Journal of Big Data and Artificial Intelligence

Volume 2 January 2024 Number 1

Journal of Big Data and
Artificial Intelligence

Journal of Big Data and Artificial Intelligence

Volume 2 January 2024 Number 1

EDITORIAL

MEMORIAL

ARTICLES

Journal of Big Data and Artificial Intelligence

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.40

EDITORIAL

A New Era of Artificial Intelligence Begins… Where Will It Lead Us?

Jim Samuel
Bloustein, Rutgers University
jim.samuel@rutgers.edu

Abhishek Tripathi
The College of New Jersey
tripatha@tcnj.edu

Ensela Mema
Kean University
emema@kean.edu

> "By far, the greatest danger of Artificial Intelligence is that people conclude too early that they understand it."
> —Eliezer Yudkowsky

We have entered the age of Artificial Intelligence (AI). Everything around us is becoming artificially intelligent: from business applications to healthcare, education to finance, and governance to art, music, and entertainment. The fact that AI has gripped public attention is evident from the steep rise in public engagement with artificial intelligence applications, explosive increase in news media coverage of AI, increasing volumes of social media posts, and the mushrooming of a range of AI ecosystem initiatives. Consider the steep rise in searches for the terms "AI" and "Artificial Intelligence" from 2022 to 2023 as presented by Google Trends[1] (Figures 1 and 2, https://trends.google.com/trends/). In some senses, this is reminiscent of the big data and data-driven-everything phenomena from around two decades ago. This perspective is also supported by a review of Google Trends for search terms on big data, which shows interest in "big data" peaking around 2014. As with AI, expectations were high then, and concerns, especially around data privacy, were also significant. What many did not realize in those early years was that the big data revolution was a forerunner to what we have just begun to experience—the new era of artificial intelligence, also popularly called the fourth industrial revolution.

Undoubtedly, there have been a series of remarkable breakthroughs in AI technologies over the past few years. However, would these have been possible without big data fueling much of the early opportunities in AI (Kersting and Meyer 2018)? Big data and high-performance computing concepts have served as the foundations for the presently visible waves of AI, driven largely through the use of supervised machine learning methods, and a converse early-stage perspective has been posited with the narrative of how AI methods have been used to create value from big data (O'Leary 2013). Big data and AI have been paired not only in the realm of opportunities but also in risks and challenges

[1] Google Trends provides a relative view to gauge trends and not exact numbers—this means that the peak is treated as 100% in each figure, and figures cannot be compared except for rate of change (https://trends.google.com/trends/).

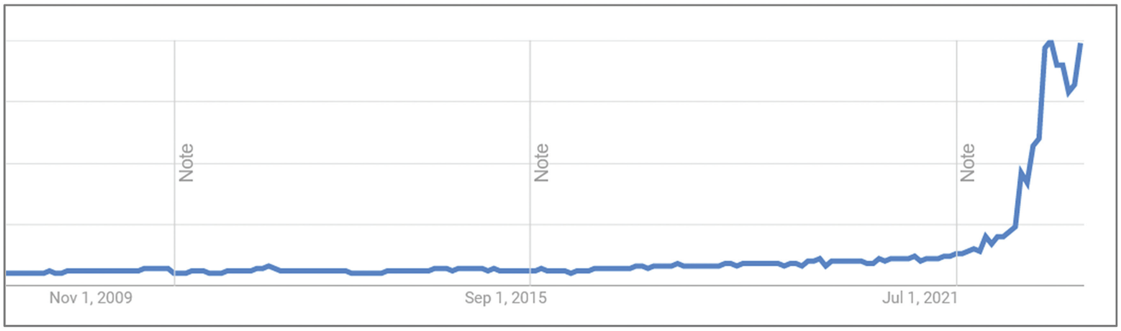


Figure 1: Google search trend for the term “AI.”

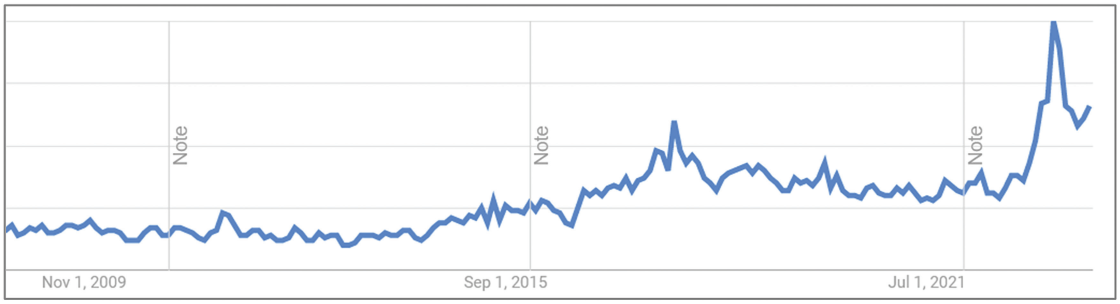


Figure 2: Google search trend for the term “Artificial Intelligence.”

to human society (Helbing et al. 2019). Despite experts like Andrew Ng highlighting the challenges of big data dependence and emphasizing the emergence of “smart data”-driven AI, the fact remains that big data is here to stay as a phenomenon (Strickland 2022). This is clearly evident from recent developments in AI and the rush to develop foundation models and large language models (LLMs) based on modeling vast quantities of data, often quipped to represent “the knowledge of the world.” However, “artificial intelligence” has become a broad and fuzzy term, making it “very difficult to mark its boundaries precisely and specify what exactly it encompasses” (Devedzic 2022).

What is artificial intelligence? A simple yet encompassing definition expands on Samuel (2021) and Samuel et al. (2022): “Artificial Intelligence is a set of technologies that mimic the functions and expressions of human intelligence, specifically cognition, logic, learning, adaptivity and creativity.” Auxiliary topics such as AI risks and AI ethics need to be defined separately as they arise from the potential impacts of AI. There is a need for continued research on various dimensions of AI. As the domain matures, it becomes imperative to develop robust AI frameworks from conceptual, ontological, and “epistemological, philosophical, ethical, technological, and social perspectives” (Devedzic 2022). The era of AI has begun with an explosive expansion that is rapidly influencing nearly every facet of human life and society—it is crucial to recognize that the current wave of AI applications is driven by developments with little philosophical or moral foundations for human interaction with advanced technologies that are simultaneously supersmart and superstupid (absence of consistent commonsense in AI). This is reflected in a number of ethics and transparency concerns surrounding the development of foundation models and LLM applications (Samuel 2023). One of the greatest present needs is the development of a robust philosophy of AI that includes interaction with human intelligence.

AI applications in major categories such as natural language processing (NLP), computer vision (CV), intelligent robotics, intelligent and connected sensors and smart quantitative models, have all used machine learning [mostly supervised, unsupervised and reinforcement learning (RL)] and some form of big data. Even though RL can operate with smaller simulated data, it has been successfully applied to big data. Consider the recent release of state-of-the-art AI foundation models such as PaLM (March 2023, with 540 billion parameters), LLaMA-2 [now freely available for downloads in multiple sizes (LLaMA-2 2023; LLaMA 2-Meta 2023)], and GPT-4 (and GPT-5 in the works), and open-source models such as BLOOM. Generative AI has been researched for over two decades, and extant research has explored generative concepts and applied it to areas such as social media (Sutskever et al. 2011; Reed et al. 2016; Garvey et al. 2021). However, the late 2022 launch of the GPT-3+-based ChatGPT application brought awareness about the power and the usefulness of generative AI to the forefront. Google (Alphabet), as an example, has continued to release new models and applications—Chirp is a cluster of "Universal Speech Models" built on 12 million hours of speech data for automatic speech recognition (ASR); Imagen, Muse and Parti are different types of text-to-image models, while Codey drafts code for humans (Google AI 2023). While we are still absorbing these, in addition to PaLM-2, Alphabet CEO Sundar Pichai announced Google's next superfoundation-model "Gemini," which is reportedly still in training (Pichai 2023). These and similar developments herald the new era of artificial intelligence—it will be wise for us to assume that our understanding is limited and we need substantial research and thought leadership on big data and AI in the years ahead to harness AI effectively for the benefit of human society.

Welcome to the Journal of Big Data and Artificial Intelligence! Keeping in line with this overarching trend towards AI-driven value creation, and the critical need to further catalyze research in artificial intelligence across multiple dimensions and domains, the Executive Editorial Board of this journal has decided to affirm this undeniably increasing importance of AI research in the context of big data: Starting with this volume, the *Journal of Big Data Theory and Practice* has been formally renamed as the *Journal of Big Data and Artificial Intelligence* (***JBDAI***). This move was deliberated over for a significant period of time and went through multiple evaluation cycles and ratification with the parent body, New Jersey Big Data Alliance (NJBDA). The *Journal of Big Data and Artificial Intelligence* now better represents the research it publishes on big data, machine learning, NLP, domain analytics, image processing, LLMs, and other AI research. We also anticipate a significant enhancement of the brand value of the journal and an increase in the attractiveness of the journal to a broader range of researchers. Please join us in this venture to create scholarly thought leadership in big data and AI by contributing at various levels to *JBDAI*.

This volume presents five intellectually stimulating articles, a memorial to one of our key scholars and founding member, and this present editorial—these articles traverse a broad range of topics and methods and we hope along with the authors that these will be useful: The "BERT based Blended approach for Fake News Detection" paper provides an interesting approach to identifying fake news on social media using BERT-LSTM and BERT-CNN. The "Investment under Uncertainty: The Role of Inventory Dynamics" paper demonstrates a novel approach to study the influence of inventory on the value of the firm and on investment decisions. Authors of the "Crime Frequency During Covid-19 and Black Lives Matter Protests" use the Holt-Winters and SARIMA models to explore changes to crime under the pandemic and social unrest conditions. Applying image classification methods, the "Machine Learning Study: Identification of Skin Diseases for Various Skin Types Using Image Classification" paper shows an approach to improving classification accuracy. Finally, the "Are Emotions Conveyed Across Machine Translations? Establishing an Analytical Process for the Effectiveness of Multilingual Sentiment Analysis with Italian Text" uses natural language processing methods within a multilingual context to generate a repeatable process for analyzing text.

In conclusion, it is necessary to highlight the irreversible connection between big data and AI—each of the above mentioned (and more) foundation models and AI applications are built on vast quantities of data, and the multimodal foundation models are each built on multiple types of big data such as text, audio, and images. Undoubtedly, we will witness the development of AIs trained on relatively smaller quantities of "smart data," and other variations. However, we are yet to observe any significant developments signaling the replacement of big data without compromising efficacy. It is more likely that smart data, small data, and all such variations to locally defined data-size optimality will eventually add up leading to a continued net increase in big data volumes and complexities. Big data and AI are, from all that we can presently observe, irreversibly linked together, with AI methods being used to generate insights and value from big data, and big data serving as foundational building blocks for a vast array of AI applications. Furthermore, AI research is rapidly expanding in the social sciences, cultural studies, socioeconomics and education, among other domains (Samuel et al. 2023). We at *JBDAI* hope to encourage and foster much high-quality research, rigor, and innovative thought leadership on big data and artificial intelligence in the years ahead, supporting human well-being, the sustainability of our natural resources, and balanced societal progress—please contribute to *JBDAI* and be a part of this exciting intellectual adventure!

References

Devedzic, V. 2022. "Identity of AI." *Discover Artificial Intelligence* **2**, no. 1: 23. doi: 10.1007/s44163-022-00038-0

Garvey, M. D., J. Samuel, and A. Pelaez. 2021. "Would You Please like my Tweet?! an Artificially Intelligent, Generative Probabilistic, and Econometric Based System Design for Popularity-Driven Tweet Content Generation." *Decision Support Systems* **144**: 113497. doi: 10.1016/j.dss.2021.113497

Google AI. 2023. "Foundation Models." Accessed November 27, 2023. https://ai.google/discover/foundation-models/

Helbing, D., B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. van den Hoven, R. V. Zicari, and A. Zwitter. 2019. "Will Democracy Survive Big Data and Artificial Intelligence?" In *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*, edited by Dirk Helbing, 73–98. Springer: Cham, Switzerland. doi: 10.1007/978-3-319-90869-4_7

Kersting, K., and U. Meyer. 2018. "From Big Data to Big Artificial Intelligence? Algorithmic Challenges and Opportunities of Big Data." *KI - Künstliche Intelligenz* **32**, no. 1: 3–8. doi: 10.1007/s13218-017-0523-7

LLaMA-2. 2023. "Model Download." Accessed November 27, 2023. https://huggingface.co/meta-llama

LLaMA 2-Meta. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models." Accessed November 27, 2023. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

O'Leary, D. E. 2013. "Artificial Intelligence and Big Data." *IEEE Intelligent Systems* **28**, no. 2: 96–99. doi: 10.1109/MIS.2013.39

Pichai, S. 2023. "Google I/O 2023: Making AI More Helpful for Everyone." Accessed November 27, 2023. https://blog.google/technology/ai/google-io-2023-keynote-sundar-pichai/#ai-products

Reed, S., Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. 2016. "Generative Adversarial Text to Image Synthesis." In *International Conference on Machine Learning*, June 19–24, New York, USA.

Samuel, J. 2021. "A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies." Scholars Strategy Network. Accessed November 27, 2023. https://scholars.org/contribution/call-proactive-policies-informatics-and

Samuel, J. 2023. "The Critical Need for Transparency and Regulation amidst the Rise of Powerful Artificial Intelligence Models." Scholars Strategy Network (SSN) Key Findings. Accessed November 27, 2023. https://scholars.org/contribution/critical-need-transparency-and-regulation

Samuel, J., R. Kashyap, Y. Samuel, and A. Pelaez. 2022. "Adaptive Cognitive Ft: Artificial Intelligence Augmented Management of Information Facets and Representations." *International Journal of Information Management* **65**: 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, Y., Brennan-Tonetta, M., Samuel, J., Kashyap, R., Kumar, V., Madabhushi, S. K. K., Chidipothu, N., Anand, I., and Jain, P. 2023. "Cultivation of Human Centered Artificial Intelligence: Culturally Adaptive Thinking in Education for AI (CATE-AI)." *Frontiers in Artificial Intelligence* **6**. https://www.frontiersin.org/articles/10.3389/frai.2023.1198180

Strickland, E. 2022. "Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big." *IEEE Spectrum* **59** (4), 22–50.

Sutskever, I., J. Martens, and G. E. Hinton. 2011. "Generating Text with Recurrent Neural Networks." In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, Washington, USA, June 28–July 2.

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.38

MEMORIAL

In Memory of Dr. David Belanger (12/8/1944–11/18/2022)

Sadly, our dear colleague, Dr. David Belanger, passed away in November, 2022.

David was a founding member of the New Jersey Big Data Alliance (NJBDA)—an alliance of New Jersey academic institutions and corporations that aims to promote Big Data education and research in New Jersey, the parent organization of this journal. "Through the last decade, as our organization grew and expanded its programs, he provided brilliant insight and guidance on our direction, offering suggestions in his thoughtful way and always ready to collaborate. David will be greatly missed," said Margaret Brennan-Tonetta, NJBDA's past president and co-founder. At NJBDA, he was most recently Vice President of the Entrepreneurship Committee.

David was an internationally known authority on Big Data and data governance.

Dr. David Belanger

During his 30 years at AT&T Bell Labs, he contributed to software and data management advances that led to the world of computing and Big Data as we know it today. From 2001 to 2012, he was the Chief Scientist at AT&T.

As Big Data emerged as a significant factor in the industry, David was called upon to lead the IEEE's "Big Data Initiative"—a worldwide effort that included defining data strategy and data standards. He also led the IEEE's DataPort project, which offers 2,500 datasets and has over 2.5 million users.

In 2012, David joined Stevens as Senior Research Fellow/Senior Lecturer. As a BIA faculty member, he developed several courses in the Big Data concentration and certificate programs. David was a superb teacher, and he touched the lives of hundreds of students who looked to him as a source of inspiration throughout their careers. He played a leading role in the BIA Industry Advisory Board, where his presence attracted leading industry experts and animated discussions of trends and issues of importance to the program. David had a positive impact on all of us. He was an inspiring, innovative thinker, role model, and constant reminder of the best academic traditions. He was always interested in exploring new ideas and excited about his research and teaching. He loved his students, was proud of their achievements, and followed them as a friend and mentor through their careers. He inspired all of us through his integrity, high intellectual standards, and collegiality.

In his memory, Stevens has established the David Belanger Fellowship Fund to provide fellowships for future generations of qualified graduate students in the area of Big Data. "Knowing David's love for his students, I cannot think of a more fitting tribute to his memory" said Ted Stohr, Professor and Academic Coordinator of the BI&A Program.

We at NJBDA and JBDAI will continue to remember David as a gentle scholar who cared for people. A colleague fittingly remembered David as being "the kindest scientist of our time."

George Avirappattu
Kean University
gavirapp@kean.edu

Mahmoud Daneshmand
Stevens Institute of Technology
mdaneshm@stevens.edu

Matthew Hale
Seton Hall University
matthew.hale@shu.edu

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Margaret Brennan-Tonetta
Rutgers University
mbrennan@njaes.rutgers.edu

Rashmi Jain
Montclair State University
jainra@mail.montclair.edu

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.27

BERT-Based Blended Approach for Fake News Detection

Shafqaat Ahmad
Penn State University
Sua845@psu.edu

Satish Mahadevan Srinivasan
Penn State University
sus64@psu.edu

Abstract

This paper presents a new approach for detecting fake news on social media. Previous works in this domain have demonstrated that context is an important factor when attempting to distinguish subtle differences within text. Fake news itself presents different level of difficulty due the vast similarity that exists between genuine and fake news contents. Therefore, we propose a collaborative approach which uses probabilistic fusion strategy to combine the knowledge gained from modelling two language models, Bidirectional Encoder Representations Transformers (BERT)-long short-term memory network (LSTM) and BERT-convolutional neural network (CNN). To achieve the fusion, we exploit the Bayesian method. Our experiments are conducted on two fake news detection datasets. The detection accuracy attained in these experiments attest to the efficiency of the proposed method, as our approach is very competitive compared to the state-of-the-art methods.

Keywords *natural language processing, language modelling, deep neural network, machine learning, BERT.*

1. Introduction

The importance of social media nowadays as a medium for human interaction cannot be overemphasized. From mere sharing of memes, posting of pictures and videos, to making live broadcast, most people tend to spend a huge amount of their time daily engaging with contents on social media. This has also become one of the reasons why most people use social media as their main source of news as opposed to the traditional news outlets. Moreover, the option of sharing, commenting, and liking news contents, coupled with the flexibility and speed at which these actions can be performed is another motivating factor changing people's interests from traditional news outlets to social media-based news. Despite the great advantage social media offers, the quality of news on social media is incomparable to traditional news. While the cost of quickly posting news on social media is extremely negligible, a considerable proportion of these news are fake, intentionally prepared to propagate false information and negative agenda (Pérez-Rosas et al. 2017). Some news is propagated for political and financial gains, and some are mainly to divide opinions and create distrust among the public. This inherently changes the belief system of people toward news that are actually genuine (Reis et al. 2019).

Fake news is challenging to detect because the fine line between genuine and fake content is only so obscure, that even for humans, detection of some fake news cannot be 100% guaranteed as context places a huge factor in fake news detection (Sharma et al. 2019). Moreover, attempting to manually detect fake news will lead to a highly laborious process that will at the same time, required huge amount of human resources.

Hence, several research studies have been focusing on development of algorithms and techniques that can be leveraged to automatically detect fake news, particularly on social media contents. This automated approach will save huge amount of time and effort, at time same being more effective. Approaches in machine learning toward fake news detection have typically been based on natural language processing (NLP) techniques, starting from the simple bag-of-words (BoW) representations to more sophisticated, advanced techniques like Word2vec (Zhang et al. 2018; Jang et al. 2019). Exploiting context within text to delineate genuine from fake is difficult, nonetheless, attempts are being made to push the boundaries of NLP toward context understanding. Recently, Bidirectional Encoder Representations Transformers (BERT) have become a revolutionary language model that has become the standard for solving several NLP problems such as machine translation, text summarization, entity extraction and so on (Sanh et al. 2019; Mozafari et al. 2020).

In this work we exploit the representation advantage of BERT, by using the model as an embedding for convolutional neural network (CNN) and long short-term memory network (LSTM). We then further propose a collaborative approach for fusing the knowledge learnt from modelling the two language models as a convoluted hybrid framework. Using BERT-LSTM and BERT-CNN as the baseline models, we propose a fusion method that relies on Bayesian theorem for score level fusion to boost fake news detection performance. Our idea is established on the fact that subtle differences that cannot be captured by a single model can be complemented for, using a hybrid model.

2. Related Works

Solving the problem of fake news detection using datasets retrieved from social media platform has been approached from different perspectives. Some research works have focused on extracting features based on solely on the content of the news articles, while others explore social context to perform feature extraction (Shu et al. 2019a).

2.1 Approaches Based on News Content Features

A lot of research works are exploring raw metainformation such as the title, body, headline, news source, and possibly digital information like videos and images to extract features (Shu et al. 2019b). Since fake news are usually intention coined to mislead the general public, it is often possible that the common words or phrases in these news are written to attract more clicks, views, and likes from social media (Shu et al. 2017). In other words, making the news go viral. Meanwhile such words or phrases can also turn out to be an advantage when extracting features, as lexical level features which are mainly words and characters can be represented using text vectorization algorithms and syntactic level features which mainly consist of sentences, statements, and phrases can be extracted using algorithms based on BoW and word2vec (Guthrie et al. 2006; Wallach 2006; Goldberg and Levy, 2014).

Furthermore, with the recent advances in deep neural networks, language models based on recurrent encoder-decoder networks have also been similarly explored, with techniques such as LSTM (Hochreiter and Schmidhuber, 1997), BERT (Devlin et al. 2018), Generalized Autoregressive Pretraining for Language Understanding (XLNET; Yang et al. 2019) making significant breakthrough in this domain. Pretrained large cased BERT model has been directly applied to detect the possibility of an author on social media to spread fake news (Baruah et al. 2020). Similarly, Rodríguez and Iglesias (2019) applied different architecture of pretrained BERT on fake news detection.

We have seen a few cases where BERT has been integrated with other linear and deep neural network models. For instance, Wu and Chien (2020) combined pretrained BERT with linear classifier by basically using the BERT model to capture the text representation and performing classification of fake news spreaders using a linear classifier. Kula et al. (2019) proposed integrating BERT model with recurrent neural network (RNN) for fake news detection. Kaliyar et al. (2021) propose fake BERT model which combines different parallel components of CNN having different kernel sizes and filters with BERT. The goal of using such an approach is to minimize the effect of ambiguity on text understanding.

BERT has somewhat attracted a considerable amount of attention in fake news detection, either using the model directly or exploring it combination with other machine learning models. However, to the best of our knowledge, we have not seen a situation where a hybrid combination of two BERT-based deep neural network models have been explored for fake news detection. Moreover, there are no reports of using fusion strategies to combine BERT model with other model, particularly probabilistic score fusion methods.

Hence, in this paper, we the follow the approach of learning representation from news contents and our key contributions are summarized as follow:

- We use BERT as an embedding for training 1D CNN and LSTM.
- We introduce using Bayesian model for performing score level fusion of BERT-CNN and BERT-LSTM.
- We assess the performance of the proposed methods under different parameter settings.
- Finally, we evaluate the performance of the proposed hybrid model, and demonstrate that it performs better than other models that use only BERT, word2vec, or BoW techniques.

3. Proposed Method

This section describes the propose learning methods for achieving fake news detection. The architecture is illustrated in Figure 1, where we follow a couple of processes including text preprocessing, text representation using languages models and postclassification score fusion using Bayesian method.

In the text preprocessing stage, we used regular expressions to remove unnecessary or unwanted characters such as hashtags, punctuations, numbers, html tags, and so on.

After obtaining the cleansed text, we pass the text to the base language models which are explained in the subsequent sections. The models perform both text representations with pretrained embedding and classification of texts into fake or genuine classes. The resulting classification scores are finally passed to Bayesian model for score fusion to obtain the final classification results.

4. BERT

BERT (Devlin et al. 2018), initially developed by Google, has its origins from pretraining contextual representations learning in NLP. It can handle NLP tasks such as supervised text classification, question-answering, text summarization, without the need for human intervention. This technique is widely popular in academics and industry because of its versatility in dealing with any corpus while producing excellent results. BERT is also an encoder-decoder type of model but adds an attention mechanism which performs mapping of a query and a set of key-value pairs to an output helping the model to maintain the relative importance of input (Liu et al. 2019). BERT has two steps, which are: pretraining and fine-tuning. In the pretraining stage, the model is trained on unlabeled data over different pretraining tasks. During fine-tuning, the BERT model is finetuned by first initializing it with the pretrained parameters using Masked LM and Next Sentence Prediction (NSP) and then fine-tuning all of the parameters using labeled data from the downstream tasks, fine-tuning is straightforward since the self-attention mechanism in the transformer allows BERT to model many downstream tasks.

In this work, we use the pretrained uncased Small BERT provided by huggingface with $L = 4$ hidden layers (i.e., Transformer blocks), a hidden size of $H = 256$, and $A = 4$ attention heads. Additional information about the pretrained uncased Small Bert can be obtained from this link https://tfhub.dev/tensorflow/small_bert/bert_en_uncased_L-4_H-256_A-4/2.

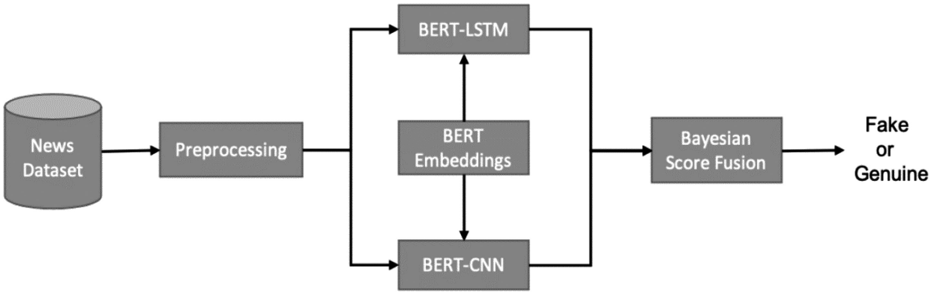


Figure 1: Illustration of the proposed system.

4.1 BERT-CNN

CNN is a one of the most successful deep learning algorithms which combines feature learning with trainable classifiers. It usually has multilayers of convolution, nonlinear transformation, pooling operation, and a fully connected network (FCN).

In CNN, convolutional layers perform operations of convolution on the input data by applying a set of filter banks with varying properties to generate some feature maps. This operation is performed in an iterative manner and the resulting feature maps from a preceding layer are transferred to the subsequent layers to learn more features via convolution. These feature maps are then approximated using an activation function, with common examples are sigmoid and rectified linear units (ReLU). Between two convolutional layers, we normally perform pooling operation to obtain features with strong affinity, which also indirectly eliminates weaker features in the feature map, and at the same time reduces the size of the feature map by replacing the values in a particular region with the statistical summarization of its neighbors.

