [Online]. Available: https://doi.org/10.1145/3290605.3300725. [Accessed: the 16th of September, 2023]. [Accessed: the 16th of September, 2023].

[7] B. Cohen, "The most popular emoji in each country around the world," The Gate, the 15th of May, 2022. [Online]. Available: https://thegatewithbriancohen.com/the-most-popular-emoji-in-each-country-around-the-world/ [Accessed: the 16th of September, 2023].

[8] M. Kejriwal, Q. Wang, H. Li, and L. Wang, "An empirical study of emoji usage on Twitter in linguistic and national contexts," Online Social Networks and Media, vol. 24, pp. 1-23, 2021. [Online]. Available: https://doi.org/10.1016/j.osnem.2021.100149.

[9] J. J. Schouteten, F. Llobell, S. L. Chheang, D. Jin, and S. R. Jaeger, "Emoji meanings (pleasure–arousal–dominance dimensions) in consumer research: Between-country and interpersonal differences," Journal of Food Science, vol. 88, no. 1, pp. 106-121, 2023. [Online]. Available: https://doi.org/10.1111/1750-3841.16374. [Accessed: the 16th of September, 2023].

 ISBN: 978-1-68558-103-9