Two of the most popular techniques are max pooling which replaces values of the feature map with the max value, and average pooling that simply computes the average of the feature map. Another operation generally applied in the learning process is regularization, and the dropout method has proven to be the most successful technique. The final layers of the network consist of FCN and loss layer. The FCN connects every single neuron in one layer to that of another layer, while the loss layer is used for making predictions.

In order to train CNN with BERT, we use the pretrained uncased BERT model embedding vectors as the embedding layer in our CNN, and the CNN model further has three convolutional layers, one global max pooling layer, one dropout layer, and one fully connected layer. We generally used ReLU activation function and sigmoid in the fully connected layer. The loss function is binary crossentropy and ADAM optimizer.

4.2 BERT-LSTM

LSTM has been introduced as a variant of RNN which incorporates memory units into the network. This effectively allows the network to determine the instances to forget previous hidden states or when to update hidden states when new data is fed into the network.

Standard RNN in its most conventional form aims to construct a model with temporal dynamics flow by mapping sequential input data to a hidden state. The hidden states are then mapped to outputs which can be expressed with the following Equation (1), given an input sequence data X:

$$\begin{gathered} h_s = f(W_{xh}X_s + W_{hh}h_{s-1} + b_s) \\ z_s = f(W_{hz}h_s + b_z), \end{gathered} \tag{1}$$

where f is a nonlinear activation function computed elementwise, h_s is the hidden state, W is the weight, b is the bias, and z_s is the output at time s. One of the major challenges of RNN is the inability to remember interaction in long-term sequence due to the problem of exploding gradients. As a result, LSTM.

For us to train LSTM with BERT, we use the pretrained uncased BERT model embedding vectors also as the embedding layer in the LSTM model and the LSTM has one spatial dropout layer, one recurrent layer, and two dense layers. The hyperparameters of the LSTM model were fine-tuned to obtain the optimal values. Similarly, the ReLU and sigmoid are used in the dense layers, and the loss is binary cross entropy and ADAM optimizer.

5. Fusion Strategy

In order to fuse the classification scores from the language models we used these techniques:

$$\text{Sum Rule}: \text{ FS} = \sum_{i=1}^{n} s_i. \tag{2}$$

$$\text{Weighted Sum Rule}: \text{ Weighted} = \sum_{i=1}^{n} w_i s_i, \tag{3}$$

where, s_i is predicted scores of a classifier and w_i is a weight value assigned to the classifier based on recognition performance.

Bayesian Fusion: this is a probabilistic approach for fusing classification scores originating from the language models. Similar approach has recently been used for fusing multiple sensors in Chen et al. (2021). In this case, we are able to compensate for the inadequacy of a model with another model. Assume we have fake news labels y, associated to each language model BERT-CNN s_1 and BERT-LSTM s_2 and that the scores from the models are conditionally independent of the news labels y. Hence, we can write the Bayesian expression as:

$$P\left(s_{1,},s_2|y\right) = P(s_1|y)P(s_2|y), \tag{4}$$

which is also equivalent to:

$$P(s_1|y) = P(s_1|s_2,y). \tag{5}$$

We can notice that conditional independence is valid in the equation because given a news label y, predicting the BERTCNN score s_1 will not have any influence on the knowledge gained from BERT-LSTM s_2.

Hence, applying Bayes rule to the above expression will result in:

$$P\left(y|s_{1,},s_2\right) = P\left(s_{1,},s_2|y\right)P(y)\,P\left(s_{1,},s_2\right). \tag{6}$$

6. Experimental Results

Experiments were conducted on two benchmark datasets, retrieved from Kaggle, which are described below. Experimental environment was setup on Google Colaboratory (also known as Colab). The python scripts used for executing the experiments discussed in this manuscript are available in a GitHub repository. The link to the GitHub repository is https://github.com/shafqaatahmad/BERT_based_Blended_approach_for_Detecting_FakeNews.

6.1 Dataset

UTK Machine Learning Club: This dataset was downloaded from Kaggle instigated by the UTK Machine Learning Club 2. It mainly consists of 20,800 news articles which are fairly distributed equally between the two classes namely genuine and fake news. However, out of the 20,800 news articles we randomly selected only 4,000 samples for this experiment. Due to the limitation in the availability of computational resources we decided not to use the entire dataset for this experiment. The distribution of the samples selected for this experiment are shown in Figure 2. In order to build the model, we split the data into 70% training set and 30% test set.

WELFake Dataset: This is a publicly available dataset for fake news detection recently published by Verma et al. (2021). The dataset contains approximately 72,000 news articles, which is a combination of four different datasets: Kaggle, McIntire, Reuters, and BuzzFeed. The dataset is compiled with the aim removing bias or class imbalance from the two news categories. The resulting dataset is 48.55% genuine news and 51.45% fake news.

Due to the large volume of the dataset and limited access to larger computing resources, we randomly select 3,575 samples from the dataset for our experiments. The data distribution is depicted in Figure 3.

6.2 Results

1) *UTK Machine Learning Club*: conducting the experiments on this dataset using the individual algorithms: BERT-LSTM and BERT-CNN, we achieved a prediction rate of 91.09% and 79.2%, respectively, as shown in

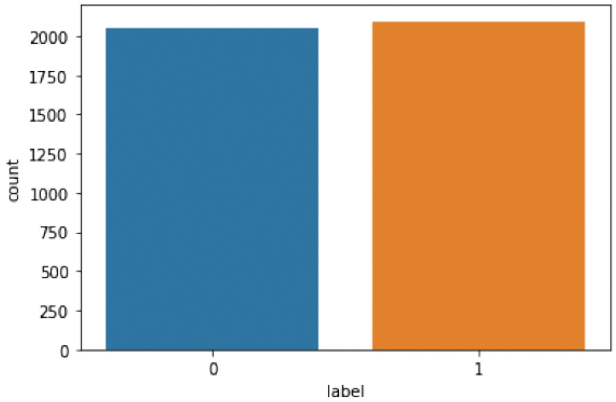


Figure 2: Distribution of samples selected from UTK dataset. 1 = Genuine, 0 = Fake.

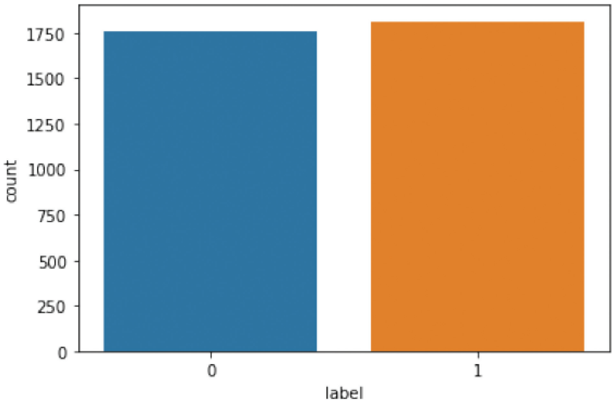


Figure 3: Distribution of samples selected from WELFake dataset. 1 = Genuine, 0 = Fake.

Table 1. In this experiment a 90% weightage was assigned to the classification output of the BERT-CNN model and the remaining 10% weightage was assigned to the classification output of the BERT-LSTM model.

Looking at the results, we did not notice a trend where Genuine class is solely predicted with higher accuracy than the Fake class and vice versa for both language models. As mentioned in the Introduction, the subtle differences between genuine and fake news could be very marginal and that it is very likely that the words used to coin fake news contents are similar to those in genuine contents. However, with the use of language models which are able to properly represent words and understand the context around those words, we can notice that both classes of news are well predicted with considerably high accuracy.

We also tried different architectures of BERT-CNN and BERT-LSTM by basically changing the parameters of the CNN and LSTM models. However, we observed similar trend in classification accuracies of the two models. Nonetheless, we attempted using our proposed score fusion method Bayesian formulation. From the experiments, we observed an improvement in performance by at least 1.5%, yielding a prediction rate of 96.7% as shown in Table 1.

Furthermore, this approach helps in bridging the gap between the two individual language model, moving the accuracy of Fake news to 97.2% and Genuine to 94.1% as opposed to the low classification rates attained using BERT-LSTM. This shows that it is quite beneficial to apply score fusion to compensate inadequacy of a model with another model. Furthermore, we attempted using the standard SUM rule for score fusion, which also produced impressive performance but not as effective as Bayesian method.

2) *WELFake*: Using this dataset, we follow the same experimental setting used on UTK dataset in terms of training/ test split, and CNN/LSTM architecture. With BERT-CNN, we attained a prediction rate of 94.5% as shown in Table 2. In this experiment a 50% weightage was assigned to the classification outputs of both the BERT-CNN and BERT-LSTM models.

In the case of BERT-LSTM, we attained a prediction rate of 80.7% as shown in Table 2. We also noticed that the BERT-CNN generally outperformed BERT-LSTM in this experiment. Also, similar classification trend can be observed where neither of the two classes significantly attained higher accuracy than the other class in all cases.

Meanwhile, using the proposed Bayesian score fusion method, the performance of the model improved to 96.2%, with the fake class increasing to 95.3%. Using SUM rule, we attained an accuracy of 92.8%. This also corroborates the point that the inadequacy of a single model can be compensated by using a hybrid model.

Table 1: Results of experiments on UTK dataset.

Methods	Genuine	Fake	Total
BERT-CNN	90.27	91.2	91.09
BERT-LSTM	80.53	78.9	79.2
SUM rule	77.8	96.2	92.9
Bayes fusion	**94.1**	**97.2**	**96.7**

Table 2: Results of experiments on WELFake dataset.

Methods	Genuine	Fake	Total
BERT-CNN	95.3	93.8	94.5
BERT-LSTM	91.9	69.6	80.7
SUM rule	89.6	95.9	92.8
Bayes fusion	**96.4**	**95.3**	**96.2**

Table 3: Results of experiments on UTK dataset.

Papers	Methods	Results (%)
Trivedi et al. (2021)	BERT	83
Barbosa et al. (2020)	MLP	96.4
Agarwal et al. (2020)	CNN-LSTM	94.7
Our proposed	**Bayes fusion**	**96.7**

6.3 Statistical Analysis

In this study we have considered two statistical techniques namely the Kruskal–Wallis test and the pairwise Wilcox test. Kruskal–Wallis is a nonparametric method for testing whether samples are from the same distribution. The null hypothesis of the Kruskal–Wallis test is that the mean ranks of the groups are the same. Kruskal–Wallis test is roughly equivalent to one-way ANOVA on ranks. The nonparametric Kruskal–Wallis test does not assume a normal distribution of the underlying data. Thus, Kruskal–Wallis test is more suitable for analysis of dataset where the sample size is small (<30). For the dataset that is not normally distributed and contain some strong outliers, it is more appropriate to use ranks rather than actual values to avoid the testing being affected by the presence of outliers or by the nonnormal distribution of data. This test also assumes that the observations are independent of each other.

The pairwise Wilcox test is performed as a *post hoc* test to determine which groups are different from other groups. Upon randomly varying the sample size of both the datasets we created 17 different samples from both the datasets. On the 34 sample datasets we applied the BERT-CNN, BERT-LSTM and our proposed Bayesian score fusion algorithm. The detection accuracy attained across all the samples of the dataset were analyzed using the Kruskal–Wallis nonparametric test. This test was performed to determine whether there is a significant difference in the performance of these algorithms. On the 17 samples of instances obtained from the UTK Machine Learning Club dataset, we observed a significant difference in the mean detection accuracy of the three algorithms at $a = 0.05$. The Kruskal–Wallis nonparametric test resulted in a p-value of 1.159e-08. In addition to that we also performed a pairwise Wilcox test to determine which group of algorithms differed from each other in terms of the detection accuracy of fake news. At $a = 0.05$, we observed a significant different in the detection accuracy of our proposed Bayesian score fusion algorithm with p-value of 0.008 and 2.6e-09 against the BERTCNN and BERT-LSTM, respectively. Also, there was a significant difference in the detection accuracy of the BERT-CNN against the BERT-LSTM (p-value = 1.1e-06, $a = 0.05$). On the 17 samples of instances obtained from the WELFake dataset, we observed a significant difference in the mean detection accuracy of the three algorithms at $a = 0.05$. The Kruskal–Wallis nonparametric test resulted in a p-value of 1.798e-08. In addition to that we also performed a pairwise Wilcox test to determine which group of algorithms differed from each other in terms of the detection accuracy of fake news. At $a = 0.05$, we observed a significant different in the detection accuracy of our proposed Bayesian score fusion algorithm with p-value of 0.024 and 1.1e-06 against the BERT-CNN and BERT-LSTM, respectively. Also, there was a significant difference in the detection accuracy of the BERT-CNN against the BERT-LSTM (p-value = 2.6e-09, $a = 0.05$).

6.4 Performance Comparison

Finally, in order to ascertain the effectiveness of the proposed method, we compared its performance with existing techniques in the literature which have been implemented on UTK and WELFake datasets as shown in Tables 3 and 4.

From the comparisons it can be noticed that our proposed Bayesian based score fusion method is highly competitive with existing methods. In most cases, we outperformed existing techniques by at least 3% increase in classification accuracy.

Table 4: Performance comparison on WELFake dataset.

Papers	Methods	Results (%)
Verma et al. (2021)	CNN	92.48
Verma et al. (2021)	BERT	93.79
Verma et al. (2021)	TFIDF+SVM	96.7
Our proposed	**Bayes fusion**	**96.2**

6.5 Internal and External Threats to Validity

Here, there are no threats to internal validity of this study since both the datasets analyzed in this study are large-scale datasets containing both real and fake news obtained from both the trustworthy and untrustworthy sources, respectively. More precisely, the trustworthiness of the source has been used as a proxy for the real labels. It is quite possible that both these data sets may suffer from false positives (since untrustworthy sources can spread a mix of real and fake news), and false negatives (false information spread by trustworthy sources, e.g., by accident). However, collecting all the news from a specified set of sources over a period of time mitigates the problems of biases in the dataset. There might be some threat to the external validity of this research because here we reported the performance measures of all the algorithms by comparing the predicted labels from the algorithms with the actual labels i.e., real, or fake which are indeed low-quality labels as the dataset is not manually curated.

7. Conclusion

In this paper, we have presented a new approach for detecting fake news from news posted on social media. We proposed using a probabilistic fusion strategy to combine the knowledge gained from two language models BERT-CNN and BERT-LSTM, at a classification score level. The experiments on these methods were conducted on two benchmarked datasets. Under varying parameter settings, the detection accuracy attained supersede the existing fake news detection methods by at least 3%.

The core objective of the fake news detection system is to effectively monitor and counter the dissemination of misleading content and misinformation with the potential to manipulate public opinion, thoughts, and behaviors on a societal level. Some real-world applications where fake news detection system can be implemented includes, Social Media Platforms, News Organizations, Government Agencies, Public Awareness Campaigns, and Fact-Checking Services.

The detection of fake news is an evolving field with very vast future potential. A particularly intriguing avenue for future research is the integration of multiple data modalities, including images and videos alongside textual data, to create advanced multimodal fake news detection system. Such an approach holds promise for detecting manipulated media, deepfakes, and complementing textual misinformation identification. Additionally, the incorporation of user feedback emerges as another fascinating enhancement for improving the accuracy and reliability of fake news detection system proposed in this paper.

References

Agarwal, A., M. Mittal, A. Pathak, and L. M. Goyal. 2020. “Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning.” *SN Computer Science* **1**, no. 3: 143. doi: 10.1007/s42979-020-00165-4

Barbosa, V., Carina de Oliveira, and B. B. Reinaldo. 2020. “AuFa-Automatic Detection and Classification of Fake News Using Neural Networks.” *8th International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE 2020)*, Universidad Autonoma De Yucatan, Cancun, January 27–29.

Baruah, A., K. A. Das, F. A. Barbhuiya, and K. Dey. 2020. “Automatic Detection of Fake News Spreaders Using BERT.” CLEF (Working paper). https://ceur-ws.org/Vol-2696/paper_237.pdf

Chen, Y. T., J. Shi, C. Mertz, S. Kong, and D. Ramanan. 2021. “Multimodal Object Detection via Bayesian Fusion.” Preprint, submitted July 2022. arXiv Preprint arXiv:2104.02904. https://arxiv.org/pdf/2104.02904.pdf

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. “Bert: Pretraining of Deep Bidirectional Transformers for Language Understanding.” Preprint, submitted May 2019. arXiv:1810.04805. https://arxiv.org/pdf/1810.04805.pdf

Goldberg, Y., and O. Levy. 2014. “word2vec Explained: Deriving Mikolov *et al.*’s Negative-Sampling Word-Embedding Method.” Preprint, submitted Feb 2014. arXiv Preprint arXiv:1402.3722. https://arxiv.org/pdf/1402.3722.pdf

Guthrie, D., B. Allison, W. Liu, L. Guthrie, and Y. Wilks. 2006. “A Closer Look at Skip-Gram Modelling.” *LREC* **6**: 1222–1225.

Hochreiter, S., and J. Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* **9**, no. 8: 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Jang, B., I. Kim, and J. W. Kim. 2019. “Word2vec Convolutional Neural Networks for Classification of News Articles and Tweets.” *PloS One* **14**, no. 8: e0220976. doi: 10.1371/journal.pone.0220976

Kaliyar, R. K., A. Goswami, and P. Narang. 2021. “FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach.” *Multimedia Tools and Applications* **80**, no. 8: 11765–11788. doi: 10.1007/s11042-020-10183-2

Kula, S., M. Choraś, and R. Kozik. 2019. “Application of the BERT-Based Architecture in Fake News Detection.” In *Computational Intelligence in Security for Information Systems Conference*, Herrero, A., Cambra, C., Urda, D., Sedano, J., Quintian, H., Corchado, E., editors, Cham, Switzerland: Springer, 239–249.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. 2019. “Roberta: A Robustly Optimized Bert Pretraining Approach.” Preprint, submitted July 26. arXiv Preprint arXiv:1907.11692

Mozafari, M., R. Farahbakhsh, and N. Crespi. 2020. “Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model.” *PloS One* **15**, no. 8: e0237861. doi: 10.1371/journal.pone.0237861

Pérez-Rosas, V., B. Kleinberg, A. Lefevre, and R. Mihalcea. 2017. “Automatic Detection of Fake News.” Preprint, submitted August 23. arXiv Preprint arXiv:1708.07104

Reis, J. C., A. Correia, F. Murai, A. Veloso, and F. Benevenuto. 2019. “Supervised Learning for Fake News Detection.” *IEEE Intelligent Systems* **34**, no. 2: 76–81. doi: 10.1109/MIS.2019.2899143

Rodríguez, A. I., and L. L. Iglesias. 2019. “Fake News Detection Using Deep Learning.” Preprint, submitted September 29. arXiv Preprint arXiv:1910.03496

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. “DistilBERT, a Distilled Version of BERT: smaller, Faster, Cheaper and Lighter.” Preprint, submitted October 2. arXiv Preprint arXiv:1910.01108

Sharma, K., F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu. 2019. “Combating Fake News: A Survey on Identification and Mitigation Techniques.” *ACM Transactions on Intelligent Systems and Technology* **10**, no. 3: 1–42. doi: 10.1145/3305260

Shu, K., D. Mahudeswaran, and H. Liu. 2019a. “FakeNewsTracker: A Tool for Fake News Collection, Detection, and Visualization.” *Computational and Mathematical Organization Theory* **25**, no. 1: 60–71. doi: 10.1007/s10588-018-09280-3

Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu. 2017. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD Explorations Newsletter* **19**, no. 1: 22–36. doi: 10.1145/3137597.3137600

Shu, K., S. Wang, and H. Liu. 2019b. “Beyond News Contents: The Role of Social Context for Fake News Detection.” *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Melbourne, VIC, Australia, February 11–19.

Trivedi, A., S. Alyssa, M. Prathamesh, M. Subhiksha, P. D. Meghana, M. Malvika, B. Meredith, S. Arathi, J. Ashish, and D. Rahul. 2021. “Defending Democracy: Using Deep Learning to Identify and Prevent Misinformation.” Preprint, submitted June 3. arXiv Preprint arXiv:2106.02607

Verma, P. K., A. Prateek, A. Ivone, and P. Radu. 2021. “WELFake: Word Embedding over Linguistic Features for Fake News Detection.” *IEEE Transactions on Computational Social Systems* **8**, no. 4: 881–893. doi: 10.1109/TCSS.2021.3068519

Wallach, H. M. 2006. “Topic Modeling: Beyond Bag-of-Words.” *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, June 25–29. doi: 10.1145/1143844.1143967

Wu, S. H., and S. L. Chien. 2020. “A BERT Based Two-Stage Fake News Spreader Profiling System.” CLEF (Working paper). https://ceur-ws.org/Vol-2696/paper_177.pdf

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. “Xlnet: Generalized Autoregressive Pretraining for Language Understanding.” *NIPS’19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver BC, Canada, December 8–14.

Zhang, S., Y. Wang, and C. Tan. 2018. “Research on Text Classification for Identifying Fake News.” *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Jinan, China, December 14–17. doi: 10.1109/SPAC46244.2018.8965536

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.28

Investment under Uncertainty: The Role of Inventory Dynamics

Xue Cui
Shenzhen University
xcuiaa@szu.edu.cn

Sudipto Sarkar
McMaster University
sarkars@mcmaster.ca

Chuanqian Zhang
William Paterson University
zhangc4@wpunj.edu

ABSTRACT

Finished-good inventory is very common under market uncertainty. We built a continuous-time model to study how the inventory will impact a firm's value and investment decisions. Our model shows that the value of a company that followed the optimal inventory policy can be significantly higher than the traditional non-inventory company, particularly if the inventory-holding cost is not large. This premium becomes smaller as the holding cost is increased and is larger when demand is volatile, and when price elasticity is large. We also show that the optimal investment size can be significantly larger than the traditional non-inventory firm, particularly when the inventory-holding cost is low, demand volatility is high, and price elasticity is low. This paper develops a simulation algorithm to solve an iterative optimization problem in a path-dependent economy.

Keywords *inventory dynamics, real options, corporate investment.*

1. Introduction

A number of papers have used the contingent-claim model for valuation of companies and to identify optimal investment decisions under uncertainty. In these papers, uncertainty is introduced by means of a stochastic underlying variable such as revenue or output price or demand strength, which follows an exogenously specified random process. The firm's valuation and the investment decisions are based on this uncertain state variable.

These papers all assume that the firm sells all of its output when it is produced and that there is no possibility of maintaining any inventory of the output; thus, they ignore any effect of inventory management on firm performance. Examples include Huberts et al. (2015), Jou and Lee (2008), Mauer and Ott (2000), Miao (2005), Dangl (1999), Bar-Ilan and Strange (1999), and many others. However, empirical evidence indicates that inventory management does in fact have a significant impact on the firm's performance, for example, Basu and Wang (2011), Elsayed (2015), Koumanakos (2008), Kroes and Manikas (2018), and Ndubuisi et al. (2020). In this paper, therefore, we address the question: how does (optimal) inventory management impact the value and the optimal investment decision of the firm?

A couple of papers have looked at the relationship between inventory and capital investment. Pindyck (1982) examines the impact of uncertainty on investment (capital stock) with and without the possibility of output inventory. However, his paper looks at incremental adjustments to capital stock and is therefore an adjustment-cost model; he shows that the directional impact of uncertainty on capital is the same with and without inventory, although the size (sensitivity) of the adjustment is smaller with inventory. Kim (2020) shows that a firm's inventory dependence (i.e., the importance of inventory in the firm's operations) makes capital investment less sensitive to important economic measures such as firm performance, industry growth, and uncertainty.

The other papers with inventory are limited to raw-material or work-in-process inventory, not output inventory, as in our paper. For instance, in Bianco and Gamba (2019), the firm uses input inventory as an operational hedge for risk management to mitigate the price risk of input materials. Cortazar and Schwartz (1993) examine the valuation of a company with a two-step manufacturing process with a work-in-process inventory.

Our paper examines a firm that has the ability to maintain inventory of the output by comparing it with the traditional model's no-inventory firm. We focus on how much this ability to maintain inventory is worth when used optimally, how this premium in value (relative to the no-inventory firm) is affected by various economic parameters, and how the ability to maintain inventory impacts the firm's investment (size and timing) decision. The firm's policy might be set so as to maximize the current profit or firm's value (because both seem to be used in practice), and we consider both scenarios.

The main results are as follows. First, if inventory holding cost is small, then the firm value with inventory can be substantially larger than in the traditional (no-inventory) models. Second, the behavior of the firm's value with respect to investment size is quite different for a firm with inventory and one without inventory; hence, the conclusions of the traditional literature with regard to investment size might have to be modified for the realistic situation of a firm with inventory. Third, if an inventory-carrying firm's inventory policy is based on maximizing profits rather than value, it might end up with a firm's value that is below the benchmark (no-inventory) firm, particularly if the demand elasticity is large and the demand level is small; thus, maximizing the firm's profits might end up causing destruction of the firm's value.

The rest of the paper is organized as follows. Section 2 describes and derives the model with and without the ability to maintain inventory. Section 3 presents the results of the model. Section 4 summarizes and concludes.

2. The Model

A firm has a production facility whose size is given by the amount of capital K. This facility allows it to produce Q units of the output per unit time, where Q is given by $K = Q^{1/\delta}$; the exponent δ can be viewed as the returns-to-scale of the technology used. The output is sold in the product market, at a price given by

$$p = \theta - \gamma q, \tag{1}$$

where q is the amount sold, θ is the random/stochastic strength of demand (or demand shock), and γ is the sensitivity of the price to the amount sold (or the elasticity of demand). This price process is commonly used in the literature to represent the output's demand curve (Aguerrevere 2003; Dangl 1999; Huberts et al. 2015). The strength of demand θ introduces uncertainty in the model and is assumed to follow the lognormal process:

$$d\theta = \mu\theta dt + \sigma\theta dZ, \tag{2}$$

where μ and σ are the trend and volatility of the demand process and Z is a standard Wiener Process.

We assume that the plant always operates at full capacity, that is, the production rate is always Q, as in, for example, Bar-Ilan and Strange (1999) and Dangl (1999). This is partly because of analytical tractability; if the firm could vary the output rate, it would make the analysis much more complicated. However, this is also a common modeling assumption, because it is a reasonable description of many real-world process industries such as paper, chemicals, etc. (Lederer and Mehta 2005). Moreover, it is consistent with the "price postponement with clearance" argument of Van Mieghem and Dada (1999). Finally, in many industries the firms make the production plans before the actual realization of market demand, and many firms find it difficult to produce below full capacity because of commitments to suppliers and because of fixed costs associated with flexibility (Goyal and Netessine 2007).

Finally, the cost of investing in the production plant has both a constant component and a component that is increasing linearly in the investment size (amount of capital, K). Let this investment cost be $I(Q) = m_0 + m_1 K = m_0 + m_1 Q^{1/\delta}$, where m_0 is the fixed investment cost and m_1 is the variable cost of investment per unit of capital (recall that capital is $K = Q^{1/\delta}$).

Because we are studying the effect of maintaining output inventory, we study both cases: the firm with and without inventory. We call the latter the "inflexible" or "benchmark" firm because it is this type of firm that has been examined in the literature so far. Also, a firm with the ability to maintain inventory might have one of two objectives when making its inventory decisions; its objective could be to maximize either the firm's profit for the period or the value of the firm. A quick look at the production economics literature indicates that, in many (if not most) cases, the objective is specified as profit-maximization; however, in the finance literature, the objective is value-maximization. We, therefore, examine both cases for the inventory-holding firm.

For high price-sensitivity γ, the profit-maximizing firm might sell a smaller amount to keep prices high (so as to maximize the current profit); however, this will increase the future inventory level and drive up inventory costs and thereby reduce the firm's value. Because it acts myopically in ignoring future costs when making decisions, we call the profit-maximizing firm a "myopic" firm. The value-maximizing firm, however, behaves strategically in considering the overall effect of its decisions on the value of the firm (taking into account all costs), hence we call it a "strategic" firm. Thus, we analyze three different firms: (i) benchmark firm (no inventory), (ii) myopic firm (with inventory), and (iii) strategic firm (with inventory).

2.1. Benchmark Firm Valuation

The benchmark firm produces and sells at a rate of Q units per unit time, that is, $q = Q$, then, its profit flow is given by

$$\pi(\theta) = pq - cq = (\theta - \gamma q)q - cq = \theta Q - (\gamma Q + c)Q \tag{3}$$

Then, the value of the plant is given by

$$V(\theta,\, Q) = E_0^Q\left[\int_0^T e^{-r\tau}\pi_\tau d\tau\right] = \frac{1 - e^{-(r-\mu)T}}{r-\mu}\theta Q - \frac{1 - e^{-rT}}{r}(\gamma Q + c)Q, \tag{4}$$

where T is the remaining life of the project.

This is the standard approach in the literature, and the above valuation is consistent with the existing models, for example, Bar-Ilan and Strange (1999). If the firm also chooses the size of the investment optimally, then it will maximize the above firm value at the time of investment less the investment cost, that is, $Q^* = \text{argmax}_Q\left\{V(\theta_0,\, Q) - \left(m_0 + m_1 Q^{1/\delta}\right)\right\}$, where θ_0 is the demand strength at the time of investment. This maximization gives the following:

$$Q^* = \frac{\theta_0 r\left(1 - e^{-(r-\mu)T}\right)}{2\gamma(r-\mu)\left(1 - e^{-rT}\right)} - \frac{c}{2\gamma} - \frac{r m_1 Q^{*\frac{1-\delta}{\delta}}}{2\gamma\delta\left(1 - e^{-rT}\right)} \tag{5}$$

2.2. Myopic Firm Valuation

Suppose the existing inventory level is N and the inventory holding cost is k per unit of product per unit time, then the profit stream is given by the following:

$$\pi(\theta, N) = pq - cQ - kN = (\theta - \gamma q)q - cQ - kN \tag{6}$$

The goal of the myopic firm is to maximize the instantaneous profit, it will set $\frac{d\pi}{dq} = 0$, which gives

$$\theta - 2\gamma q^* - k\frac{dN}{dq} = 0, \quad \text{or} \quad q^* = \frac{\theta - k\frac{dN}{dq}}{2\gamma} \tag{7}$$

Note that the optimal q is a function of θ and N, for example, $q^* \equiv q^*(\theta, N)$. To simplify it, we note that, when one more unit is sold (i.e., q is up by 1), the inventory balance will decline by 1 (i.e., N will fall by 1), hence $\frac{dN}{dq} = -1$. This gives us the optimal amount that the myopic firm will sell at any instant:

$$q^* = \frac{\theta + k}{2\gamma} \tag{8}$$

Here, the instant sales q^* could be larger than capacity Q due to the inventory. However, there is an upper limit on how many units the firm can sell at any point in time; suppose it can draw down inventory at a rate of $\dot{N}\left(= \frac{dN}{dt}\right)$

units per unit time, then the sales amount is limited by $q \leq Q + \dot{N}$. The demand threshold where the sales amount reaches the upper limit is given by $\overline{\theta} = 2\gamma\left(Q + \dot{N}\right) - k$.

Moreover, please notice that from the inventory dynamic, we can solve its value up to time t

$$N_t = tQ - \int_0^t q_\tau d\tau \tag{9}$$

When substituting this into Equation (6), we can rewrite instantaneous profit as follows:

$$\pi(\theta) = (\theta–\gamma q)q – cQ - ktQ - k\int_0^t q_\tau^* d\tau \tag{10}$$

The valuation of the myopic firm cannot be expressed analytically, but the general approach is as follows: firm's value $V(\theta,\ t)$ is given by

$$V(\theta,\ t) = E_t^Q\left[\int_t^\infty e^{-r\tau}\pi\left(\tau,\ \theta|q_\tau^*\right)d\tau\right] \tag{11}$$

Under Feynman–Kac, a partial differential equation (PDE) can be written as follows:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2\theta^2\frac{\partial^2 V}{\partial\theta^2} + \mu\theta\frac{\partial V}{\partial\theta} + \pi\left(t, \theta|q_t^*\right) = rV \tag{12}$$

which is subject to three boundary conditions:

$$V(\theta \downarrow 0, t) = 0 \tag{13}$$

$$V(\theta \uparrow \infty, t) = Q\left(\frac{\theta}{r-\mu} - \frac{\gamma Q + c}{r}\right) \tag{14}$$

$$V(\theta, T) = 0 \tag{15}$$

Equation (13) states that the firm's value approaches zero when demand falls to very low levels. Equation (14) states that, when demand is very large, the firm will sell its entire output and keep no inventory. Equation (15) states that, when the firm is liquidated (at time $t = T$), its liquidation value will be zero (remaining inventory value will be offset by storage costs).

Unfortunately, the PDE still cannot be solved directly because it is path-dependent, that is, current profit depends on previous inventory history. We will therefore turn to the Monte Carlo simulation to solve this; the numerical details are in Appendix A.

2.3. Strategic Firm Valuation

The strategic firm will select optimal sales *ex ante* to maximize the present value of all future profit flows under demand uncertainty. Mathematically, the general form of profit process for strategic firm is the same as that for the myopic firm (e.g., Equation (10)), although they differ in the choice of optimal sales q_t^*. However, the optimal sales level for the strategic firm is difficult to express explicitly because we do not yet know the expression for the firm's value *ex ante*, which in turn varies with q_t^*.

To begin with, the value of the strategic firm after investment, $V(\theta, t,\ q)$, is governed by following equation:

$$V(\theta, t,\ q) = \max_q E_t^Q\left[\int_t^\infty e^{-r\tau}\pi(\theta, \tau,\ q)d\tau\right] \tag{16}$$

It can be observed that the Equation (16) is similar to Equation (11) except for the search process of optimal q_{t}. It can be similarly reformulated to the following PDE:

$$\max_q\left[\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2\theta^2\frac{\partial^2 V}{\partial\theta^2} + \mu\theta\frac{\partial V}{\partial\theta} + \pi(t, \theta)\right] = rV \tag{17}$$

Technically, the profit flow at any time t depends on previous optimal sales q_{t-1}^* (to calculate inventory), which also has to be solved in the entire PDE domain. Due to the complex nature of path-independence, we also turn to simulation to solve it. It has similar boundary conditions as for myopic firms:

When demand approaches zero, the firm's value becomes zero

$$V(\theta \downarrow 0, t) = 0 \tag{18}$$

When θ is very large, the firm will sell its full output, hence no new inventory is generated

$$V(\theta \uparrow \infty, t) = Q\left(\frac{\theta}{r-\mu} - \frac{\gamma Q + c}{r}\right) \tag{19}$$

Suppose there is a life limit $t = T$ such that the firm will be liquidated with zero

$$V(\theta, T) = 0 \tag{20}$$

These boundary conditions will help establish the accuracy and convergence of simulation results. The detailed algorithm is listed in Appendix B.

2.4. A brief discussion of the numerical method

The computation for the strategic firm is challenging because the firm's current (optimal) inventory is an outcome of past production and optimal sales, and the past optimal sales are linked to the current and future (optimal) inventories. There are two ways to solve for the optimal q_t^*, the "lumpy" approach and the "stepwise" approach. Under the lumpy approach, we forecast sales at all points in time $\boldsymbol{q_{1\sim T}} = (q_1, q_2, \ldots, q_T)$ by iterating over all possible sales combinations (e.g., sales could be any real non-negative numbers) to maximize the expected firm's value $E_0^Q(\theta_t, q_t^*)$. In the stepwise approach, we only consider solving optimal sales up to time t, $\boldsymbol{q_{t\sim T}} = (q_t, q_{t+1}, \ldots, q_T)$ to the maximize expected firm's value at current t: $E_t^Q(\theta_{t\sim T}, q_{t\sim T}^*)$ and iterate from $t = T$ to 1. From our numerical results, there does not seem to be much of a difference between the two methods in computation time.

However, simulation is costly in computation time, which makes it less suitable for the strategic-type firm. In addition, a longer time series (e.g., a longer firm life) will exponentially increase central processing unit (CPU) time. For example, if we simply search for optimal sales at each time spot to maximize the firm's value at entry, it may cost several minutes to a half hour to find the solution for a short firm life, even for a single path! Such a method cannot solve the problem for a large batch of simulation paths (say, 10,000 paths).

Our proposed searching algorithm can be described as follows:

Step 1. To start with, we assume that all periods of inventories are positive and compute the corresponding optimal sales, then we recalculate the updated inventories. If all of them are positive, then it is one of our solutions, if not, go to step 2.

Step 2. Suppose the first period of zero inventory is $t = i$, then we check if the inventories of all previous periods, for example, $t = 2, 3, \ldots, i-1$, are positive. If not, then we move the search to $t = i + 1$ and redo this part; if yes, then do the following:

1. Assume all inventories after $t = i$ are positive, and update if under such assumption both $N_i = 0$, and $N_{i+1,\ldots T} > 0$ (e.g., all assumptions are valid). If yes, we find the solution. If not, we do next.
2. Then we will have two possible outcomes:

   If $N_i > 0$, then our assumption on $N_i = 0$ is invalid and we move a new search to $t = i + 1$, the reason is that the assumption of all positive future inventories has maximized q_i (e.g., minimize N_i).

   If $N_i = 0$, but not all updated future inventories are positive, then it indicates that there should be a second period, $t = j\,(j > i)$, at which $N_j = 0$, for example, repeat step 2 to find out j period.

Step 3. Because step 2 could generate multiple solutions for $N_i = 0$ and/or $N_j = 0$, given $max(i,j) < T$, we need to repeat both steps to find all possible solutions until the end of the time series. To facilitate this process, we can, first, find the latest time when the inventory becomes zero, then all previous periods should also be candidates.

Although to prove the convergence and stability of our algorithm is beyond our capacity. It should not be hard to convince because we will have a finite set of solutions, for example, 2^N, there always exists the best set to maximize the equity value *ex ante*, in an iterative procedure that always improves value function, we should get the optimal solution in a certain number of iterations.

Under our proposed searching algorithm, the CPU time to calculate a single path can be decreased to 0.01 second, which is still very time-consuming, particularly for comparative static analysis or searching for optimal investment decisions, thus we have to limit to a max firm life of 10 years with 1-year time step.[1] To enhance the accuracy, we then simulate 20 batches to get the standard error down to the 1% level.

2.5. Analysis

Before presenting the numerical results, we discuss, briefly, what we can expect from the computations. As discussed above, all the firms will produce at the same rate Q but will sell different amounts. The inflexible (benchmark) firm does not have the ability to maintain inventory, hence it will sell all that is produced (Q). Both the myopic and the strategic firm can sell at a different rate because they have the ability to maintain inventory; therefore, they will sell more than or less than the amount they produce (Q).[2] However, their objective will be different: the myopic firm wants to maximize profits every period, whereas the strategic firm wants to maximize the firm's value. Because its decisions are designed to maximize the firm's value, it is clear that the strategic firm's value will be the highest of the three types; the next should be the myopic firm because it also has flexibility with regard to sales but uses its flexibility to myopically maximize instantaneous profits. The inflexible (benchmark) firm should have the lowest value because it is not able to maintain inventory and hence has no flexibility with regard to sales. Thus, the usual ordering of the firm's value is strategic, myopic, and benchmark, in order of decreasing value. There, however, is one exception, discussed below.

If the myopic firm sells a smaller quantity (i.e., q^* is smaller), then it will build up more inventory, which might destroy the firm's value because of inventory costs (recall that the myopic firm is maximizing profits, not the firm's value). Therefore, because of the excess inventory carrying cost, it is possible (although perhaps counterintuitive) that, when q^* is small enough, the value of the myopic firm falls below the value of the benchmark firm. Recall that the myopic firm sells the output at the rate of $q^* = \frac{\theta+k}{2\gamma}$; hence q^* is increasing in θ and k, while it is decreasing in the demand elasticity γ. Therefore, the myopic firm's value is more likely to be below the benchmark value when θ and k are small and when γ is large; alternatively, the myopic firm's premium is less likely to be negative when the demand level (θ) is large and the demand elasticity (γ) is small. Both of these are confirmed by our numerical results below.

Thus, the premium for the strategic firm's value over the benchmark firm will always be positive. However, the value of the ability to keep inventory will decline as the inventory holding cost rises, thus the premium should be decreasing in k (and approaching zero for a large enough k). However, the premium for the myopic firm over the benchmark firm can be positive or negative. Moreover, for all cases (positive or negative), this premium should approach zero for a large enough k because the ability to maintain inventory will make no difference if inventory holding costs are very high.

3. Numerical Results from the Model

3.1. Base-Case Parameter Values

In this section, we present numerical results from the simulation. We start with a comparison of the firm's value as a function of inventory holding cost k, for the myopic firm and the strategic firm as well as the benchmark firm (for the benchmark firm, the value will obviously be independent of k). For numerical results, we need to specify the input variables. We use a set of "base case" values for the input variables, as follows: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0$, $m_1 = 5$, $K = 8$, $\mathrm{T} = 10$, $k = 0.2$, and $\theta_0 = 6$. We choose them based on the following economic literature:

For the interest rate, we adopted $\mathrm{r} = 4\%$ as in Aretz and Pope (2018), in line with the average 10-year U.S. treasury rate. For the expected drift of the output price, we use $\mu = 0$, which conforms to Bayer (2007) and Lambrecht (2001). The cash flow volatility σ is set at 0.2, in following Lyandres and Zhdanov (2013) and Arnold (2014). The return-to-scale parameter δ is set to 0.5, the approximate average of past estimates.[3] All others are structural parameters and can be studied in a comparative statics analysis.

[1] In fact, MATLAB only allows a minimum time step = 0.5 year for such simulation because it only can iterate max 2^{20} inventory policies for a single simulated path. However, the smaller step size will not only cost much more time but might also make the simulation results non-convergent.

[2] Note that it can sell more than it produces only if there is inventory on hand.

[3] The return to scale varies in a large range in the past papers to entertain corresponding calibration performance. However, varying its value does not alter our main conclusion, so we take an approximately average of past values, to name a few, they are Miao (2005) ($\gamma = 0.4$), Danis and Gamba (2018) ($\gamma = 0.475$), and Riddick and Whited (2009) ($\gamma = 0.75$).

3.2. Valuation and Premium over the Benchmark Firm

3.2.1. Base case results

The firm's value as a function of inventory holding cost k for the three firms, shown in Figure 1(a): benchmark or inflexible firm (black line), myopic firm (blue line), and strategic firm (broken red line), and the same results in terms of premium over the benchmark firm's value are shown in Figure 1(b). As expected from the above discussion, the benchmark firm's value is independent of k and the strategic firm's value is the highest over the entire range. Also, the value of the strategic firm is a decreasing function of k; for a small k, the difference in the firm's value is substantial (about 11% above the benchmark firm's value for $k = 0$); but the difference falls rapidly with k, and it becomes negligible levels for k that exceeds 1.5. This is not surprising because the inventory becomes more expensive to support as k is increased, hence the ability to maintain inventory becomes less valuable; thus, the strategic firm value (and premium) is decreasing in k.

The behavior of the myopic firm value is a little different. First, it is a U-shaped function of k, falling from 6.75% for $k = 0$ to –12.1% for $k = 2$ and then rising slightly as k is increased further. Second, the myopic firm's value falls below the benchmark firm value when k is large enough (for k exceeding 0.4).

Thus, for a small inventory holding cost, the myopic firm value can also be significantly higher than the benchmark firm's value. However, for a larger inventory holding cost, the ability to maintain inventory can destroy value if it is not used optimally, for example, if maximizing short-term profit instead of value as in the myopic firm; in the base case, up to 12% of the firm's value can be destroyed this way. The U-shaped relationship can be explained as follows. As k is increased from 0, there are two opposing effects on the firm value: (i) direct effect: a higher k means higher inventory holding cost, which lowers the firm value and thus results in a downward-sloping curve; and (ii) indirect effect through q^*: as discussed above, a higher k results in higher q^*, that is, the firm starts reducing inventory; the resulting lower inventory holding cost will increase firm value and this will result in an upward-sloping curve. The latter effect dominates for a larger k, giving rise to the overall U-shaped relationship between inventory holding cost and the myopic firm's value, observed in Figure 1.

There are two points worth noting in these results. First, the benefits of maintaining inventory diminish rapidly as the inventory holding cost rises. Therefore, in industries where the inventory holding costs are high, it would make sense to use more inventory-minimizing techniques such as JIT (just in time). Second, an inventory management policy of choosing inventory levels based on maximizing profits (as in the myopic firm) could, in fact, end up reducing the firm's value relative to an inflexible (a no-inventory) firm. This is an important point because, in practice, inventory policy is often implemented out by operating managers whose objective is to maximize annual profits (Bassamboo et al. 2020; Canyakmaz et al. 2022; Li et al. 2021; Ma et al. 2022; Transchel et al. 2022; Zhao 2008). Our result indicates that profit-maximizing might not, in fact, be optimal because it could result in reduced firm value (particularly for large inventory holding cost). This suggests that it is better for inventory policy to be set on the basis of the firm's value rather than profits.

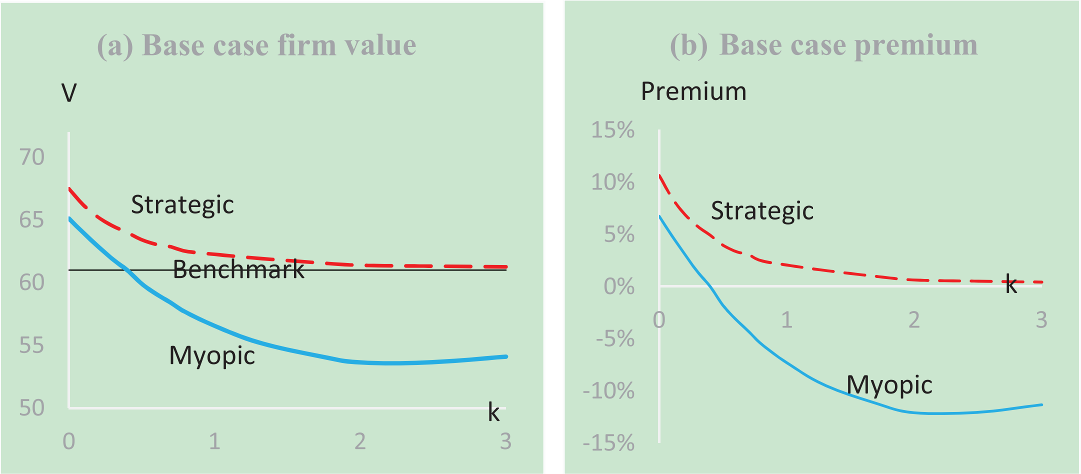


Figure 1: Shows the firm values for the three firm types (benchmark, myopic, and strategic) and the premium over the benchmark for the myopic and strategic firms. The base-case parameter values are used: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0, m_1 = 5$, $K = 8$, $\mathrm{T} = 10$, and $\theta_0 = 6$.

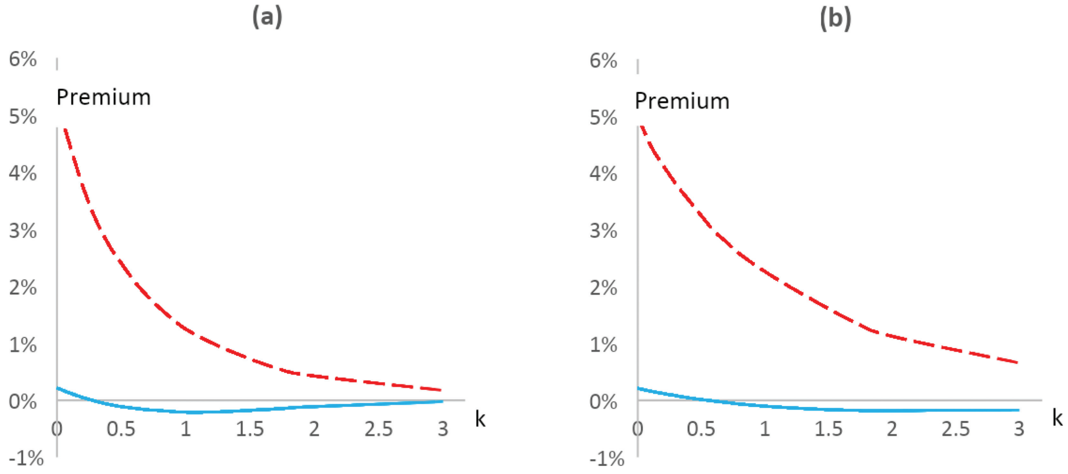


Figure 2: Shows the premium over the benchmark firm for two special cases, $\gamma = 0.5$ and $\theta_0 = 12$ (base-case value for all other parameters). The broken red line shows the strategic firm, and the solid blue line shows the myopic firm. Apart from the above parameters, the base-case values are used: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0, m_1 = 5$, $K = 8$, T = 10, and $\theta_0 = 6$.

Next, recall from our discussion above that the myopic firm's premium is less likely to be negative when demand level (θ) is large and demand elasticity (γ) is small because q^* will be larger and the firm will be less likely to accumulate large quantities of inventory, and, with lower inventory holding costs, there will be less value destruction. Thus, for high θ and/or low γ, we would expect the negative premium in the myopic firm's value to be smaller than in the base case. Repeating the numerical computations with these two scenarios ($\theta = 12$ and $\gamma = 0.5$), we find that this is indeed the case. As Figure 2 shows, in both cases, the negative premium that we noted in Figure 1(b) becomes much smaller; in fact, the premium is very close to zero. This is because, with a larger q^*, the myopic firm will be selling more and leaving less in inventory, which makes it resemble more and more the benchmark firm; thus, with a high θ and low γ, the difference between myopic and benchmark firms will shrink significantly and the premium will be closer to zero, consistent with Figure 2.

3.2.2. Comparative statics

Because there is no inventory with the benchmark firm, the benchmark firm has to sell everything it produces, even if it means selling at low prices during low-demand periods. However, both the myopic and the strategic firms can maintain inventory rather than selling the output at low prices, the former doing it in such a way as to maximize (short-term) profit, whereas the latter maximizes the firm's value. Thus, as explained in Section 2.4, both firms will generally be valued at a premium over the benchmark firm (although there might be exceptions for the myopic firm, as explained above). In this section, we find that this is indeed the case; the numerical results show that, for reasonable parameter values, the premium over the no-inventory firm is positive for both firms (Figure 3(a)–(g)).

We now look at how the various input parameters affect the premium for the myopic and strategic firms over the value of the benchmark no-inventory firm. Anything that reduces the need to maintain inventory (thereby reducing the value of the option to maintain inventory), for example, greater demand level, will move the premium closer to zero. Conversely, anything that makes the option more important (e.g., higher demand volatility) will move the premium away from zero.[4] This is indeed what our numerical results (below) confirm.

First, we look at the effect of demand volatility (σ). For both strategic and myopic firms, the premium is an increasing function of σ. As we know from the standard option theory, greater volatility increases option values. Because both firms have the option to maintain inventory, it is not surprising that the firm's value (and premium over the benchmark firm) is increasing in volatility in both cases, as shown in Figure 3(a).

Next, a higher demand growth rate (μ) has a negative effect on both strategic and myopic firm value, as shown in Figure 3(b), and both premiums approach zero when the growth rate is very large. This is also as expected because

[4] Recall that, in the base case (Figure 1), the premium for the myopic firm could be positive or negative, depending on the inventory holding cost k. For a positive (negative) premium, moving closer to zero implies that the premium will be decreasing (increasing); moving away from zero will imply just the opposite. Therefore, for the myopic firm, whether the premium is increasing or decreasing will depend on whether it is positive or negative. This is illustrated in Figure 4.

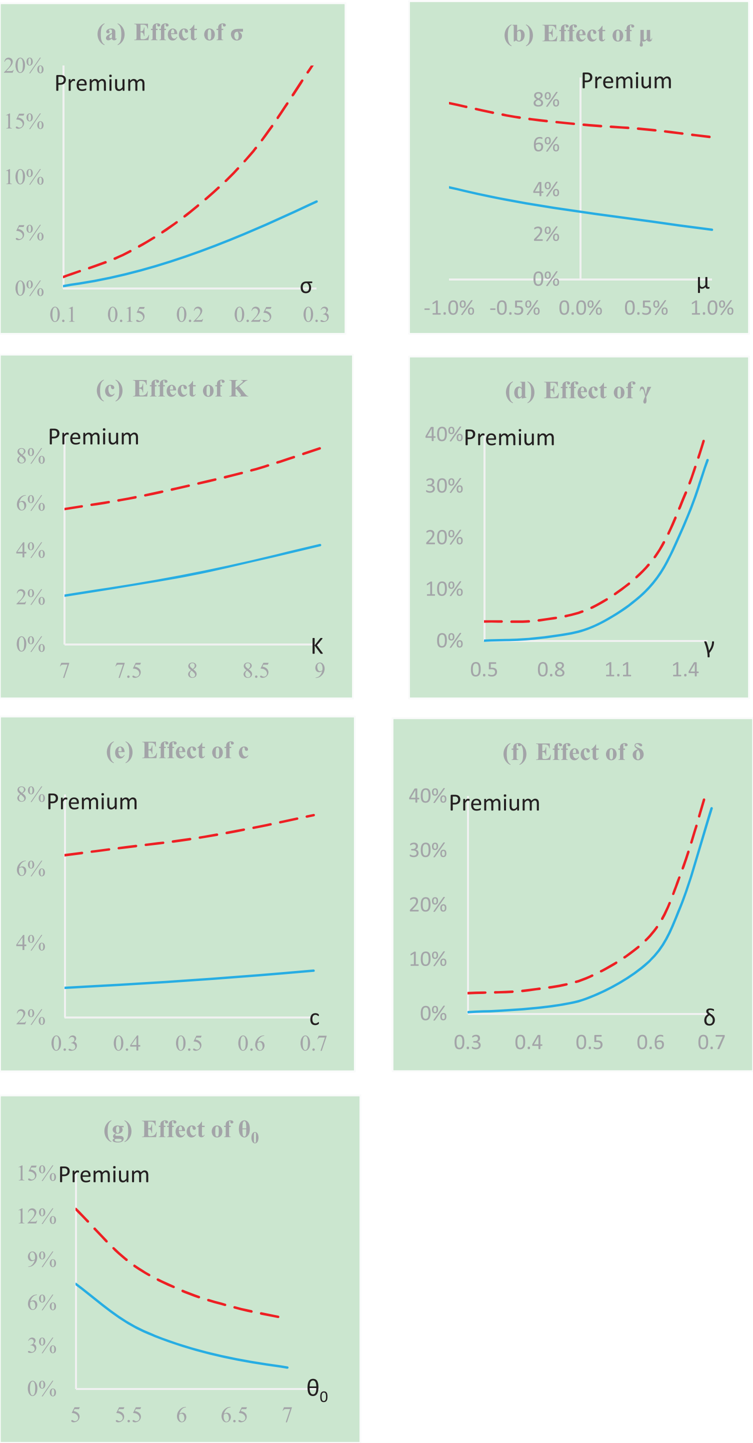


Figure 3: Comparative static results for the premium relative to the benchmark firm. The broken red line shows the strategic firm, and the solid blue line shows the myopic firm. The base-case parameter values are used: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0$, $m_1 = 5$, $K = 8$, $\mathrm{T} = 10$, $\mathrm{k} = 0.2$, and $\theta_0 = 6$.

a higher growth rate will cause both firms to sell more, hence inventory will play a diminished role and the premium will trend toward zero as a result (and they are decreasing functions of μ because both premiums are positive).

A larger capacity (K) means that the firm is producing more because it always produces at full capacity; when it is producing more, there will be more inventory, hence inventory will play a larger role. Thus, the magnitude of the premium (resulting from the ability to maintain inventory) will rise. As a result, premiums will move away from zero as K is increased, that is, because the premium is positive, it will increase with K in both cases. As we see in Figure 3(c), this is indeed the case.

Next, greater demand price sensitivity (γ) will result in a lower current output price (all else remaining unchanged), hence both flexible companies will sell less and therefore maintain higher inventory levels. This means that a higher γ will cause the inventory effect to be larger; hence the magnitude of the premium will be larger (i.e., it will move away from zero) as γ is increased. That is, the premium, if positive (negative), will be increasing (decreasing) in price sensitivity. As we see in Figure 3(d), this is indeed the case, with the premium for both firms being positive and increasing in γ.

For the operating cost (c), note that a higher c means that the margin is lower, which has the same effect as a lower price or a higher γ; thus, as in the case of γ, a higher c will result in the inventory becoming more important, hence the magnitude of the premium will increase with c, that is, the premium will move away from zero as c is increased or the premium, if positive (negative), will be increasing (decreasing) in c. As shown in Figure 3(e), this is exactly what we find, with the premium in both cases being positive and increasing in c.

A larger returns-to-scale parameter (δ) implies that a greater quantity will be produced with the same amount of capital, therefore, the ability to maintain inventory will be more valuable. Thus, the premium should be moving away from zero (if positive, an increasing function of δ), which is consistent with our numerical results shown in Figure 3(f).

When the demand level at investment (θ_0) is higher, both firms will sell more and thus have less in inventory; therefore, inventory will play a reduced role and the magnitude of the premium will fall, that is, the premium will move toward zero. As a result, a positive premium will be a decreasing function of θ_0. This is consistent with our numerical results, as shown in Figure 3(g). Finally, the two parameters interest rate (r) and investment cost (m_1) do not have a noticeable effect on the two premiums, because the myopic and strategic firms are impacted the same way as is the benchmark firm.

In the above comparative static results, the myopic firm premium was positive in all cases; with negative premium, the relationship will seem somewhat difference. To illustrate, we show in Figure 4 the effect of μ and γ when the myopic premium is negative (by setting inventory-holding cost $k = 0.5$ instead of 0.2). We note that (for the myopic firm premium) the effect is now different from the previous case (Figure 3), that is, increasing in μ and decreasing in γ; this is because the myopic premium is negative, as discussed above.

To summarize, the above comparative static results show that the premium over the benchmark firm's value can vary a lot when the input parameters are varied. Moreover, we also find (not shown) that the response of the firm's

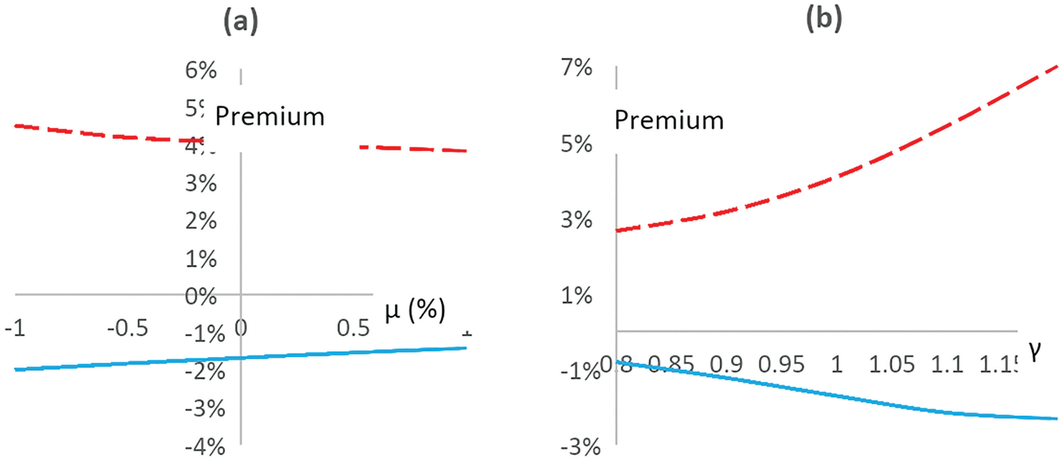


Figure 4: Some comparative static results for the premium *when the myopic firm's premium is negative*. The broken red line shows the strategic firm value, and the solid blue line shows the myopic firm value. The base-case parameter values are used: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0$, $m_1 = 5$, $K = 8$, $\mathrm{T} = 10$, $k = 0.5$, and $\theta_0 = 6$.

value to certain parameter values (e.g., investment size) can be quite different for firms with inventory and those without inventory; the implication is that investment size and its sensitivity to various parameters can be quite different from the traditional models if we include the ability to maintain inventory.

3.3. Optimal Investment Size

Real-option models study the timing and size of investment (e.g., Bar-Ilan and Strange 1999; Besanko et al. 2010; Shibata and Nishihara 2018). However, as discussed by Li and Mauer (2016), firms often do not have the freedom to choose investment timing because external circumstances, such as competition or a short life of investment opportunity, determine the timing of investment. The value of such an investment opportunity can dissipate rapidly if not acted upon immediately, particularly in high-technology industries or even traditional industries with substantial competition. In such situations, the firm does not have the luxury of waiting for the optimal time to invest. However, the firm can in general choose the size or scale of the investment (Dangl 1999; Decamps et al. 2006; Dixit 1993). Therefore, the choice of investment size, given exogenously specified timing, is an important corporate decision; see, for instance, Jou and Lee (2004) or Moyen (2007).

In this section, we examine the investment size (or production capacity) decision and how it is impacted by the firm's ability to maintain inventory. The benchmark firm must sell everything it produces right away, hence it will be more hesitant to invest in a large capacity relative to the myopic firm or the strategic firm. Thus, the optimal investment size should be higher when the firm has the inventory option.

The firm value for all three cases as a function of investment size K is shown in Figure 5. We note that the optimal K for the myopic and strategic firm is significantly larger than that of the myopic firm (9.4 versus 7.6 units of capital), as expected from the above discussion; note that there is no difference in optimal size between the myopic and strategic firms. Clearly, the investment size can be significantly impacted by the ability to maintain output inventory.

We also take a look at how the optimal investment size is affected by certain parameter values and whether the inventory option affects this relationship. In particular, we focus on the following parameters: inventory holding cost k, demand volatility σ, and demand elasticity γ. These results are shown in Figure 6. Not surprisingly, the benchmark firm's K^* is unaffected by k, whereas K^* for both myopic and strategic firm are decreasing in k. The strategic firm's K^* converges to that of the benchmark firm as k is increases sufficiently; however, the myopic firm's K^* can be smaller than that of the benchmark firm for a high enough k. As discussed in Section 3.2, the myopic firm value can be below the benchmark firm's value if k is high enough because it maximizes profit instead of value. For the same reason, the myopic firm will choose a smaller size than the benchmark firm for large enough k, as shown in Figure 6(a).

Next, as Figure 6(b) shows, the benchmark K^* is independent of σ; this is because it has no embedded options, hence the decision is unaffected by volatility. However, both myopic and strategic firms have the inventory option, hence the value increases with volatility, thus K^* in both cases is increasing in σ; we also note in Figure 6(b) that the

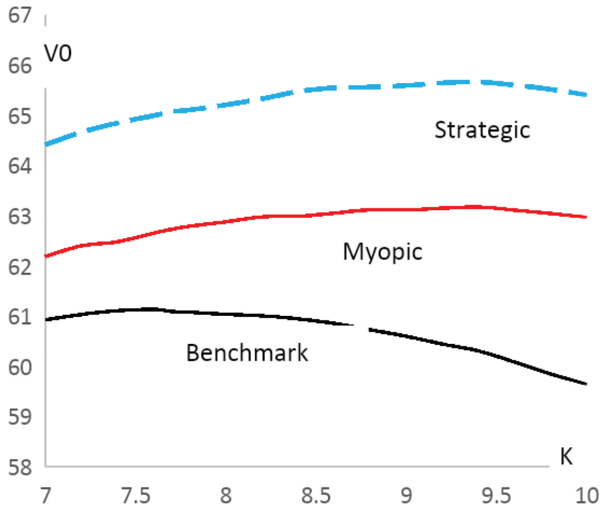


Figure 5: Shows firm value V_0 in all three cases as a function of investment size K by using the base case parameter values: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0$, $m_1 = 5$, $T = 10$, and $\theta_0 = 6$. The optimal investment size is 7.6 for the benchmark firm and 9.4 for both the myopic firm and the strategic firm.

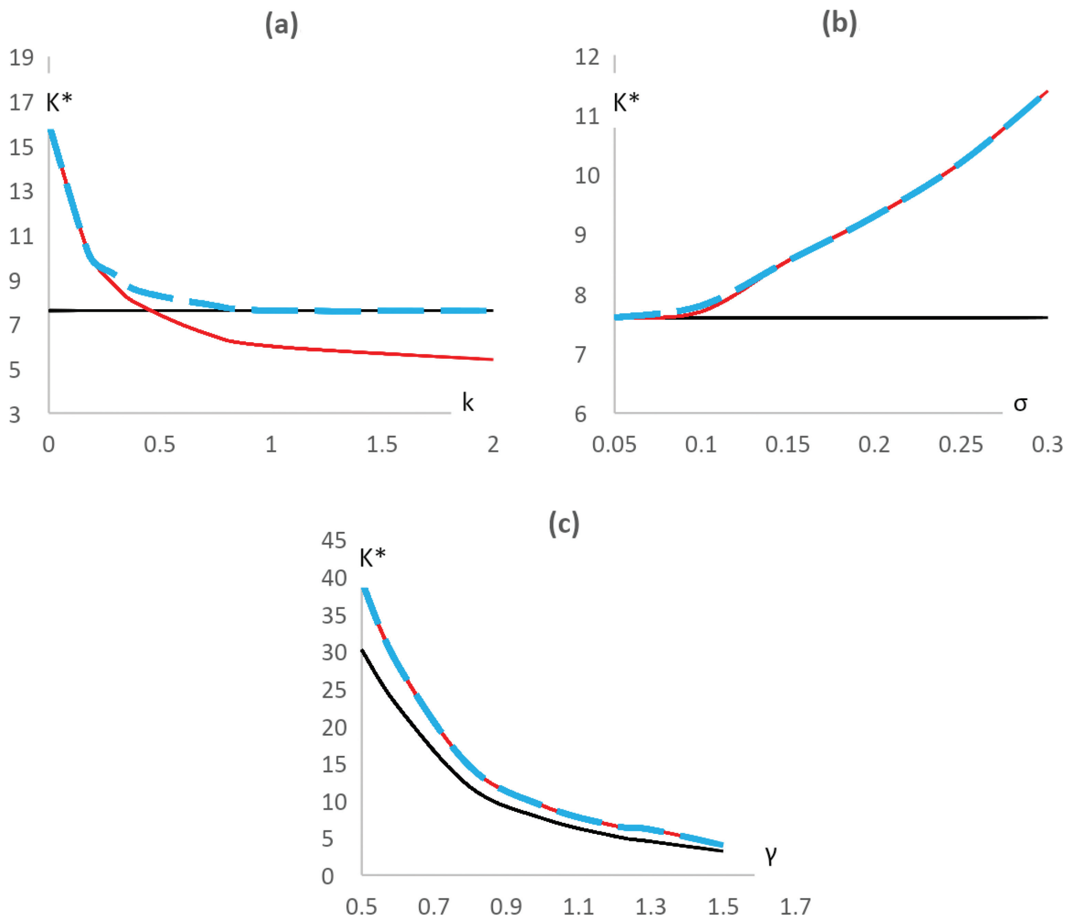


Figure 6: Shows optimal investment size K^* as a function of inventory holding cost k (part a), demand volatility σ (part b) and demand elasticity γ (part c). The base case parameter values are used: $c = 0.5$, $\gamma = 1$, $r = 0.04$, $\mu = 0$, $\sigma = 0.2$, $\delta = 0.5$, $m_0 = 0, m_1 = 5$, T = 10, and $\theta_0 = 6$.

optimal size is virtually identical for myopic and strategic firms for all σ. Finally, Figure 6(c) shows the effect of price elasticity γ. A larger γ means that greater output will be relatively less attractive because the price decline will be larger. It is then not surprising that K^* is decreasing in γ for all three firms. Once again, there is no difference in K^* for the myopic and strategic firms. Overall, it seems that there is no difference between optimal investment size for the myopic firm and the strategic firm, although there might be substantial differences in the valuation of the two.

4. Conclusion

This paper studied the contingent-claim valuation of a company that can maintain an inventory of its output. Existing models ignore the possibility of maintaining inventory. We show that the value of a company by following the optimal inventory policy can be significantly higher than the traditional non-inventory company, particularly if the inventory-holding cost is not large. This premium shrinks as holding cost is increased and disappears for large enough holding cost, and is particularly large when demand is volatile, when demand level is low, and when price elasticity is large. Thus, the ability to maintain inventory could potentially have a significant impact on the company's investment and financing decisions.

It is also shown that, when the company's inventory policy is set myopically so as to maximize the current profit rather than long-term value (as is often the case in practice), it is possible that the firm's value actually falls below the traditional no-inventory firm, that is, the premium turns negative. We also show that the optimal investment size for a firm that follows the optimal inventory policy can be significantly larger than the traditional no-inventory firm, particularly when the inventory-holding cost is low, demand volatility is high, and price elasticity is low.

This paper heralds new directions on the application of artificial intelligence (AI) in the dynamic valuation problem. For example, Liu et al. (2023) develops a deep learning-based numerical method (The Seven League scheme) on graphics processing units. Their method improves the computational and convergence speed for large-scale Monte Carlo simulation on stochastic differential equations. Rhijn et al. (2023) have developed generative adversarial networks algorithm to solve stochastic differential equations. They found that the supervised generative adversarial networks outperformed the Euler and Milstein schemes in strong error on a discretization with large time steps. However, their methods can only be applied to well-defined economic boundaries, whereas our model is built on an undefined inventory process, which needs to be optimized along with the iterative valuation. It is obvious that a longer time series simulation with optimized inventory management will lead to more accurate assets valuation. Therefore, the above-mentioned deep leaning algorithm will have great potential on thus topics.

Appendix A: Valuation of Myopic Firm

In the simulations, we need to transform previous continuous time modelling of profit and inventory dynamics to a discrete-time setting. The discrete approximation to Equation (1) is as follows:

$$\theta_{i,j} = \theta_{i,j-1} \exp\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}\varepsilon_{i,j} \tag{A.1}$$

where the subscript i represents the i^{th} path and j denotes the time period, the profit flow thereby can be rewritten as

$$\pi_{i,j} = \left(\theta_{i,j} - \gamma q_{i,j}\right)q_{i,j} - cQ - kN_{i,j} \tag{A.2}$$

The current inventory $N_{i,j}$ can be expressed as

$$N_{i,j} = N_{i,j-1} + \left(Q - q_{i,j}\right)\Delta t \tag{A.3}$$

Note that $N_{i,j-1}$ is the inventory stock at previous period.

The instantaneous profit at time spot j can be rewritten as

$$\pi_{i,j} = \left(\theta_{i,j} - \gamma q_{i,j}\right)q_{i,j} - cQ - k\left(N_{i,j-1} + \left(Q - q_{i,j}\right)\Delta t\right) \tag{A.4}$$

with the following non-negative requirement

$$min\left(N_{i,j}, N_{i,j-1}\right) \geq 0 \tag{A.5}$$

The firm value will be calculated by

$$V_{i,j} = \exp\left(-r\Delta t\right) V_{i,j+1} + \int_0^{\Delta t} \pi_{i,j} \tag{A.6}$$

The last item is given by

$$\int_0^{\Delta t} \pi_{i,j} = \left(\left(\theta_{i,j} - \gamma q_{i,j}\right)q_{i,j} - cQ - kN_{i,j-1}\right)\Delta t - \int_0^{\Delta t}\left(k\int_0^x \left(Q - q_{i,j}\right)d\tau\right)dx \tag{A.7}$$

And it can be further written as

$$\int_0^{\Delta t} \pi_{i,j} = \left(\left(\theta_{i,j} - \gamma q_{i,j}\right)q_{i,j} - cQ - kN_{i,j-1}\right)\Delta t - \frac{1}{2}k\left(Q - q_{i,j}\right)\Delta t^2 \tag{A.8}$$

Note the last integrand captures the cumulative inventory storage cost over Δt.

Recall that the profit flow over Δt is

$$\int_0^{\Delta t} \left(pi_j | N_{j-1}\right) = \left(\left(\theta_j - \gamma q_j\right)q_j - cQ - kN_{j-1}\right)\Delta t - \frac{1}{2}k\left(Q - q_j\right)\Delta t^2 \tag{A.9}$$

Take first-order condition with respect to q_j, we have optimal sales for the current time spot j

$$q_j = \frac{2\theta_j + k\Delta t}{4\gamma} \tag{A.10}$$

The upper sales amount should be constrained by upper limit $\overline{q}_j\Delta t < N_{j-1} + Q\Delta t$, or we have

$$q_j = \begin{cases} \dfrac{2\theta_j + k\Delta t}{4\gamma} \; if \; \theta_j < \overline{\theta}_j \; \dfrac{N_{j-1}}{\Delta t} + Q \; if \; \theta_j \geq \overline{\theta}_j \end{cases} \tag{A.11}$$

here $\overline{\theta}_j$ is the upper demand threshold

$$\overline{\theta}_j = 2\gamma\left(\frac{N_{j-1}}{\Delta t} + Q\right) - \frac{k\Delta t}{2} \tag{A.12}$$

Appendix B: Valuation of the Strategic Firm

The algorithm for the strategic firm is more complex due to the unknown inventory decisions. Before presenting optimal solutions to the entire time series in a simulated discrete time periods, we start with presenting a very simple case. We assume that the entire life only has four periods: the demand shocks are $\theta = \theta_t$, for $t = 0$, 1, 2, 3, 4. Note that production begins at $t = 1$. The demand shocks in the time series will be produced in a randomness generator in MATLAB, The MathWorks, Inc., Massachusetts, United States.

The firm's value, as a function of sales at four periods q_1, q_2, q_3, q_4 are

$$V_0 = max_{q_1, q_2, q_3, q_4}\left(e^{-r\Delta t}\Pi_1 + e^{-2r\Delta t}\Pi_2 + e^{-3r\Delta t}\Pi_3 + e^{-4r\Delta t}\Pi_4\right) \tag{A.13}$$

here $\Pi_j = \int_0^{\Delta t} \pi_j$ represents cumulative profit flow over period of Δt.

The firm value can be expended easily in the following for a general case, with subscript j as the j^{th} time period

$$V_0 = max_{q_j = 1,2,3,\ldots T} \sum_{j=1}^{T} \left\{ e^{-jr\Delta t}\left[\left(\theta_j - \gamma q_j\right)q_j - cQ - kN_{j-1}\right]\Delta t - k\left(Q - q_j\right)\Delta t^2/2 \right\} \tag{A.14}$$

We list for up to four periods for the purpose of *induction and deduction*. In what follows, we first present detailed solution to the reduced case, then we go to implementation details in general case.

$$\begin{aligned} V_0 = {} & e^{-r\Delta t}\left\{[(\theta_1 - \gamma q_1)q_1 - cQ]\Delta t - k(Q - q_1)\Delta t^2/2\right\} \\ & + e^{-2r\Delta t}\left\{[(\theta_2 - \gamma q_2)q_2 - cQ - kN_1]\Delta t - k(Q - q_2)\Delta t^2/2\right\} \\ & + e^{-3r\Delta t}\left\{[(\theta_3 - \gamma q_3)q_3 - cQ - kN_2]\Delta t - k(Q - q_3)\Delta t^2/2\right\} \\ & + e^{-4r\Delta t}\left\{[(\theta_4 - \gamma q_4)q_4 - cQ - kN_3]\Delta t - k(Q - q_4)\Delta t^2/2\right\} \end{aligned} \tag{A.15}$$

Note that inventory at $t = 0$ is zero, and we need the following constraints

$$\begin{aligned} N_1 &= max\,(Q - q_1, 0)\Delta t \\ N_2 &= max\,[N_1 + (Q - q_2)\Delta t, 0] \\ N_3 &= max[N_2 + (Q - q_3)\Delta t, 0] \\ N_4 &= max\,[N_3 + (Q - q_4)\Delta t, 0] \end{aligned}$$

They will ensure non-negative inventory conditions and regulate the maximum sales at each time. This is a super-large (particularly for full times) constrained optimization problem. Even for the four-period case, a traditional optimization such as the Kuhn-Tucker method with regard to all of q_1, q_2, q_3, q_4, N_1, N_2, N_3, N_4, are difficult to implement here because the max-type function makes the entire value non-differentiable at all domains and it depends on inventory status. We, therefore, adopt an iteration method, overall, our optimization for each simulated path can simply be written as

$$V_0 max_{q_1,\ldots,q_T, Q} = \sum_{i=0}^{T} e^{-ri\Delta t}\Pi_i(q_i, N_{i-1}(q_1, ., q_{i-1})), \tag{A.16}$$

$$\text{subject to } N_{i=Q, i\in[1,T]} > 0 \text{ and}$$

$$N_{i\neq Q, i\in[1,T]} = 0 \tag{A.17}$$

where Q represents the collection of all non-zero inventories. Note that the challenge here is that the set Q is unknown *ex ante* in the constraints and it has to be solved along with the optimization problem, and we call this *iterative optimization*.

In particular, at initial production $t = 1$, the sales must be less or equal than capacity, for example, $q_1 \leq Q$. Define inventory status as binary: zero or nonzero. We could totally have $2^3 = 8$ scenarios. Notice that the last period inventory N_4 has no impact here. In what follows, we only discuss five sample cases for the sake of exhibition. All other cases can be derived in a similar way:

Scenario 1: $N_1 > 0, N_2 > 0, N_3 > 0$

The first-order condition with respect to q_i are as follows:

$$\frac{\partial V_0}{\partial q_1} = e^{-r\Delta t}(\theta_1 - 2\gamma q_1)\Delta t + \left(e^{-r\Delta t}/2 + e^{-r2\Delta t} + e^{-r3\Delta t} + e^{-r4\Delta t}\right)k\Delta t^2 = 0 \tag{A.18}$$

$$\frac{\partial V_0}{\partial q_2} = e^{-2r\Delta t}(\theta_2 - 2\gamma q_2)\Delta t + \left(e^{-2r\Delta t}/2 + e^{-r3\Delta t} + e^{-r4\Delta t}\right)k\Delta t^2 = 0 \tag{A.19}$$

$$\frac{\partial V_0}{\partial q_3} = e^{-3r\Delta t}(\theta_3 - 2\gamma q_3)\Delta t + \left(e^{-3r\Delta t}/2 + e^{-r4\Delta t}\right)k\Delta t^2 = 0 \tag{A.20}$$

$$\frac{\partial V_0}{\partial q_4} = e^{-4r\Delta t}\left((\theta_4 - 2\gamma q_4)\Delta t + k\Delta t^2/2\right) = 0 \tag{A.21}$$

Then we have

$$q_1^* = \frac{\theta_1 + \left(\frac{1}{2} + e^{-r\Delta t} + e^{-2r\Delta t} + e^{-3r\Delta t}\right)k\Delta t}{2\gamma} \tag{A.22}$$

$$q_2^* = \frac{\theta_2 + \left(\frac{1}{2} + e^{-r\Delta t} + e^{-2r\Delta t}\right)k\Delta t}{2\gamma} \tag{A.23}$$

$$q_3^* = \frac{\theta_3 + \left(\frac{1}{2} + e^{-r\Delta t}\right)k\Delta t}{2\gamma} \tag{A.24}$$

$$q_4^* = min\left(\frac{\theta_4 + k\Delta t/2}{2\gamma}, Q + \frac{N_3}{\Delta t}\right) \tag{A.25}$$

Therefore, we have optimal sales at each instant time $q_j^* = \frac{\theta_j + \left(\sum_{i=j}^{T} e^{-r(T-i)\Delta t} - \frac{1}{2}\right)k\Delta t}{2\gamma}$, the idea is simple: each optimal sales equals demand shock plus all current and future savings on the inventory storage. Moreover, it is interesting that each optimal sales is independent of future demand shocks. However, the precondition is that inventories at all periods should be positive, that is,

$$N_j = N_{j-1} + \left(Q - q_j\right)\Delta t > 0 \tag{A.26}$$

Actually this condition also equivalently regulates the upper sales quantities

$$\overline{q}_j\Delta t < N_{j-1} + Q\Delta t \tag{A.27}$$

or an upper-demand threshold beyond which the firm will clear inventories, which is obtained by substituting previous solutions on the optimal sales

$$\overline{\theta}_j < 2\gamma\left(\frac{N_{j-1}}{\Delta t} + Q\right) - \left(\sum_{i=j}^{T} e^{-r(T-i)\Delta t} - \frac{1}{2}\right)k\Delta t \tag{A.28}$$

For simplicity, we will not discuss the negative profits, which will generate lower sales (demand) boundaries $\underline{q}_j\left(\underline{\theta}_j\right)$ below which the sales will be zero. In fact, it is doable, for example, substitute the optimal sales solution back to profit function and solve the positive root because the quadratic equation will be convex shaped. However, the multiple switching thresholds will make our algorithm very messy. In what follows, we discuss when the precondition of all positive inventories is violated:

Scenario 2 : $N_1 = N_2 = N_3 = 0$

$$q_1^* = q_2^* = q_3^* = Q,\ q_4^* \text{ is same as in Scenario 1}$$

Scenario 3 : $N_1 = 0, N_2 > 0, N_3 > 0$

$$q_1^* = Q,\ q_2^*,\ q_3^*,\ q_4^* \text{ are same as in Scenario 1}$$

Scenario 4 : $N_1 > 0,\ N_2 = 0,\ N_3 > 0$

This scenario is a little more complicated. Because, when $N_2 = 0$, it means we could not freely optimize q_2^*, instead, q_2^* is constrained by $q_2^* = 2Q - q_1^*$. So, this case becomes a constrained optimization problem.

$$max\, V_0(q_1, ., q_4),\ \text{subject to}\ 2Q - q_1 - q_2 = 0 \tag{A.29}$$

So we have to re-substitute the updated q_2^* to optimize q_1^*:

$$q_1^* = \frac{\theta_1 - e^{-r\Delta t}\theta_2 + 4\gamma Q e^{-r\Delta t} + k\Delta t(1 + e^{-r\Delta t})/2}{2\gamma(1 + e^{-r\Delta t})} \tag{A.30}$$

This equation can be further rewritten as

$$q_1^* = \underbrace{\frac{\theta_1 + k\Delta t/2}{2\gamma}}_{\textit{myopic effect}} + \underbrace{\left(2Q - \frac{\theta_1 + \theta_2}{2\gamma}\right)\frac{e^{-r\Delta t}}{1 + e^{-r\Delta t}}}_{\textit{inventory effect}} \tag{A.31}$$

This equation has two parts: the first part is the same as the myopic case. The second part captures the inventory effect: for example, when the future demand θ_2 increases, then q_1^* decreases, that is, the previous sales should decrease to leave some inventory for future sales, which is intuitive because the future price will become higher. The affiliated item $\frac{e^{-r\Delta t}}{1+e^{-r\Delta t}}$ can be considered as the "weight ratio" to capture the discount weight of future effect. Last, q_3^* and q_4^* are the same as in Scenario I.

$\text{Scenario 5}: N_1 > 0,\ N_2 > 0,\ N_3 = 0$

The logic is similar to Scenario 4. We have freedom to optimally select q_1^* and q_2^*, while leaving the equality constraint $q_3^* = 3Q - q_1^* - q_2^*$. We substitute this constrain to the value function and take the first derivative to both q_1 and q_2:

$$q_2^* = \frac{\theta_2 + k\Delta t/2}{2\gamma} + \left(3Q - q_1^* - \frac{\theta_2 + \theta_3}{2\gamma}\right)\frac{e^{-r\Delta t}}{1 + e^{-r\Delta t}} \tag{A.32}$$

and at the first period

$$q_1^* = \frac{\theta_1 + g(k\Delta t/2)}{2\gamma} + \left(3Q - q_2^* - \frac{\theta_1 + \theta_3}{2\gamma}\right)\frac{e^{-2r\Delta t}}{1 + e^{-2r\Delta t}} \tag{A.33}$$

here, the coefficient

$$g = 1 + \frac{2e^{-r\Delta t}}{1 + e^{-2r\Delta t}} \tag{A.34}$$

Obviously, we can solve the two-variable equations for q_1^* and q_2^*, however, when there are many periods until the inventory encounters zero at the first time (e.g., the constraint becomes $q_j^* = iQ - \sum_{z=1}^{i-1} q_{z\neq j}^*$), we cannot do this. For example, suppose $N_1, N_2, N_3, \ldots, N_{i-1} > 0$ and $N_i = 0$ and let us derive $q_1^*, q_2^*, q_3^*, \ldots, q_{i-1}^*$:

$$q_j^* = \frac{\theta_j + g_j(k\Delta t/2)}{2\gamma} + \left(iQ - \sum_{z=1}^{i-1} q_{z\neq j}^* - \frac{\theta_j + \theta_i}{2\gamma}\right)\frac{e^{-(i-j)r\Delta t}}{1 + e^{-(i-j)r\Delta t}} \tag{A.35}$$

The coefficient g_j can be expressed as

$$g_j = \begin{cases} 1 + \dfrac{2\displaystyle\sum_{v=1}^{i-2} e^{-vr\Delta t}}{1 + e^{-(i-j)r\Delta t}} & \text{for } j > i - 1 \\ 1 & \text{for } j = i - 1 \end{cases} \tag{A.36}$$

The equation system can be rewritten as

$$\boldsymbol{A}_{(\boldsymbol{i}-1)\times(\boldsymbol{i}-1)} q^*_{(\boldsymbol{i}-1)\times 1} = \boldsymbol{B}_{(\boldsymbol{i}-1)\times 1} \tag{A.37}$$

where the bold matrices are as following

$$q^*_{(\boldsymbol{i}-1)\times 1} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{i-1} \end{bmatrix} \tag{A.38}$$

$$\boldsymbol{B}_{(\boldsymbol{i}-1)\times 1} = \begin{bmatrix} \dfrac{\theta_1 + g_1(k\Delta t/2)}{2\gamma} + \left(iQ - \dfrac{\theta_1 + \theta_i}{2\gamma}\right)\dfrac{e^{-(i-1)r\Delta t}}{1 + e^{-(i-1)r\Delta t}} \\ \dfrac{\theta_2 + g_2(k\Delta t/2)}{2\gamma} + \left(iQ - \dfrac{\theta_2 + \theta_i}{2\gamma}\right)\dfrac{e^{-(i-2)r\Delta t}}{1 + e^{-(i-2)r\Delta t}} \\ \vdots \\ \dfrac{\theta_{i-1} + g_{i-1}(k\Delta t/2)}{2\gamma} + \left(iQ - \dfrac{\theta_{i-1} + \theta_i}{2\gamma}\right)\dfrac{e^{-r\Delta t}}{1 + e^{-r\Delta t}} \end{bmatrix} \tag{A.39}$$

$$\boldsymbol{A}_{(\boldsymbol{i}-1)\times(\boldsymbol{i}-1)} = \begin{bmatrix} 1 & a_1 & a_1 & \cdots & a_1 \\ a_2 & 1 & a_2 & \cdots & a_2 \\ a_3 & a_3 & \ddots & a_3 & a_3 \\ \vdots & a_4 & a_4 & 1 & \vdots \\ a_{i-1} & a_{i-1} & a_{i-1} & \cdots & 1 \end{bmatrix} \tag{A.40}$$

here $a_j = \frac{e^{-(i-j)r\Delta t}}{1+e^{-(i-j)r\Delta t}}$ is the last coefficient in the equation. To solve the linear system, we use the Gauss-Seidel method.

Finally, suppose we simulate M paths, the above calculation will be repeated M times. This simple four-period case informs some intuitions:

1. The optimal sales (or inventory) level at the current time t depends on both the previous and future optimal sales (or inventory) level, which should be difficult to produce a closed form solution.
2. Luckily, all future storing costs of optimal sales expression will cut off at the first timing of zero inventories.
3. The solutions to optimal sales can be converted to solutions to determine the optimal timing of depleting inventories, which may still be impossible to solve (for N time periods, we will have 2^N cases of inventories status (e.g., zero or positive)).

References

Aguerrevere, F. L. 2003. “Equilibrium Strategies and Output Price Behavior: A Real-Options Approach.” *Review of Financial Studies* **16**: 1239–1272.

Aretz, K., and P. F. Peter. 2018. “Real Options Models of the Firm, Capacity Overhang, and the Cross Section of Stock Returns.” *Journal of Finance* **73**, no. 3: 1363–1415. doi: 10.1111/jofi.12617

Arnold, M. 2014. “Managerial Cash Use, Default, and Corporate Financial Policies.” *Journal of Corporate Finance* **27**: 305–325.

Bar-Ilan, A., and W. C. Strange. 1999. “The Timing and Intensity of Investment.” *Journal of Macroeconomics* **21**, no. 1: 57–77. doi: 10.1016/S0164-0704(99)00090-7

Bassamboo, A., A. Moreno, and I. Stamatopoulos. 2020. “Inventory Auditing and Replenishment Using Point-of-Sales Data.” *Production and Operations Management* **29**, no. 5: 1219–1231. doi: 10.1111/poms.13153

Basu, N., and X. Wang. 2011. “Evidence on the Relation between Inventory Changes, Earnings and Firm Value.” *International Journal of Business and Finance Research* **5**: 1–14.

Bayer, C. 2007. “Investment Timing and Predatory Behavior in a Duopoly with Endogenous Exit.” *Journal of Economic Dynamics and Control* **31**, no. 9: 3069–3109. doi: 10.1016/j.jedc.2006.10.006

Besanko, D., Doraszelski, U. Lu, and X. Satterthwaite. 2010. “Lumpy Capacity Investment and Disinvestment Dynamics.” *Operations Research* **58**, no. 4, part2: 1178–1193. doi: 10.1287/opre.1100.0823

Bianco, M., and A. Gamba. 2019. “Inventory and Corporate Risk Management.” *The Review of Corporate Finance Studies* **8**, no. 1: 97–145. doi: 10.1093/rcfs/cfy007

Canyakmaz, C., S. Ozekici, and F. Karaesmen. 2022. “A News Vendor Problem with Markup Pricing in the Presence of within-Period Price Fluctuations.” *European Journal of Operational Research* **301**, no. 1: 153–162. doi: 10.1016/j.ejor.2021.09.042

Cortazar, G., and E. S. Schwartz. 1993. “A Compound Option Model of Production and Intermediate Inventories.” *The Journal of Business* **66**, no. 4: 517–540. doi: 10.1086/296616

Dangl, T. 1999. “Investment and Capacity Choice under Uncertain Demand.” *European Journal of Operational Research* **117**, no. 3: 415–428. doi: 10.1016/S0377-2217(98)00274-4

Danis, A., and A. Gamba 2018. “The Real Effects of Credit Default Swaps.” *Journal of Financial Economics* **127**: 51–76.

Decamps, J.-P., T. Mariotti, and S. Villeneuve. 2006. “Irreversible Investment in Alternative Projects.” *Economic Theory* **28**, no. 2: 425–448. doi: 10.1007/s00199-005-0629-2

Dixit, A. 1993. “Choosing among Alternative Discrete Investment Projects under Uncertainty.” *Economics Letters* **41**, no. 3: 265–268. doi: 10.1016/0165-1765(93)90151-2

Elsayed, K. 2015. “Exploring the Relation between Efficiency of Inventory Management and Firm Performance: An Empirical Research.” *International Journal of Services and Operations Management* **21**, no. 1: 73–86. doi: 10.1504/IJSOM.2015.068704

Goyal, M., and S. Netessine. 2007. “Strategic Technology Choice and Capacity Investment under Demand Uncertainty.” *Management Science* **53**, no. 2: 192–207. doi: 10.1287/mnsc.1060.0611

Huberts, N. F. D., K. J. M. Huisman, P. M. Kort, and M. N. Lavrutich. 2015. “Capacity Choice in (Strategic) Real Options Models: A Survey.” *Dynamic Games and Applications* **5**, no. 4: 424–439. doi: 10.1007/s13235-015-0162-2

Jou, J.-B., and T. Lee. 2004. “Debt Overhang, Costly Expandability and Reversibility, and Optimal Financial Structure.” *Journal of Business Finance & Accounting* **31**, no. 7-8: 1191–1222. doi: 10.1111/j.0306-686X.2004.00572.x

Jou, J.-B., and T. Lee. 2008. “Irreversible Investment, Financing, and Bankruptcy Decisions in an Oligopoly.” *Journal of Financial and Quantitative Analysis* **43**, no. 3: 769–786. doi: 10.1017/S0022109000004282

Kim, K. 2020. “Inventory, Fixed Capital, and the Cross-Section of Corporate Investment.” *Journal of Corporate Finance* **60**: 101528. doi: 10.1016/j.jcorpfin.2019.101528

Koumanakos, D. P. 2008. “The Effect of Inventory Management on Firm Performance.” *International Journal of Productivity and Performance Management* **57**, no. 5: 355–369. doi: 10.1108/17410400810881827

Kroes, J. R., and A. S. Manikas. 2018. “An Exploration of ‘Sticky’ Inventory Management in the Manufacturing Industry.” *Production Planning & Control* **29**, no. 2: 131–142. doi: 10.1080/09537287.2017.1391346

Lambrecht, B. 2001. “The Impact of Debt Financing on Entry and Exit in a Duopoly.” *Review of Financial Studies* **14**, no. 3: 765–804. doi: 10.1093/rfs/14.3.765

Lederer, P. J., and T. D. Mehta. 2005. “Economic Evaluation of Scale Dependent Technology Investments.” *Production and Operations Management* **14** (1): 21–34.

Li, D., J. Chen, and Y. Liao. 2021. “Optimal Decisions on Prices, Order Quantities, and Returns Policies in a Supply Chain with Two-Period Selling.” *European Journal of Operational Research* **290**, no. 3: 1063–1082. doi: 10.1016/j.ejor.2020.08.044

Li, J. Y., and D. C. Mauer. 2016. “Financing Uncertain Growth.” *Journal of Corporate Finance* **41**: 241–261. doi: 10.1016/j.jcorpfin.2016.09.006

Liu, S., G. Colonna, L. Grzelak, and C. Oosterlee. 2023. “GPU Acceleration of the Seven-League Scheme for Large Time Step Simulations of Stochastic Differential Equations.” Preprint, submitted February 10. https://www.sciencedirect.com/science/article/abs/pii/S1386418112000444?via%3Dihub

Lyandres, E., and A. Zhdanov. 2013. “Investment Opportunities and Bankruptcy Prediction.” *Journal of Financial Markets* **16**, no. 3: 439–476. doi: 10.1016/j.finmar.2012.10.003

Ma, S., Z. Jemai, and Q. Bai. 2022. “Optimal Pricing and Ordering Decisions for a Retailer Using Multiple Discounts.” *European Journal of Operational Research* **299**, no. 3: 1177–1192. doi: 10.1016/j.ejor.2021.10.004

Mauer, D. C., and S. Ott. 2000. “Agency Costs, Under-Investment, and Optimal Capital Structure: The Effect of Growth Options to Expand.” In *Project Flexibility, Agency, and Competition*, edited by Michael J. Brennan and Lenos Trigeorgis, 151–180. New York: Oxford University Press.

Miao, J. 2005. “Optimal Capital Structure and Industry Dynamics.” *The Journal of Finance* **60**, no. 6: 2621–2659. doi: 10.1111/j.1540-6261.2005.00812.x

Moyen, N. 2007. “How Big is the Debt Overhang Problem?” *Journal of Economic Dynamics and Control* **31**, no. 2: 433–472. doi: 10.1016/j.jedc.2005.10.008

Ndubuisi, A. N., Chinyere, O. J. E., Beatrice, O., and E. P. Uche. 2020. “Inventory Management and Firm Performance: Evidence from Brewery Firms Listed on Nigeria Stock Exchange.” *International Journal of Research in Business, Economics and Management* **2**: 72–93.

Pindyck, R. S. 1982. “Adjustment Costs, Uncertainty, and the Behavior of the Firm.” *American Economic Review* **72**: 415–427.

Riddick, L.A., and T. M. Whited. 2009. “The Corporate Propensity to Save.” *Journal of Finance* **64**: 1729–1766.

Rhijn, J., C. Oosterlee, L. Grzela, and S. Liu. 2023. “Monte Carlo Simulation of SDEs Using GANs.” *Japan Journal of Industrial and Applied Mathematics* **40**, no. 3: 1359–1390. doi: 10.1007/s13160-022-00534-x

Shibata, T., and M. Nishihara. 2018. “Investment Timing, Reversibility, and Financing Constraints.” *Journal of Corporate Finance* **48**: 771–796. doi: 10.1016/j.jcorpfin.2017.12.024

Transchel, S., M. E. Buisman, and R. Haijema. 2022. “Joint Assortment and Inventory Optimization for Vertically Differentiated Products under Consumer-Driven Substitution.” *European Journal of Operational Research* **301**, no. 1: 163–179. doi: 10.1016/j.ejor.2021.09.041

Tserlukevich, S. 2008. “Can Real Options Explain Financing Behavior?” *Journal of Financial Economics* **89**, no. 2: 232–252. doi: 10.1016/j.jfineco.2007.11.003

Van Mieghem, J. A., and M. Dada. 1999. “Price versus Production Postponement: Capacity and Competition.” *Management Science* **45**, no. 12: 1639–1649. doi: 10.1287/mnsc.45.12.1631

Zhao, X. 2008. “Coordinating a Supply Chain System with Retailers under Both Price and Inventory Competition.” *Production and Operations Management* **17**, no. 5: 532–542. doi: 10.3401/poms.1080.0054

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.26

Crime Frequency during COVID-19 and Black Lives Matter Protests

Aylin Kosar
Rutgers University Camden
akosar@scarletmail.rutgers.edu

Mehmet Turkoz
William Paterson University
turkozm@wpunj.edu

ABSTRACT

COVID-19 disrupted daily life within the United States and around the world when government restrictions were implemented. During the onset of the pandemic, social unrest developed after the death of George Floyd. Our objective was to study the crime rate during the pandemic and social unrest that resulted after the death of George Floyd. We used data from four cities heavily affected by the pandemic and social unrest: Seattle, San Francisco, Los Angeles, and Philadelphia. Holt-Winters and SARIMA models were used to see if there was any change in crime during the pandemic and social unrest in addition to before and after the social unrest. Los Angeles had the lowest crime frequency of the four cities, whereas Philadelphia had the highest frequency. All Holt-Winters models and SARIMA models showed that around January 2020, when the first COVID-19 case occurred, crime was the same for all four cities except Philadelphia, where crime had dropped for a particular time until it increased again. There was no clear evidence to suggest that crime was affected during the COVID-19 pandemic and the social unrest during the protests.

Keywords *COVID-19, crime, Black Lives Matter, protests, social unrest, SARIMA, Holt-Winters.*

1. Introduction

As COVID-19 spread from Wuhan, China, to the entire world, the United States not only had to start dealing with the spread of the virus but also dealt with the rise in social unrest over police brutality. The first reported case was on January 20, 2020, in Seattle, Washington, cited in Ashby (2020). Various observations have been made about the number of reported cases and deaths in the United States. USAFacts (n.d.) estimated 24,603,888 known cases, 170,957 reported cases, 409,728 known deaths, and 3,342 newly reported deaths on January 23, 2021. The Centers for Disease Control and Prevention COVID Data Tracker (n.d.) reported 24,876,261 cases and approximately 171,844 new cases; 416,010 deaths and approximately 3,414 newly reported deaths. The COVID Tracking Project (n.d.) launched by *The Atlantic* had reported 24,800,354 cases, 142,949 new cases, and 410,212 reported deaths.

Last, the John Hopkins University COVID-19 Dashboard reported 25,128,825 cases and 419,228 deaths (Johns Hopkins University 2021). On February 29, 2020, the United States recorded its first death and announced travel restrictions (Campedelli et al. 2021). To prevent the spread of COVID-19, social distancing and lock-down policies were issued throughout the United States as it began to spread. Social distancing measures included a mandate that individuals maintain a distance from one another when in public, limitations on gatherings, operation of businesses, and instructions to remain at home (Mohler et al. 2020). Distancing measures simultaneously affected the daily routines and social interactions of millions of people (Campedelli et al. 2021). Daily commuters were forced to spend their days at home; household members shared the same living spaces throughout the entire day; people could not connect to their peers in person but only telematically (Campedelli et al. 2021). Eventually, these restrictions changed when social unrest occurred.

The occurrence of social unrest within the United States started on May 25, 2020, when George Floyd was killed by a police officer from the Minneapolis Police Department, cited in Dave et al. (2020); the officer was eventually charged with murder, cited in Dave et al. (2020). The protests demanded police reform. The first protest took place in Minneapolis, Minnesota, the day after the incident on May 26, 2020 (Dave et al. 2020). By June 16, 2020, the rise of protests had begun nationwide, primarily within larger cities; many lasted more than 3 days and one third saw at least 1,000 participants (Dave et al. 2020). Researchers at the Armed Conflict Location and Event Data Project (2020) analyzed more than 7,750 demonstrations from 2,400 locations between May and August, and found that less than seven percent of the protests were violent (Li 2020). Despite the media focus on looting and vandalism during the protests, there is little evidence to suggest that the demonstrators engaged in widespread violence (Armed Conflict Location & Event Data Project 2020). In addition, when demonstrations did turn hostile, there were reports of infiltrators having instigated violence (Armed Conflict Location & Event Data Project 2020). Between May 24 and August 22, more than 360 counter-protests were recorded around the country, accounting for nearly 5% of all demonstrations (Armed Conflict Location & Event Data Project 2020). Of these, 43–nearly 12%–turned violent (Armed Conflict Location & Event Data Project 2020).

During the pandemic, many states had issued a state of emergency and had set out guidelines to reduce the spread of the virus. Since these guidelines were issued, some academic literature has claimed that crime had decreased, whereas others have claimed that there was no decrease for particular crimes. San Francisco experienced a 43% decline, cited in Felson et al. (2020). New York, San Francisco, Los Angeles, Chicago, and Philadelphia all reported declines in assault/battery and robbery (McDonald and Balkin 2020). According to Abrams (2021), Pittsburgh, New York City, San Francisco, Philadelphia, Washington D.C., and Chicago experienced a drop in overall crime rates above 35%; Cincinnati and Seattle did not see a notable difference. The specific type of crimes that experienced a large decline was theft (28%), simple assault (33%), and rape (38%) Abrams (2021). Meanwhile, other crimes had major increases: commercial burglaries (38%) and car theft in certain cities (Abrams 2021). Violent and property crime saw a 19% decline (Abrams 2021).The sharp rise of nonresidential burglaries is likely associated with property damage and looting at the beginning of protests against police violence, cited in Zhang et al. (2020) For Philadelphia, Los Angeles, and Chicago, there was at least a 25% decrease in stops by the police except for Seattle (Abrams 2021). Cassell (2020) reported an increase in homicides and shootings across the country, starting in late May and continued through June and July 2020. Ashby (2020) reported that there was no evident relationship between COVID-19 and any type of crime between the first case within the United States on January 20 and the beginning of March.

This study examined the changes in the rate of crime during the pandemic and periods before and after social unrest. Based on data collected from January 2018 to after the onset of the social unrest, Seattle, San Francisco, Los Angeles, and Philadelphia were chosen. Holt-Winters and SARIMA models were used to measure crime during the pandemic and before and after social unrest. Data visualizations were then used to determine the areas where crime was most prevalent. This might help reduce crime even further if we know what areas are heightened by crime and what prevents individuals from committing crime.

2. Main Text and Analysis

Crime data recorded by each city’s police department was collected by using the Open Data sites of the four cities mentioned in this study. To acquire the needed information for Los Angeles, we had to combine two datasets because the Open Data site did not contain one dataset with the relevant data. The pre-processing procedures were executed by using the Python programming language in a Jupyter notebook. Crimes were studied in a generalized sense rather than by selecting a specific one. Because both the pandemic and social unrest are considered different types of events, analyzing the effects of crime rate for both may not provide sufficient information. To address this issue, we developed two cases to see a proper estimate of the frequency of crime before and after the pandemic and

Table 1: The training and test data frequencies for each city and both cases.

Case	Training Set	Test Set	Dates for Training Set	Dates for Test Set	Dates for Test Set
1	Seattle	750	367	Jan 1, 2018–Jan 20, 2020	Jan 21, 2020–Jan 21, 2021
	San Francisco	750	370	Jan 1, 2018–Jan 20, 2020	Jan 21, 2020–Jan 24, 2021
	Los Angeles	750	364	Jan 1, 2018–Jan 20, 2020	Jan 21, 2020–Jan 18, 2021
	Philadelphia	750	370	Jan 1, 2018–Jan 20, 2020	Jan 21, 2020–Jan 24, 2021
2	Seattle	876	241	Jan 1, 2018–May 25, 2020	May 26, 2020–Jan 21, 2021
	San Francisco	876	244	Jan 1, 2018–May 25, 2020	May 26, 2020–Jan 24, 2021
	Los Angeles	876	238	Jan 1, 2018–May 25, 2020	May 26, 2020–Jan 18, 2021
	Philadelphia	876	244	Jan 1, 2018–May 25, 2020	May 26, 2020–Jan 24, 2021

the recent civil unrest. Case 1 focuses on the dates before the pandemic started (January 1, 2018) and the date of the first COVID-19 case in the United States (January 21, 2020). The test data was obtained between the dates of January 21, 2020, and the present time. Through the analysis of case 1, it is expected to see how the coronavirus pandemic had affected the crime rate. In case 2, we focused on the dates between before the pandemic began (January 1, 2018) and when George Floyd died (May 25, 2020). Dates between May 26, 2020, and the present were used to generate the test data. The training and test data frequencies for case 1 and case 2 are shown in Table 1.

The Holt-Winters model for case 2 (Figure 2) shows identical results as case 1 (Figure 1). For case 2, the focus is the day of George Floyd's death (May 25, 2020) and the subsequent social unrest. Seattle's crime rate spiked to nearly 800, with an eventual decline after May 2020. San Francisco experienced a slight increase, with a frequency near 400. Los Angeles had the lowest crime rate, below 200, with a slight uptick between May 2020 and September 2020. Philadelphia had a frequency of 400, with an eventual increase after May 2020 to above 800, and sometime after September 2020, reached precisely 800. The SARIMA models (refer to Figures 3 and 4) demonstrate identical outcomes as the Holt-Winters models in both cases. Bar charts and lollipop charts were used to indicate crime-prone areas because the aforementioned models did not reveal such information. The first bar chart (Figure 5) shows the number of crimes versus the types of locations in Los Angeles. The Public Transportation - MTA Line Specific area is the most crime-prone in Los Angeles. Religious institutions, government, and entertainment are the areas with the least crime. It would have been interesting to see what specific crimes occurred where. In Seattle and Los Angeles, we examined which neighborhoods had the highest crime rate. The 77th Street neighborhood in Los Angeles (Figure 6) had the highest incidence of crimes, exceeding 6000, whereas the Devonshire neighborhood had fewer. Meanwhile, Seattle's Queen Anne neighborhood (Figure 7) had the highest number of crimes, exceeding 8000, whereas the Commercial Harbor Island and Commercial Duwamish neighborhoods had fewer. San

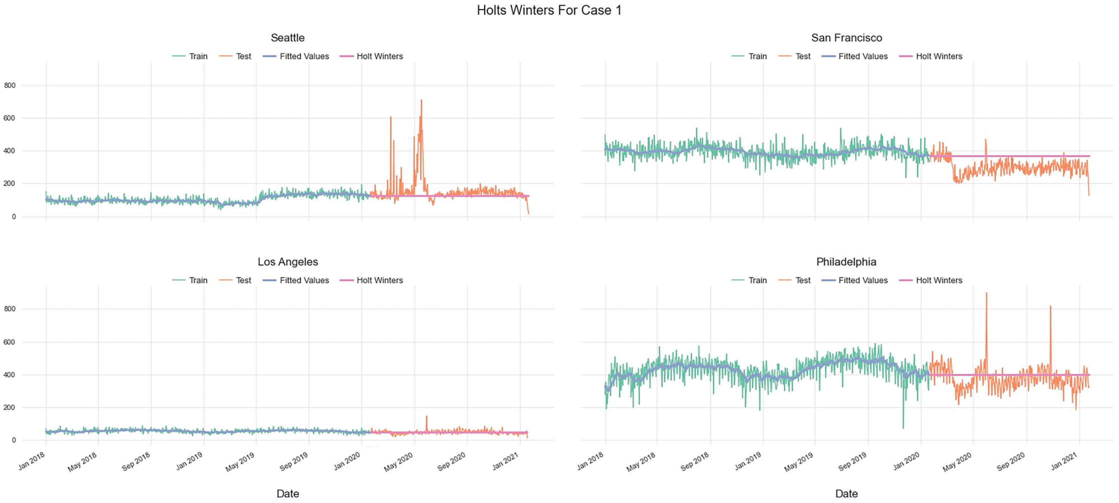


Figure 1: Holt Winters for case 1. Holt-Winters plots for all four cities for case 1.

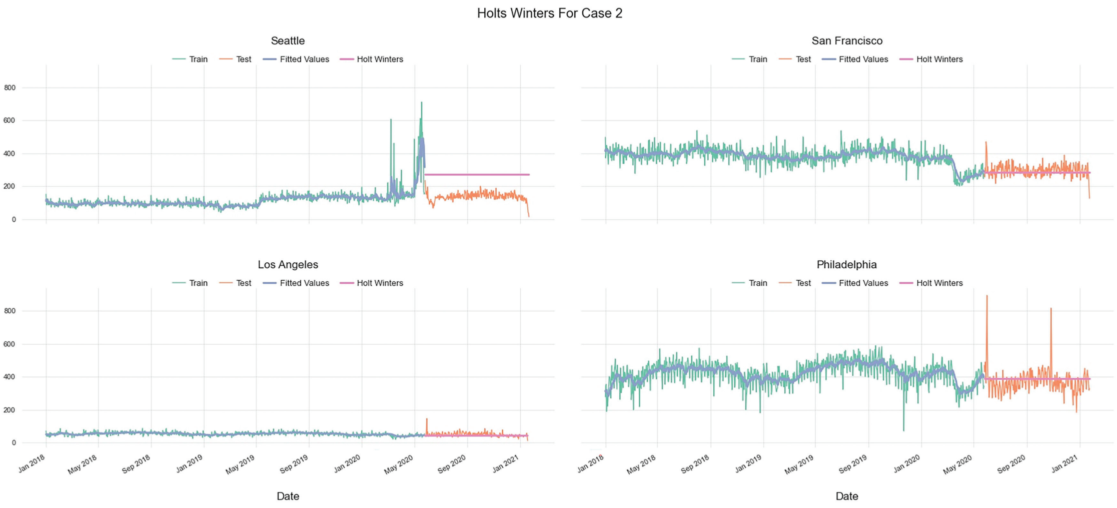


Figure 2: Holt Winters for case 2. Holt-Winters plots for all four cities for case 2.

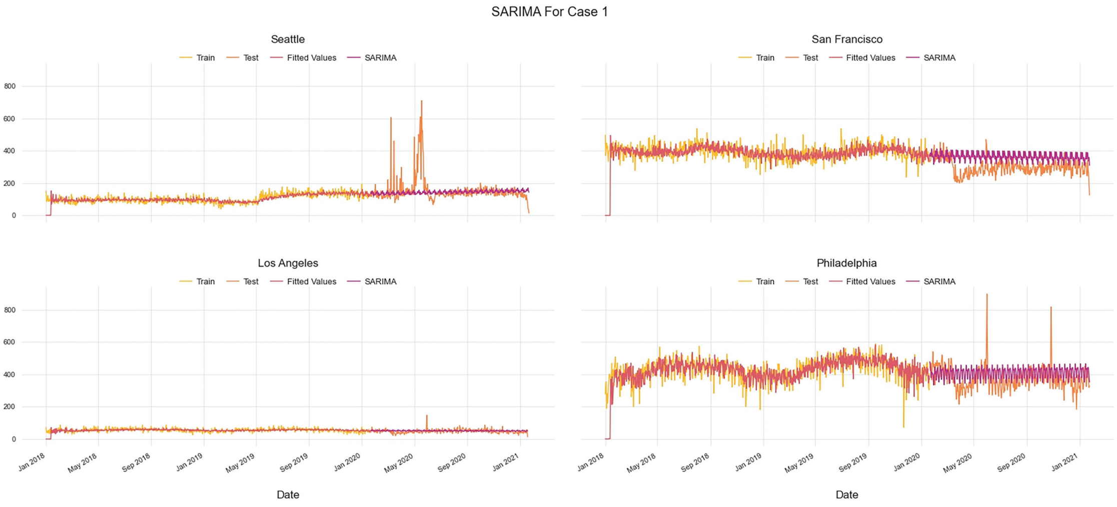


Figure 3: SARIMA for case 1. SARIMA plots for all four cities for case 1.

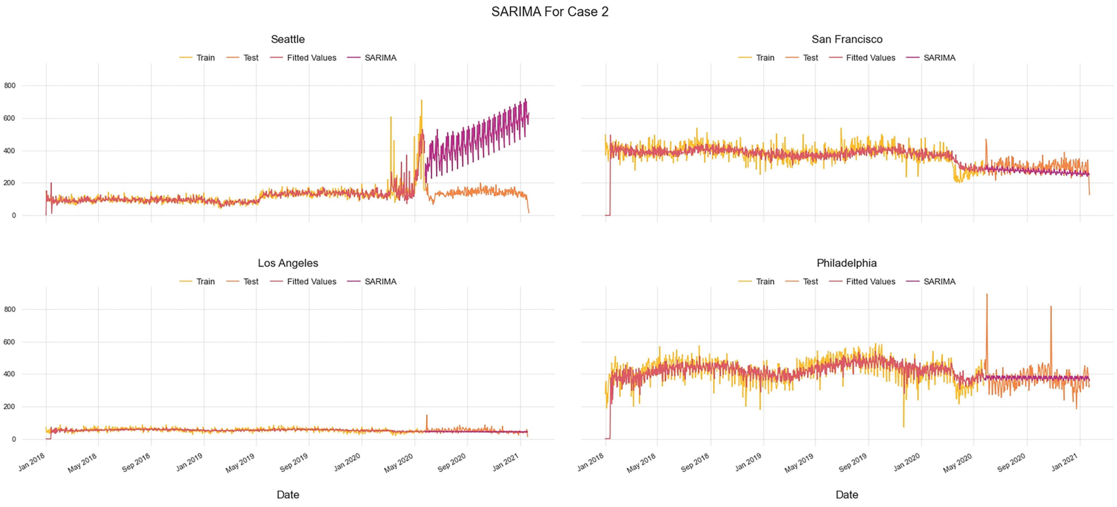


Figure 4: SARIMA for case 2. SARIMA plots for all four cities for case 2.

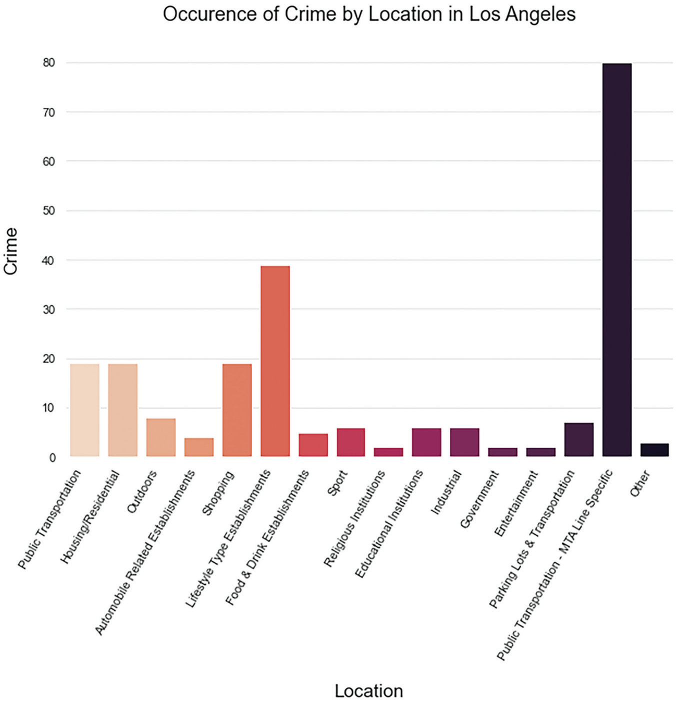


Figure 5: Occurrence of crime by location in Los Angeles. Bar plot, showing the crime frequency based in each type of physical location in Los Angeles.

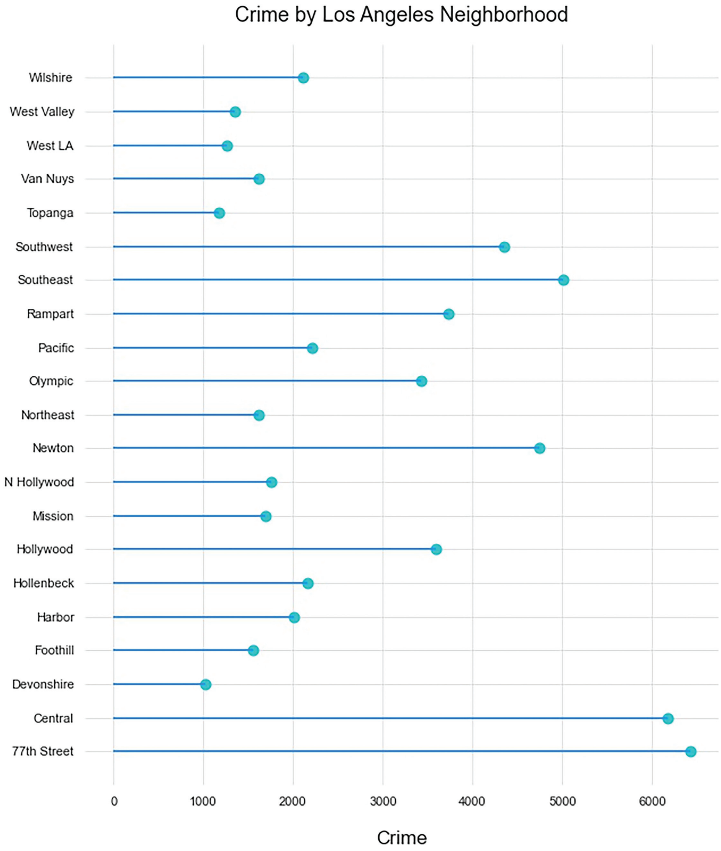


Figure 6: Crime by Los Angeles neighborhood. Lollipop chart, showing the number of crimes within each Los Angeles neighborhood.

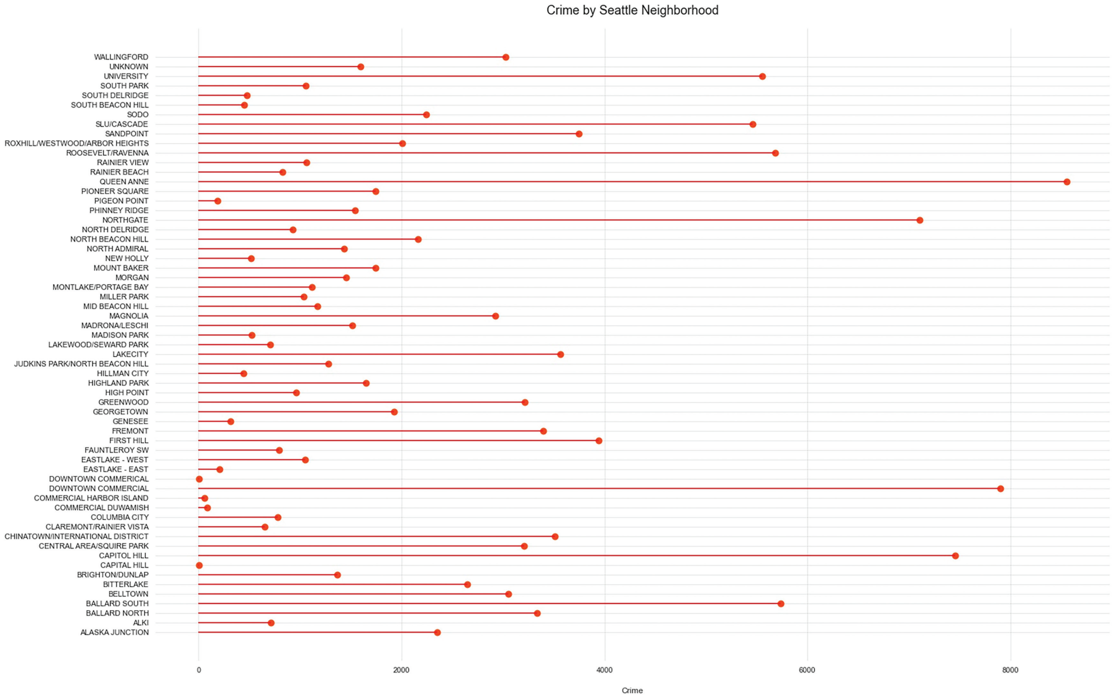


Figure 7: Crime by Seattle neighborhood. Lollipop chart, showing the number of crimes within each Seattle neighborhood.

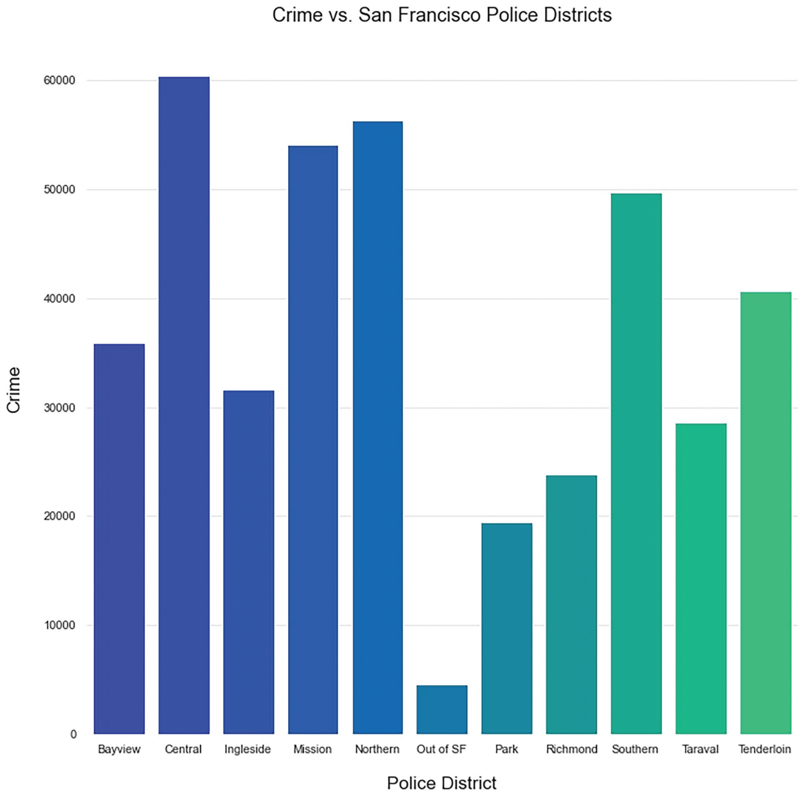


Figure 8: Crime vs. San Francisco police districts. Bar chart, showing the number of crimes within each San Francisco Police District.

Francisco's Central police district had the most crimes compared with other police districts in the city, as shown in Figure 8. The Holt-Winters model used for Los Angeles' training data had the best mean error value, of –0.05, for both case 1 and case 2 when evaluating the models presented in Tables 2 and 3 (as shown in the Appendix section).

2.1. Internal and External Threats to Validity

There are no internal validity threats present in this research because the research question was answered by using available datasets. It is uncertain if these datasets obtained from existing resources contain inaccurate information. In addition, the external validity of this research may be at risk due to the use of only four datasets. To perform a comprehensive analysis and generalize this research, more datasets from various cities would be required.

3. Conclusion

Based on our analysis, Philadelphia had the highest crime frequency, whereas Los Angeles had the lowest crime frequency. Public areas, including public transportation, especially MTA transportation lines were hot spots for crime in Los Angeles. The 77th Street neighborhood in Los Angeles and the Queen Anne neighborhood in Seattle had the largest number of crimes. San Francisco's Central police district had the most crimes. Despite widespread media coverage of the rioting and vandalism that occurred, the claims were not entirely accurate. There was no significant evidence that indicated any changes in crime within any of the four cities during the pandemic, although we did see a brief uptick in crime during the protests.

There is no actual evidence that suggests how crime behaves during a massive pandemic, however, there is evidence that suggests that crime does change during a massive change in routines (Abrams 2021). For a crime to occur, there must be an opportunity, a potential victim, and a proper location. When neither of these are available, a crime cannot be committed. Due to the restrictions in place, it is difficult for many types of criminal offenses to occur. Social distancing has made it difficult for attractive areas (bars, nightclubs, stores, malls, etc.) for criminal offenses to occur because these places have attracted fewer people. According to Campedelli et al. (2021), members of a community can be modeled as potential offenders, potential victims, and potential guardians, moving around and interacting in a socio-geographical space. In moving from these premises, routine activity theory postulates that offenders and victims (or targets) usually meet during daily, non-criminal activities (Campedelli et al. 2021). Behavioral decisions then determine how the various agents react to each other's presence and actions (Campedelli et al. 2021). Crime occurs in the context of the everyday routines as the three factors converge in space and time: a motivated offender, a victim or potential target, and the absence of a capable guardian (Campedelli et al. 2021). Crime is known to be heavily context dependent, and the contexts of different cities vary considerably (Ashby 2020). Crime opportunities and places where crimes occur are likely to change drastically from past observations and experiences (Ashby 2020).

Understanding relationships between COVID-19 and crime requires some estimate of how much crime would be expected to occur in the absence of the pandemic (Ashby 2020). This is difficult because so many factors influence how much crime occurs (Ashby 2020). There is considerable variation in the nature of violent crimes, from those with greater financial incentive, such as robbery, to those that are often associated with alcohol or drug use, such as assault (Abrams 2021). Assaults with a deadly weapon, homicides, burglaries, intimate partner assaults, and stolen vehicles do not report any significant effect (Campedelli et al. 2021). There is an extensive agreement over the fact that burglars prefer to target unoccupied homes (Campedelli et al. 2021). People forced at home will guard their houses for longer hours, minimizing their exposure to burglaries (Campedelli et al. 2021). There was an increase in non-residential burglaries because individuals spent more time at home and other buildings were left less occupied. The results for car theft and theft from cars varied substantially by city (Abrams 2021). Drops in drug crimes were by far the greatest, with most cities reporting data that showed massive declines of more than 65% (Abrams 2021). Police departments modified policies, including de-emphasizing particular types of crimes, such as drugs (Abrams 2021), and no longer making arrests for some crimes (Abrams 2021). Jails and prisons have seen some of the most severe outbreaks and, as such, a number have released inmates early (Abrams 2021). Courts shut down and deferred cases (Abrams 2021), which may result in fewer prosecutions (Abrams 2021). Together, this has resulted in a change in the opportunities for crime, probability of observation, capture, arrest, prosecution, and penalty (Abrams 2021).

Results of this study suggest that policing should be focused on areas most likely prone to crime, even during a pandemic, to prevent crime from ever happening. We were not able to find clear data on the type of victims, depending on the specific crime and location. This type of data would have further helped see more information on the sort of crimes that would have affected the population within all four cities. For instance, if the 77th street neighborhood in Los Angeles were having more theft specifically within public areas, then more guardianship or policing within that area would be needed. It would have been interesting to see what specific crimes occurred in which sort of location. There was no significant evidence that suggested that crime was severely impacted during the pandemic. All the cities showed a temporary rise in crime frequency during the protests and social unrest, returning to their previous levels thereafter. Future research should further investigate which locations were possibly impacted by crime during the pandemic because areas with a lack of policing might have been impacted during the pandemic and even the social unrest that previously took place. This sort of data would help understand which areas needed more policing so crimes would eventually decrease overtime. Crime rose as a result of police resources being redirected to areas where the protests occurred, while areas that were typically policed were neglected. When those areas were being policed, there was a control on the frequency of crimes that took place, hence making communities much safer and livable.

Appendix

Table 2: Accuracy score for all models for each city for case 1.

		Training Set					Test Set				
City	Model	MAPE	MAE	ME	MPE	RMSE	MAPE	MAE	ME	MPE	RMSE
Seattle	Holt-Winters	26.25	12.62	0.299	–5.96	84.42	20.33	37.12	28.49	6.21	84.42
	SARIMA	27.09	15.79	1.35	–4.89	22.58	22.71	36.56	10.59	–7.18	81.84
San Francisco	Holt-Winters	9.27	29.63	–93.88	–1.38	39.01	27.81	74.96	–70.25	–26.70	84.89
	SARIMA	11.74	41.98	6.47	0.487	68.96	26.06	70.75	–62.49	–24.05	82.85
Los Angeles	Holt -Winters	17.54	7.57	–0.05	–3.96	11.65	18.81	8.39	–0.165	–6.08	11.65
	SARIMA	18.73	8.76	1.07	–1.86	11.98	21.85	9.15	–3.16	–12.67	12.43
Philadelphia	Holt-Winters	16.05	43.88	0.885	–3.11	75.30	16.05	56.21	–20.69	–9.02	75.30
	SARIMA	18.61	52.56	6.24	–1.82	73.74	19.23	67.69	–34.49	–12.81	85.59

Table 3: Accuracy score for all models for each city for case 2.

		Training Set					Test Set				
City	Model	MAPE	MAE	ME	MPE	RMSE	MAPE	MAE	ME	MPE	RMSE
Seattle	Holt -Winters	35.57	18.96	0.709	–11.80	139.32	115.59	136.99	–136.99	–115.59	139.32
	SARIMA	37.98	22.84	0.198	–12.29	43.97	269.34	332.17	–332.17	–269.34	351.49
San Francisco	Holt -Winters	14.55	29.92	–84.83	–2.85	38.99	9.46	27.95	10.49	1.95	37.99
	SARIMA	15.18	39.58	5.63	–1.05	65.42	11.72	34.82	25.38	7.05	44.91
Los Angeles	Holt -Winters	19.77	7.45	−0.05	–4.65	12.86	17.39	8.69	3.29	1.18	12.86
	SARIMA	20.38	8.48	0.902	–2.79	11.57	17.89	9.11	5.29	5.46	13.53
Philadelphia	Holt -Winters	17.36	43.77	0.588	–3.42	74.65	14.14	50.35	–16.74	–7.73	74.65
	SARIMA	18.08	50.51	6.20	–2.04	70.29	13.11	48.12	–4.29	–4.28	73.10

References

Abrams, D. S. 2021. "COVID and Crime: An Early Empirical Look." *Journal of Public Economics* **194**: 104344. doi: 10.1016/j.jpubeco.2020.104344

Armed Conflict Location & Event Data Project (ACLED). 2020. "Demonstrations & Political Violence in America: New Data for Summer." Accessed January 25, 2021. https://acleddata.com/2020/09/03/demonstrations-political-violence-in-america-new-data-for-summer-2020/

Ashby, M. P. J. 2020. "Initial Evidence on the Relationship between the Coronavirus Pandemic and Crime in the United States." *Crime Science* **9**, no. 1: 6. doi: 10.1186/s40163-020-00117-6

Campedelli, G. M., A. Aziani, and S. Favarin. 2021. "Exploring the Immediate Effects of COVID-19 Containment Policies on Crime: An Empirical Analysis of the Short-Term Aftermath in Los Angeles." *American Journal of Criminal Justice* **46**, no. 5: 704–727. doi: 10.1007/s12103-020-09578-6

Cassell, P. G. 2020. "Explaining the Recent Homicide Spikes in U.S. Cities: The "Minneapolis Effect" and the Decline in Proactive Policing." *Federal Sentencing Reporter* **33**, no. 1–2: 83–127. https://ssrn.com/abstract=3690473. doi: 10.1525/fsr.2020.33.1-2.83

Centers for Disease Control and Prevention (CDC). n.d. "COVID Data Tracker." Accessed January 24, 2021. https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days

Dave, D. M., A. I. Friedson, K. Matsuzawa, J. J. Sabia, and S. Safford. 2020. "Black Lives Matter Protests and Risk Avoidance: The Case of Civil Unrest during a Pandemic" (Working paper no. w27408). *National Bureau of Economic Research*. doi: https://doi.org/10.3386/w27408

Felson, M., S. Jiang, and Y. Xu. 2020. "Routine Activity Effects of the Covid-19 Pandemic on Burglary in Detroit, March, 2020." *Crime Science* **9**, no. 1: 1–7. doi: https://doi.org/10.1186/s40163-020-00120-x

Johns Hopkins University. 2021. "COVID-19 Case Tracker. John Hopkins Coronavirus Resource Center." Accessed January 23, 2021. https://coronavirus.jhu.edu/map.html

Li, W. 2020. "Is Violent Crime Rising In Cities Like Trump Says? Well, It's Complicated." The Marshall Project. Accessed January 25, 2021. https://www.themarshallproject.org/2020/09/25/is-violent-crime-rising-in-cities-like-trump-says-well-it-s-complicated

McDonald, J. F., and S. Balkin. 2020. "The COVID-19 and the Decline in Crime." Available at SSRN 3567500. doi: https://doi.org/10.2139/ssrn.3567500

Mohler, G., A. L. Bertozzi, J. Carter, M. B. Short, D. Sledge, G. E. Tita, C. D. Uchida, and P. J. Brantingham. 2020. "Impact of Social Distancing during COVID-19 Pandemic on Crime in Los Angeles and Indianapolis." *Journal of Criminal Justice* **68**: 101692. doi: 10.1016/j.jcrimjus.2020.101692

The COVID Tracking Project. n.d. "The Data." Accessed January 23, 2021. https://covidtracking.com/data/

USAFacts. n.d. "US COVID-19 Cases and Deaths by State." Accessed January 23, 2021. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Zhang, Z., D. Sha, B. Dong, S. Ruan, A. Qiu, Y. Li, J. Liu, and C. Yang. 2020. "Spatiotemporal Patterns and Driving Factors on Crime Changing during Black Lives Matter Protests." *ISPRS International Journal of Geo-Information* **9**, no. 11: 640. Nov. doi: 10.3390/ijgi9110640

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.32

Machine Learning Study: Identification of Skin Diseases for Various Skin Types By Using Image Classification

Gulhan Bizel
Saint Peter's University
gbizel@saintpeters.edu

Albert Einstein
Saint Peter's University
aeinstein@saintpeters.edu

Amey G. Jaunjare
Saint Peter's University
ajaunjare@saintpeters.edu

Sharath Kumar Jagannathan
Saint Peter's University
sjagannathan@saintpeters.edu

Abstract

Increased machine learning methods have helped improvise human interaction with digital devices, which helps in skin disease identification, prediction, and classification by using algorithms. Image classification for skin disease application algorithms can detect Caucasian skin tones but poorly performs when analyzing other skin colors. In this research, a deep learning algorithm was used to address the problem that other applications perform poorly with the classification of skin disease types.

Convolutional neural network, a machine-learning algorithm was used to classify images and add the predicted images within the data set. The images in the data set covered a lot of patient factors, such as age, sex, disease site (e.g., hand, feet, head, nail), skin color (white, yellow, brown, black), and different periods of lesions (early, middle, or late). Multiple private applications can detect skin diseases during the analysis. For the darker color skin population, the performance was poor, and skin cancer detection was not possible even with the help of image recognition. This research aims to conduct an analysis of visual searches within skin-related health searches to identify opportunities to provide digital health consumers with visual search results that are more representative of America's diverse populations.

Keywords *convolutional neural network (CNN), neural network, deep learning, machine learning, image classification.*

1. Introduction

A widely used dermatology diagnostic technique is the examination of diseased skin against healthy skin. A field that focuses on skin is now recognizing the importance of skin color (Enderling 2019). Images are the critical resource for diagnoses in dermatology. The lack of images of darker skin tones creates a barrier for proper treatment. Skin conditions that present unusual color patterns, for example, redness in light skin, can be harder to see in dark skin. Physicians who lack diagnostic experience with such image patterns may struggle with diagnosing people of color. It is unknown whether "Dreamscape Immersive," a company that focuses on creating immersive virtual reality (VR) experiences, will improve the diagnostic accuracy for all types of pigmented skin lesions or only for those that are melanocytic (Errichetti 2020). The issue is becoming more serious for dermatologists and may influence different outcomes for a different color of skin. The lack of images is one reason why dermatologists may misdiagnose skin illnesses for patients who are darker skinned, which may result in serious health-threatening medical issues.

A skin lesion is a nonspecific term that refers to any change in the skin surface. Skin lesions may have color (pigment); be raised, flat, large, small, or fluid filled; or exhibit other characteristics (Skin Lesions 2022). With current technology, smartphones are showing promise as diagnostic tools. During dermatology appointments, patients can send their physicians pictures of their skin lesions before their visit. Patients who are comfortable with using this technology seem to have adopted this new tool for their health care. Smartphone photographs have provided valuable relevant information for a physician's diagnosis and treatment decisions, which is usually based only on the medical history reported by patients (Hubiche 2016).

Google's artificial intelligence initiatives, often referred to as Google AI, (with its primary location in Mountain View, California and offices and research centers around the world) encompasses a wide range of tools, research, and applications aimed at pushing the boundaries of what AI can do. Google's mission with AI is to make it universally accessible and useful, ensuring that the benefits of AI are spread broadly across all sectors of society. Google AI is used to detect common skin cancer problems, which has improved the clinical trials to treat their patients effectively. There are, according to a Google search history, billions of people searching for answers and information related to skin, hair, and nails. Approximately 2 billion people worldwide experience dermatologic issues, and there is a shortage of dermatologists to diagnose and treat them. Using a search bar makes it difficult to understand what type of skin diseases they have because many of the terms are scientifically complex and difficult for a patient to understand.

Google has developed a web-based application tool in which the users can take their pictures by using their personal devices and can then upload them in that tool. The model analyzes the images and has been programmed to identify 288 common skin disease problems. It then provides information about the disease, which helps the patient research his or her skin problems and suggests common questions asked to a dermatologist. This tool is not for a proper diagnosis, but it assists users in making their decision on what further steps are required (Bui 2021). Although social media is a powerful tool for users to share their information with regard to their health, there is a risk of misleading and inaccurate information when dermatologists are encouraged to present their answers in multiple social media and other applications, which counteract other misleading information (Powell 2019).

In recent years, machine learning has been pivotal in advancing skin cancer diagnosis and understanding. Das et al. (2021) explored the utility of machine learning in skin cancer identification, which has become a significant focus given the rising incidence of skin cancer globally. Their work, published in the *International Journal of Environmental Research and Public Health*, has shed light on the profound implications of machine learning techniques in early skin cancer detection and treatment planning. On a similar note, Li et al. (2021) delved into the application of deep learning for skin disease diagnosis in their review published in *Neurocomputing*. Their comprehensive review underscores the remarkable progress and the challenges encountered in deploying deep learning algorithms for skin disease diagnosis, which contributes to a broader understanding of the potential and the limitations of these technologies.

The quest for efficient and accurate skin disease image classification has led to the development of innovative algorithms and frameworks. Chen et al. (2021) introduced the "Interactive Attention Sampling Network for Clinical Skin Disease Image Classification," presented at the 4th Chinese Conference on Pattern Recognition and Computer Vision. Their work delineates the efficacy of the Interactive Attention Sampling Network in categorizing clinical skin disease images, thus offering a promising avenue for enhancing diagnostic accuracy. Furthermore, the challenge of class imbalance in skin disease recognition was addressed by Yang et al. (2020) through their proposed self-paced balance learning algorithm. Published in the "IEEE Transactions on Neural Networks and Learning Systems," their research presents a robust methodology to mitigate the class imbalance issue, which is prevalent in many classification tasks, especially in clinical skin disease recognition. Through these diversified studies, it is evident that the fusion of machine learning and dermatologic expertise holds immense promise for the evolution of skin disease diagnosis and management.

A skin disease, also known as dermatosis, refers to any condition that affects the skin. These conditions can range from temporary to chronic, mild to severe, and can be caused by various factors, including genetics, environmental factors, allergens, and pathogens. Skin diseases are very often experienced during one's lifetime. Skin diseases prevade all cultures and affect between thirty to seventy percent of individuals (Hay et al. 2014). People can be affected by skin disease anytime during their lifetime. Skin disease is twofold: skin infection and skin neoplasm with thousands of specific skin conditions (Hurt 2012). Skin disease may impact the quality of a person's life, mostly related to his or her psychosocial situation. However, only a small portion of people can recognize these diseases without seeing a physician or dermatologist.

There are several over-the-counter medications for treating the regularly occurring skin diseases in daily life. So, people who decide to use these treatments need to choose the correct medicines without the benefit of seeing a physician. Should people want to take this route, a visual analyzing system would be useful, even if it is not fully accurate. For example, if a skin disease presents itself, a person can submit a photo of the skin condition to his or her physician for an initial diagnosis. Interestingly, there are already several applications that use computer visual techniques to recognize many common skin diseases based on simple photo images.

Although there are several related tools, skin disease recognition has not yet been a hundred percent accurate by image recognition. Skin disease images have no consistently distinct pattern, like a fingerprint. For instance, it is still very difficult to find an accurate description of a simple allergic skin disease such as eczema. In addition, there are many variations, such as contact issues between lesions and surrounding skin or the coloring inside the lesion, which makes it complicated for skin disease recognition. A recent report indicates that performing clinical skin disease recognition by image analysis is of major importance because skin disease is one of the most common diseases that appear in medicine (Esteva et al. 2017).

There are some related developments in skin disease recognition such as disease classification (Barata et al. 2021; Yu et al. 2017) and detection and localization (Marchetti et al. 2018). Examining clinical skin disease images is economical and getting the digital image from a portable electronic device is convenient for patients who can easily run a self-diagnosis. One drawback to self-diagnosis is that there are few open data sources that are needed to develop deep learning technology in this field. The challenge that researchers face is that clinical imaging is easily affected by light intensity, camera angle, uncertain background, and other light-related factors and interferences (Yang et al. 2018). Moreover, most current research addresses binary skin disease recognition problems.

Smartphones can be considered a new potential source of medical data. It has been observed that an increasing number of patients present pictures of their skin problems to their physicians. To benefit from their full utility, the application determined the proportion of patients whose smartphone photos provided information relevant to their diagnosis. Patients are now practicing this type of new technology as part of their own health care. Relevant information by using mobile phone photos can be provided for diagnostic and treatment decisions. A result is determined based on the patient's medical history. Because these reports may influence the patient's diagnosis and treatment, it is crucial to keep the photos in the patient's medical records. With these digital applications, patients can better partner with their physicians and become more self-aware of their health status.

The paper is outlined as follows: the methodology section gives the overview of the convolutional neural network (CNN) model approach used in skin cancer detection of multiple skin types. The results section shows the graphic representation of the number of trainings performed and researched about the model findings. The final section includes the best results found in the model and the best batch size when considering the best accurate score.

2. Material and Methods

2.1 Dataset

In this model, the image data sets used are from skin diseases, ranging from various types of eczema and acne to diverse cancerous conditions. The dimensions of the pixels vary from 640 × 480 pixels to 1640 × 1181 pixels. The images collected were publicly available and have been obtained legally. The objective is to extract the class names from the images and train the model on different skin colors, such as white, black, brown, and the male and female genders. The key task was to import the dataset of images into the dataset and divide the dataset into two parts, training (80% of images) and validation (20% of images) (Skin Disease Images 2023). The different types of classes and datasets that are being used in this research are shown in Figure 1.

2.2 Research Model

The programming language used in this research is Python, developed by Guido van Rossum and first released in 1991, which is one of the most frequently used high-level programming languages. Its high-level, built-in data structure, combined with its dynamic typing and binding, make it very attractive for rapid application development

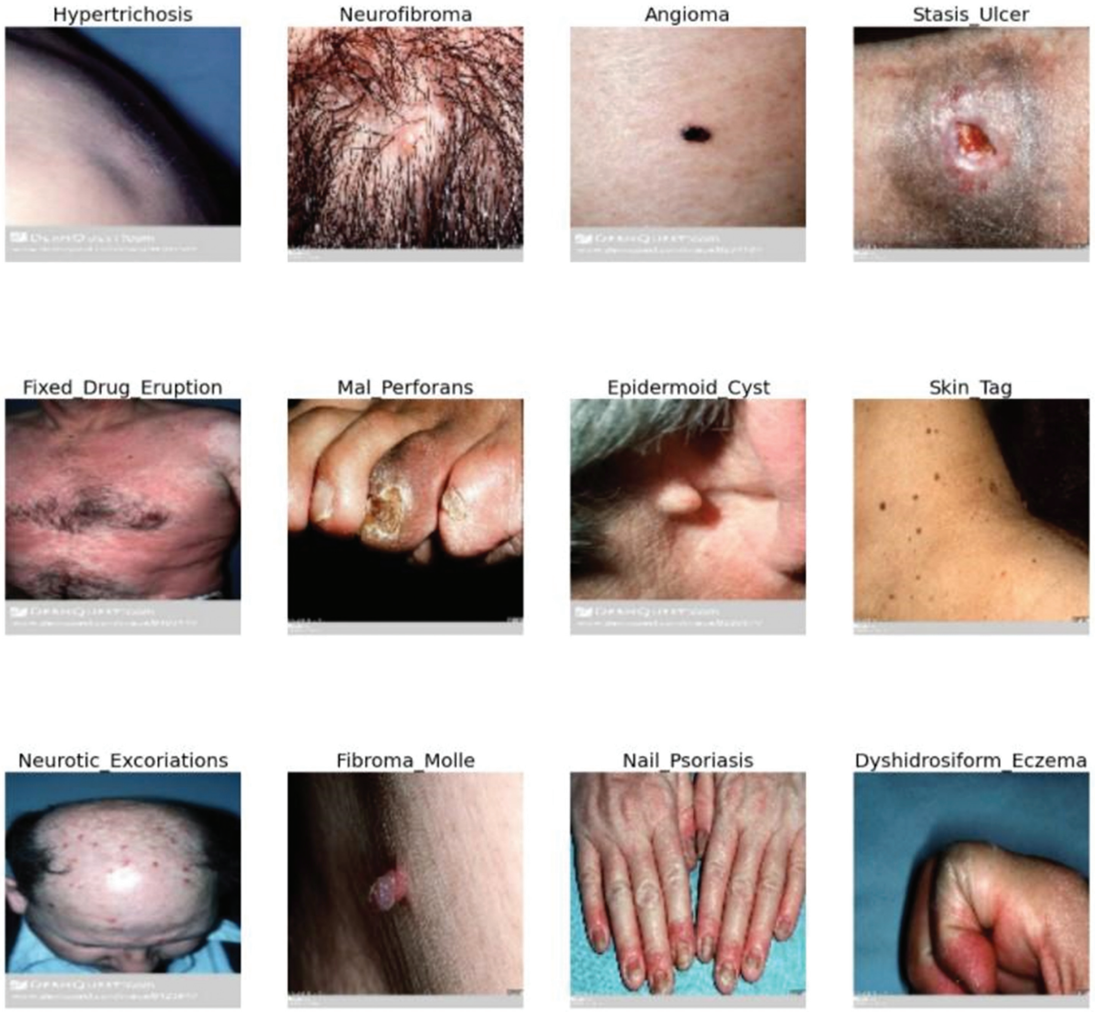


Figure 1: Classes of images within the dataset.

as well as scripting or glue language to connect existing components (https://www.python.org/). The application used is Jupyter Notebook, which is an open-source web-based application that allows the creation and sharing of documents that integrate live code, equations, visualizations, and other multimedia resources. The Jupyter Notebook evolved from the IPython project in 2014. Project Jupyter is a nonprofit initiative primarily based in the U.S. and operates under NumFOCUS, a 501(c)(3) charitable organization in Austin, Texas. Tensor Flow is used to create large-scale neural networks with many layers (Introduction to Convolutional Neural Networks 2022; Abadi et al. 2016) It is mainly used for deep learning or machine learning problems, such as classification, perception, understanding, discovering, prediction, and creation (Kiran 2020). The flow of the model that has been applied in this research is shown in Figure 2, which covers the steps in the following order: training image, validation data, data augmentation, and CNN. Eventually, the applied model led to the output of the study.

2.2.1 Training image and validation data

After extracting class names and dividing the dataset, the images were resized into 128 × 128 pixels, so all the pixels were set to a common size for all images, which made the model train better. A batch size of sixty four was used to help train the models to minimize the loss function (Chang 2021). Mathematically, it is calculated and defined as in Equation (1). The theta shows the number of parameters used in the model, n represents the number of training samples used for training, i is the single element of the training dataset, and xi is the loss function of the single element of each training sample. Using batch with the option of other hyperparameters will help improve training performance and increase the accuracy of the model. The data were then shuffled by using the default true

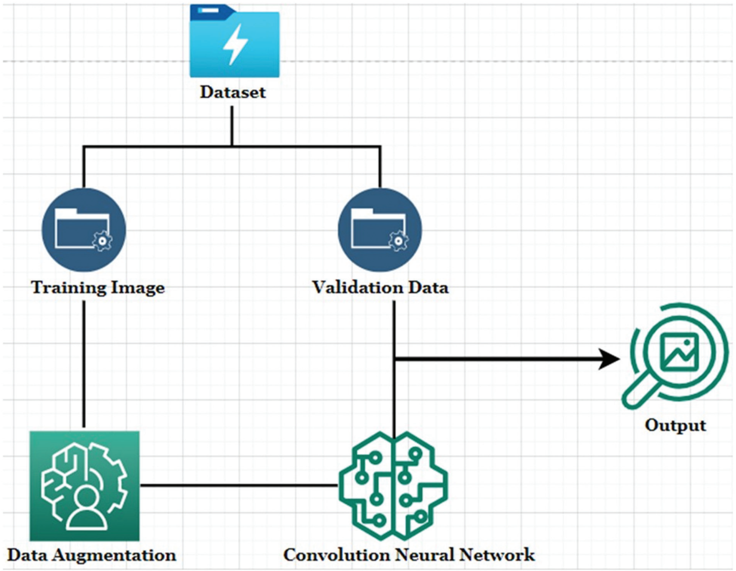


Figure 2: Flow chart of the research model.

function, which determined if the data needed to be shuffled, and arranged the data in alphanumeric order. On creating data training and validation cache function,

$$z(\theta) = 1\Big/n \sum_{(i=1)}^{n\,x_{i(\theta)}} \qquad (1)$$

2.2.2 Data augmentation

Overfitting is the main concern while training the dataset. When the dataset trains too long, it starts to learn noise in the dataset, which matches the irrelevant data once the model is trained. To overcome this problem, data augmentation was used to make the data more stable. Data augmentation was not only used for overfitting but also to increase the accuracy performance of the model. Data augmentation was used for audio, text, images, and other types of data. In this research, data augmentation was used for images when images were rotated vertically and horizontally so that the cancer names and augmented images were trained together to reduce overfitting and increase the model's performance. Data augmentation is useful for improving the performance and outcomes of machine learning models by forming new examples for training of the datasets. If the dataset in a machine learning model is rich and sufficient, then the model performs better and more accurately (Takimoglu 2021). The image shown in Figure 3 is an example of data augmentation performed in the dataset before applying to the model training. It shows the augmentation image of basal cell carcinoma disease from various angles.

2.2.3 Deep learning and CNN classification

A neural network works like the human brain where each neuron is programmed to solve complex problems where neurons learn from other neurons to provide the required outcome. Neurons are the connection between the hidden layers and output layers, which split into parts of the image that are true or false (1 or 0). Neurons have many hidden layers that help in the segmentation of images into kernels to detect the type of image. The output layer is the output of all the segmented images performed by the hidden layers (Shinozaki 2021). In this paper, the CNN classification is being used, which is a part of deep learning for the prediction of images with regard to skin cancers.

Deep learning is a subset of a machine learning model used to perform complex tasks. This means that algorithms are trained like the human brain, and can perform speech recognition, image identification, and prediction (What is Deep Learning 2022). Deep Learning uses layers of neurons to predict the patterns in the input datasets to present the outcome. One of the deep learning models used in this research was the CNN. The CNN is described as the heart of deep learning, which is primarily used for the classification of images, clustering images, and object detection. The process of detection of images is processed through tensors, which are known to be an array of numbers with additional dimensions, is shown in Figure 4. CNN receives the input of images and converts them into RGB

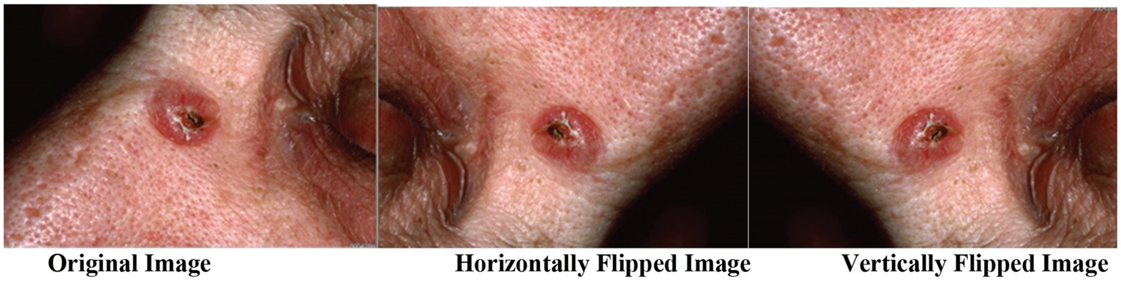

Original Image **Horizontally Flipped Image** **Vertically Flipped Image**

Figure 3: Augmentation image of basal cell carcinoma disease from various angles.

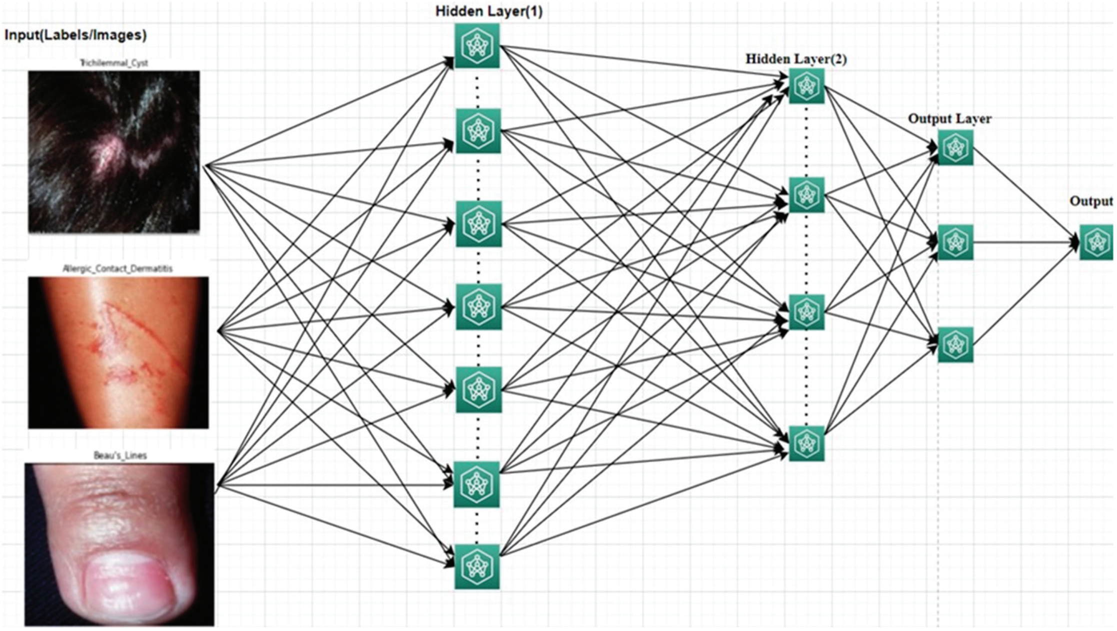


Figure 4: Example of tensors work in the convolutional neural network (CNN).

(red, blue, and green) stacked one above the other, where they are measured by the number of pixels converting them into matrices of multiple dimensions (Liu et al. 2021).

CNN tensors convert the image into an array of matrices which in turn are divided into RGB colors stacked on top of one another. The filter then weights the increases in the strength of the connection as shown in Figure 4. Bias is known to be constantly represented by the number bias 1, which is marked in the right bottom corner in red. According to CNN, bias vector is used to extract the features of the images and determine what classes need to be assigned for that image. CNN takes the images from the raw pixels and trains the images to extract their meaningful features for better prediction and classify the whole image.

CNN is supervised machine learning in which the labels of the classes are trained together for the outcome respective of the class to which the object belongs. CNN trains through the inputted images and works through the class labels and compresses them together. CNN is a multilayer neural network, which has several hidden layers stacked up one after the other, which allows the neural network to learn complex features. The hidden layers in the CNN comprise convolutional layers that have activation layers (ReLU), max pool layer, fully connected layer used for learning and predicting the output (Introduction to Convolutional Neural Networks 2022).

Convolution layers as the primary parameter for a CNN model are represented in Figure 5. These layers are composed of multiple filters that are defined by the height, width, and depth of the input image, which converts the image into weighted matrices known as kernels. Filters range in between $3 \times 3''$ to $11 \times 11''$ which determines the dimension's size of the image. Filters will perform element-wise multiplications, which result in values that will determine the edge and batch of the color in an image. It is computed as in Equation (2) (McDermott 2022),

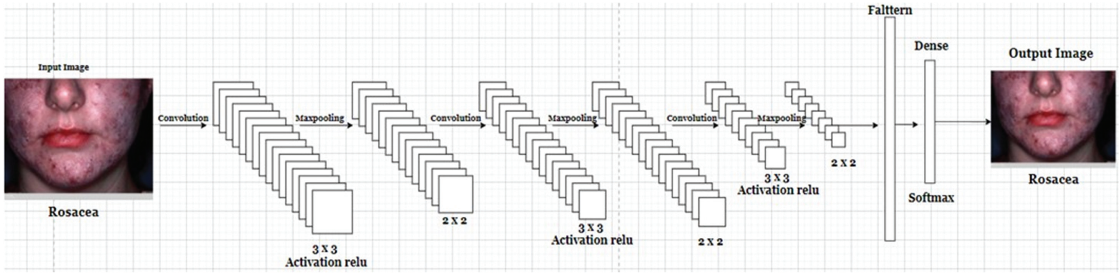
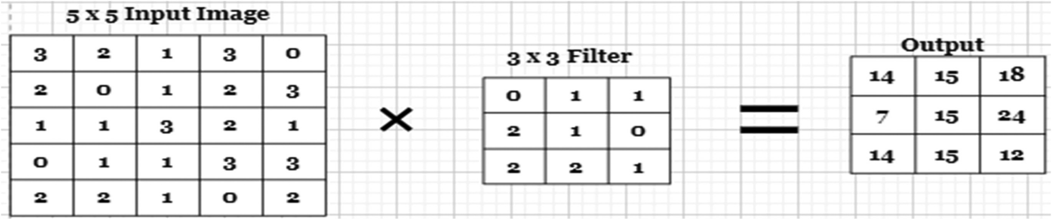


Figure 5: Convolutional neural network (CNN) model flow used in this research with 13 layers.



5 x 5 Input Image

$$\begin{bmatrix} 3 & 2 & 1 & 3 & 0 \\ 2 & 0 & 1 & 2 & 3 \\ 1 & 1 & 3 & 2 & 1 \\ 0 & 1 & 1 & 3 & 3 \\ 2 & 2 & 1 & 0 & 2 \end{bmatrix} \times \underset{\text{3 x 3 Filter}}{\begin{bmatrix} 0 & 1 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix}} = \underset{\text{Output}}{\begin{bmatrix} 14 & 15 & 18 \\ 7 & 15 & 24 \\ 14 & 15 & 12 \end{bmatrix}}$$

(3x0) + (2x1) + (1x1) + (2x2) + (0x1) + (1x0) + (2x1) + (3x1) = 14

Figure 6: Example of input image filtered by using con2d with a 3 × 3 kernel.

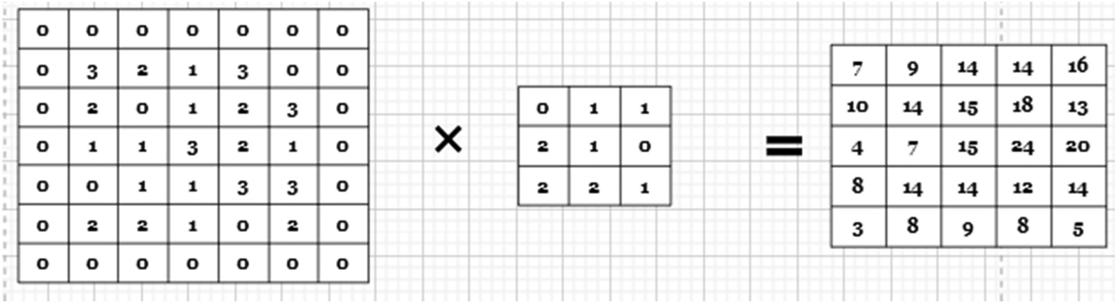
in which the input image is denoted by *i*, the filter of the image is known as *f*, and *m* and *n* are the rows and columns of the kernels given, respectively. Features of the images are extracted by using kernels in which convolution is a two-dimensional array, which contains correlations of the image with respect to the number of filters applied.

$$z[\mathrm{m,n}] = (\mathrm{i} * \mathrm{f})[\mathrm{m,n}] \tag{2}$$

The input image, which is n × n matrix, and the filtered image with m × m matrix, which results in the output size as (n – m + 1) × (n – m + 1), is shown in Figure 6, and shows the calculation of the values slide of one column to the right each time and slides down one row at a time. This is called striding, which is represented as stride = k and the following result is written as ([n – m]/k + 1) × ([n – m]/k + 1) (DeepAI 2020).

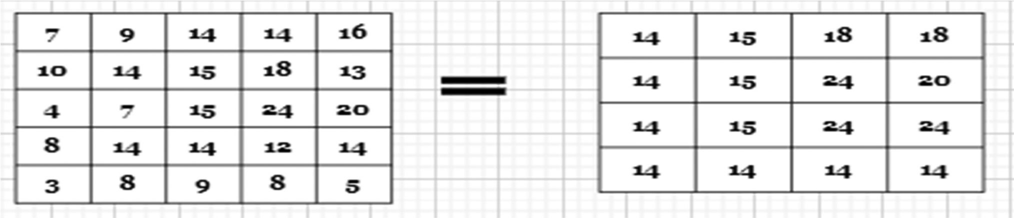
From Figure 6. Strides with the 3 × 3 matrix reduce the output size so to retain the same input image size, which is the 5 × 5 matrix. Padding is used to obtain the original input size by adding zeros at the edges of the image array, as shown in Figure 7.

As shown in Figure 7, the padding is added in the first layer because the image size remains the same in which the equation is represented as input of the image size, which is n × n matrix, and filter size, which is m × m matrix, when following the output size as (n + 2p – m + 1) × (n + 2p – m + 1). Convolutions with the striding and



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 2 & 1 & 3 & 0 & 0 \\ 0 & 2 & 0 & 1 & 2 & 3 & 0 \\ 0 & 1 & 1 & 3 & 2 & 1 & 0 \\ 0 & 0 & 1 & 1 & 3 & 3 & 0 \\ 0 & 2 & 2 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 1 \\ 2 & 1 & 0 \\ 2 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 9 & 14 & 14 & 16 \\ 10 & 14 & 15 & 18 & 13 \\ 4 & 7 & 15 & 24 & 20 \\ 8 & 14 & 14 & 12 & 14 \\ 3 & 8 & 9 & 8 & 5 \end{bmatrix}$$

Figure 7: Example of padding of the filter to retain the original input size.



7	9	14	14	16
10	14	15	18	13
4	7	15	24	20
8	14	14	12	14
3	8	9	8	5

=

14	15	18	18
14	15	24	20
14	15	24	24
14	14	14	14

Figure 8: Example of max pooling filter by using 2 × 2 kernel.

padding need an activation function, which is relu, a multi-layer neural network, which is represented as Equation (3) (DeepAI 2020).

$$\mathrm{f}(\mathrm{x}) = \max(0, \mathrm{x}) \tag{3}$$

According to the equation, ReLU is a non-linear activation function that represents a sigmoid where, if the value is negative, then it is referred to as zero and, if the value is positive, then the value remains the same. The following, Equation (4), can be written

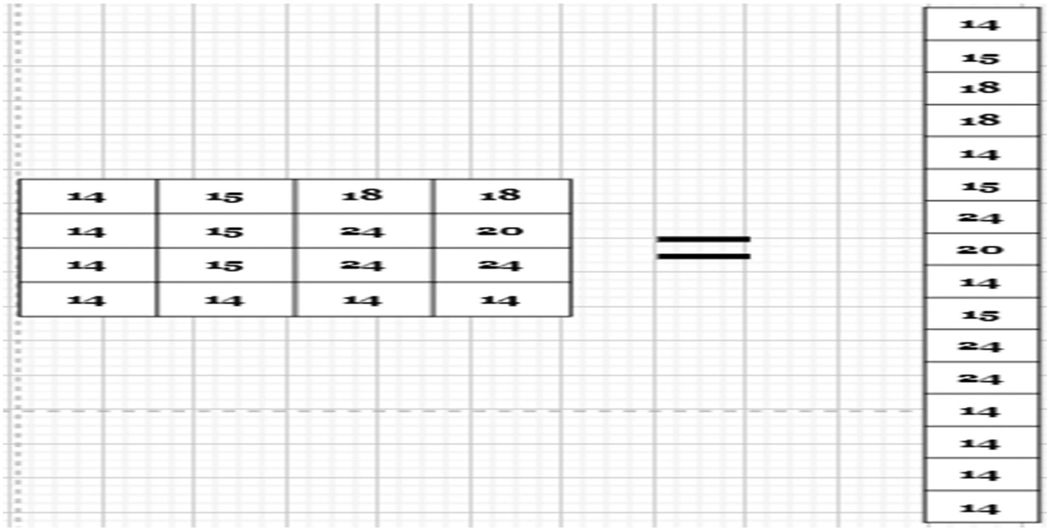
$$\begin{gathered}\mathrm{f}(\mathrm{x}) = 0,\ \text{if}\ \ \mathrm{x} \leq 0 \\ \mathrm{X},\ \text{if}\ \ \mathrm{x} \geq 0\end{gathered} \tag{4}$$

From Equation (4), if the value is f(–1) is < 0, then it is restricted to the value 0; conversely if the value is f(2), it remains the same as the input as the value of $x > 0$. The Relu activation function is used to accelerate the training speed of the neural networks, which saves additional computation time than traditional activations. Pooling layers are used to reduce the size of feature maps, which helps in compressing the dimensions of the features in the image. The subsampling takes the minimum, maximum, or average of the image array cells to proceed further with the output, as shown in Figure 8.

In Figure 8, the max pooling operation with a 2 × 2 window is illustrated. For every 2 × 2 region in the feature map, the maximum value is extracted. Flattening is the step in which the important parts of the images are captured before applying the final filter. The final step is to create a vector where the classification can work on that part of the values provided, which converts a multidimensional array n × n matrix to n × 1 matrix.

In Figure 9, it is seen that the layer that is processed after flattening is the dense layer, which uses softmax as the activation layer. It is well known as the extension for logistic regression where the predicted output is the probability between 0 and 1, it is determined by Equation (5) (Saluja 2022).

$$\mathrm{p}(y = \mathrm{j}|x^{(i)}) = e^{x^{(i)}} \Big/ \sum\nolimits_{j=0}^{k} \quad e^{x_k^{(i)}} \tag{5}$$



14	15	18	18
14	15	24	20
14	15	24	24
14	14	14	14

=

14
15
18
18
14
15
24
20
14
15
24
24
14
14
14
14

Figure 9: Example of flatten filter n × 1 kernel.

2.2.4 Cost function

In this prediction model after softmax, the loss of the function must be calculated, which is cross-entropy or is known as a log loss function, which calculates the distance between the probabilities created by the softmax layer. The use of the loss function is to determine if the prediction has predicted the correct observation. In this study, sparse cross-categorical entropy is used, which is like cross-entropy to determine the labels that belong to two or more classes in which the labels are transformed into integers by using one-hot representation. It is represented as in Equation (6) (Anis 2021).

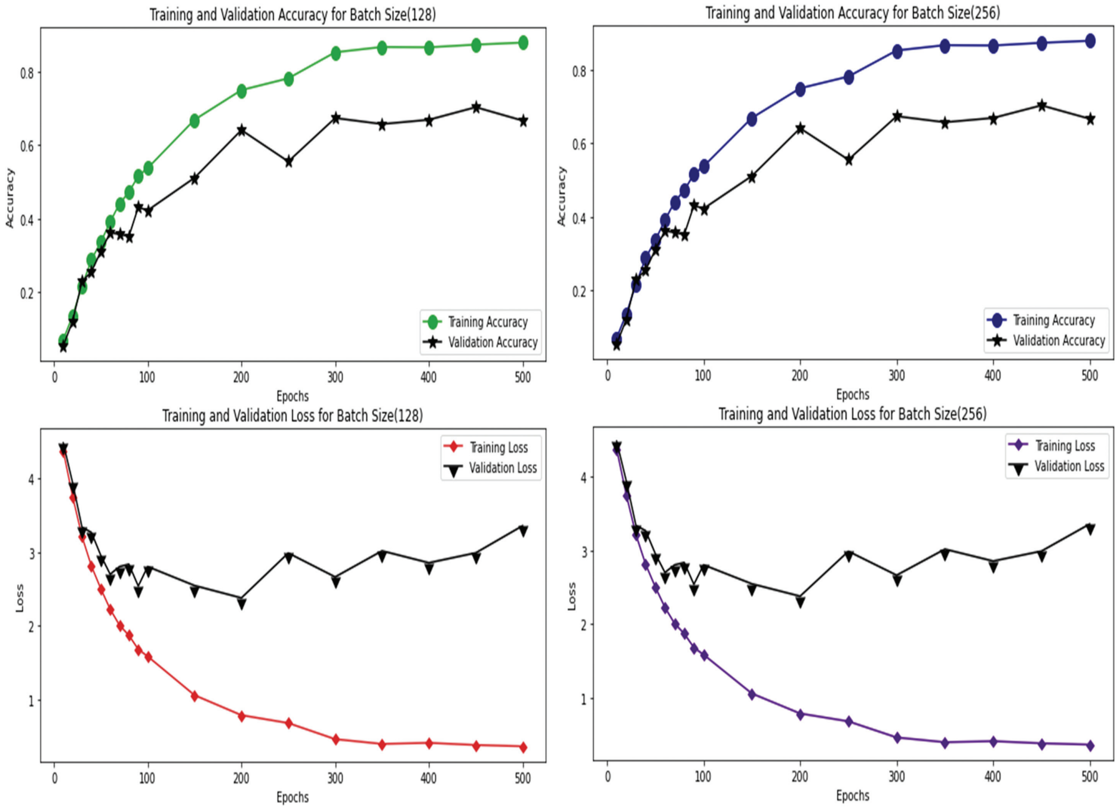
3. Results

In this section, the research results will be presented in detail in different aspects, starting from finding the right batch size to follow up accuracy and loss trends. The next step will be checking the training and validation accuracy comparison to verify the selected batch size. Eventually the selected results of the image classification will be introduced with further analysis of other random image data, which will be explained in greater detail.

3.1 Accuracy and Loss Trend

When the model is trained with several layers of accuracy, a loss check is required to notice which batch size can be used for skin cancer recognition. The trend for accuracy between 0 and 1 while the trend for loss is from 0 to 100 percent is represented in Figure 10. Both accuracy and loss metrics are tracked over a span of 0 to 500 epochs for training and validation sets. This tracking helps determine the presence of any errors or issues with the model's learning process. It helps for understanding model solidity.

The validation loss is a metric, and it is calculated after training each epoch. Validation loss is encountered when under- and overfitting the dataset. If the training loss is close to zero and the validation loss is greater than the



(a) 128 batch size

(b) 256 batch size

Figure 10: Accuracy and loss trend for different batch sizes (a) 128 batch size, (b) 256 batch size.

training loss, then it is considered as overfitting as the predicted model. The model then starts to accidentally predict some unseen data as true. If the training loss is greater than the validation loss, then that means that the model is inaccurate, which causes large errors.

A batch size for thirty two understanding accuracy and loss on the training and validation are demonstrated in Figure 10(a). The accuracy is greater than seventy percent, and the validation accuracy is nearly sixty percent. As per the loss, the training line is less than 1.0 percent and validation losses are too high (approximately 2.9 percent). Although the accuracy is good, the loss represents high impact for the model.

In the research, a batch size of sixty four is used. The graph shown in Figure 10(b) indicates the distance between the training and validation accuracy. When considering the loss function after training the model, it shows that the sixty-four–batch size was better in terms of the distance between the training and the validation loss. It indicates that the model is good for predicting the training accuracy of 86.34 percent, validation accuracy of 64.22 percent. the validation loss of 2.4 percent, and training loss of 0.4 percent.

$$j(W) = -1/n \sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \tag{6}$$

3.2 Training and Validation Accuracy Comparison for Different Batch Sizes

After considering the accuracy results and losses on the four different batch sizes, it is important to compare them to finalize the best batch size and move on with the model application. The training and validation accuracy percentage comparison is explained in Figure 11, and it shows the accuracy for all the batch sizes used to determine which model to use for this research.

The thirty-two–batch size shows a training accuracy of seventy-nine percent and sixty-five percent on validation, in which the images taken for the training in this batch were 206 items. For the 64-batch size, the images taken for training were 103 items, which is a reasonable amount and shows an accuracy of 86.34 percent and validation accuracy of 64 percent. Checking batch size of 128, the images used for training were fifty-two items, which makes the model skip some patterns in skin cancer detection, even though the accuracy is above ninety percent and validation is sixty-seven percent. The loss function is too high for the detection of skin cancer. As for a batch size of 256, the images used for training are thirty-two items, which has fewer chances for the model to understand the labels for the skin cancer and the pattern of the images. When considering all the evidence retrieved from the models of batch sizes in this research, a batch size of sixty four is optimal.

3.3 Image Prediction

Once the batch size has been settled, the next step is testing the model with the image dataset created for the research. The image predictions performed by the model for the validation data, which are twenty percent of the image dataset and eighty percent for training dataset when uploading unseen data for the final prediction, are shown in Figure 12. The objective was to see if the actual label of the skin cancer and the predicted skin cancer label were

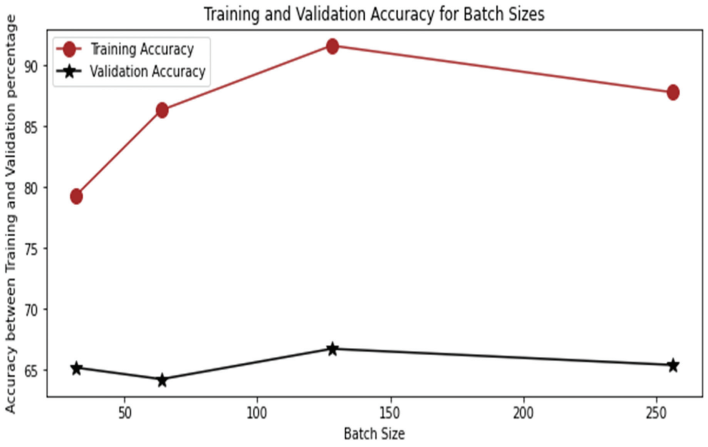


Figure 11: Training and validation accuracy percentage check for batch sizes 32, 64, 128, and 256.

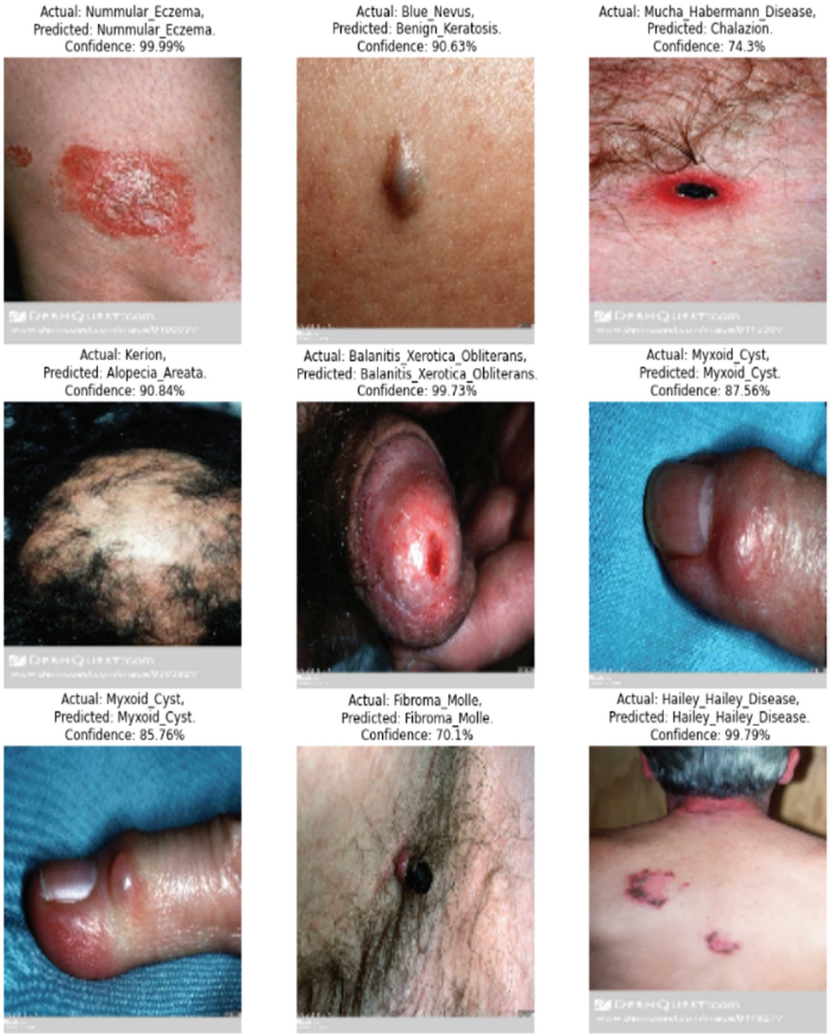


Figure 12: Prediction results for the dataset.

the same. For example, Hailey disease shows the actual label and the predicted label are the same, with the confidence of 99.79 percent. Also, the kerion label shows the predicted label as alopecia. Although the image looks like alopecia, the labels did not match due to overfitting. Also, the Blue Nevus disease class label was predicted incorrectly, with a confidence of 90.63 percent, which helps to conclude that the model is encountering overfitting issues for the same diseases that have very similar skin patterns. To overcome this problem, a data augmentation library can be used in the model to get the required accuracy with the image class labels. The lowest confidence level observed was 70.1 percent for fibroma molle disease, which still shows a strong pattern for medical image classification.

It was decided to test the model image dataset library that has been built for this research. Random disease images could give the indication whether the model was responding to only the dataset library or whether it was coming out with the results also for unintroduced images. Several images were gathered randomly from the internet by using the Google search engine to check if the model can predict the unseen data as the class labels are added to the twenty cancer images shown six of them in Figure 13. The expectation was the model will recognize the class names for some diseases, such as actinic solar damage yet predict acute eczema. However, the prediction results show that the highest confidence level of prediction is 99.22 percent and the lowest prediction confidence is 76.19 percent, with an outlier of a confidence level of 47.32 percent. However, the model needs to be tested by adding more data to understand if the labels are predicted correctly and if there needs to be any other additional layers implemented in the future.

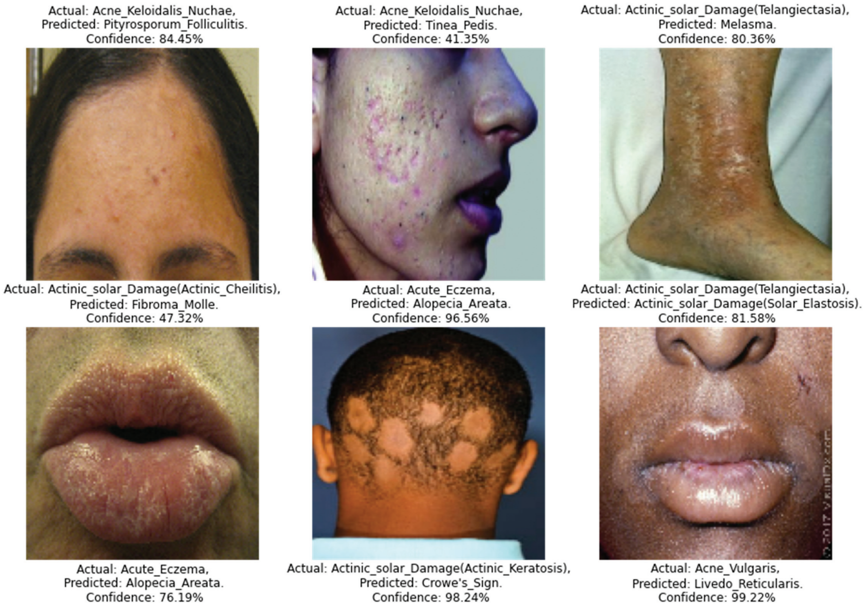

Actual: Acne_Keloidalis_Nuchae,
Predicted: Pityrosporum_Folliculitis.
Confidence: 84.45%
Actual: Acne_Keloidalis_Nuchae,
Predicted: Tinea_Pedis.
Confidence: 41.35%
Actual: Actinic_solar_Damage(Telangiectasia),
Predicted: Melasma.
Confidence: 80.36%
Actual: Actinic_solar_Damage(Actinic_Cheilitis),
Predicted: Fibroma_Molle.
Confidence: 47.32%
Actual: Acute_Eczema,
Predicted: Alopecia_Areata.
Confidence: 96.56%
Actual: Actinic_solar_Damage(Telangiectasia),
Predicted: Actinic_solar_Damage(Solar_Elastosis).
Confidence: 81.58%
Actual: Acute_Eczema,
Predicted: Alopecia_Areata.
Confidence: 76.19%
Actual: Actinic_solar_Damage(Actinic_Keratosis),
Predicted: Crowe's_Sign.
Confidence: 98.24%
Actual: Acne_Vulgaris,
Predicted: Livedo_Reticularis.
Confidence: 99.22%

Figure 13. Prediction results for sample images.

4. Conclusion

According to the training and validation accuracy of the model, it was found that the model built for this research performed better predictions with the batch size of 64 compared with batches with 32, 128, and 256 sizes. Although the validation loss is favorable compared to other batch sizes, there are still other models worth testing, including ReLU, VGG-16, and Unet for object detection. When considering that the CNN model has set an important baseline for image classification, masking skin cancers would assist in the prediction with less loss and high accuracy.

Analysis of the research has shown that, with an image library built by the dataset, the results were satisfying in terms of the confidence level starting from 70.1 percent. Despite the same image classifications that may be misread, the confidence levels are within acceptance levels not only for the dataset but also for the random images. The second check point of the model was to test it with some random images retrieved from the Google search engine by simply searching for images under the “common skin diseases images” category. The confidence level for this study began at 76.19 percent, a commendably acceptable level. This suggests that the model is accurate and can effectively handle datasets it has not been previously trained on. The dataset used for the research contains 6,584 images of 198 fine-grained skin disease categories. Medical image banks are hard to compile, so it is a challenging task to label them.

This study also raises a challenging problem for automatic visual classification of clinical skin disease images. An obvious fact is that traditional methods are not sufficient in terms of effectiveness. The challenging work was to build a clinical skin disease images dataset, including samples of real-world images from twenty categories. Each sample in the benchmark is well labeled. The research intended to release the labeled dataset to the community to promote related research and facilitate its expansion. Hence, to increase efficiency of the model, more data are required to train it. Increasing the data in the CNN model would increase the accuracy of the model and minimize the loss function in this research model that is made for the image recognition of skin diseases.

This research helped establish some proven record analysis based on clinical disease images. However, all these works are built on smaller datasets, which only contain very few species and are not publicly available. The absence of benchmark datasets is a barrier to compare the outcomes of the completed research. Consequently, in this research, one of the important milestones would be considered as the introduction of a new publicly available dataset for real-world skin disease image recognition. It would also be a strong baseline for different skin colors, which was the starting point of this research. The challenge is to improve image recognition for the dark colors between brown and black because it makes it more complicated to get the correct disease identification.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2016. "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." Preprint, submitted March 16. https://doi.org/10.48550/arXiv.1603.04467

Anis, A. 2021. "How to Build Custom Loss Functions in Keras for Any Use Case." Cnrg.io blob. Accessed February 11, 2022. https://cnvrg.io/keras-custom-loss-functions/

Barata, C., M. E. Celebi, and J. S. Marques. 2021. "Explainable Skin Lesion Diagnosis Using Taxonomies." *Pattern Recognition* **110**: 107413. doi: 10.1016/j.patcog.2020.107413

Bui, P. 2021. "Using AI to Help Find Answers To Common Skin Conditions." Google AI. Accessed February 4, 2022. https://blog.google/technology/health/ai-dermatology-preview-io-2021/

Chang, D. 2021. "Effect of Batch Size On Neural Net Training—Deep Learning Experiments." Accessed February 9, 2022. https://medium.com/deep-learning-experiments/effect-of-batch-size-on-neural-net-training-c5ae8516e57

Chen, X., D. Li, Y. Zhang, and M. Jian. 2021. "Interactive Attention Sampling Network for Clinical Skin Disease Image Classification." *Proceedings of the Pattern Recognition and Computer Vision, 4th Chinese Conference, PRCV 2021*, Beijing, China, October 22.

Das, K., C. J. Cockerell, A. Patil, P. Pietkiewicz, M. Giulini, S. Grabbe, and M. Goldust. 2021. "Machine Learning and Its Application in Skin Cancer." *International Journal of Environmental Research and Public Health* **18**, no. 24: 13409. doi: 10.3390/ijerph182413409

DeepAI. 2020. "ReLu." Accessed February 11, 2022. https://deepai.org/machine-learning-glossary-and-terms/relu

Enderling, H. 2019. "Towards a Quantitative Personalised Oncology." *Research Outreach* **107**, no. 107: 130–133. doi: 10.32907/ro-107-130133

Errichetti, E. 2020. "Dermoscopy in Monitoring and Predicting Therapeutic Response in General Dermatology (Non-Tumoral Dermatoses): An Up-to-Date Overview." *Dermatology and Therapy* **10**, no. 6: 1199–1214. doi: 10.1007/s13555-020-00455-y

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* **542**, no. 7639: 115–118. doi: 10.1038/nature21056

Hay, R. J., N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J. L. Murray, and M. Naghavi. 2014. "The Global Burden of Skin Disease in 2010: An Analysis of the Prevalence and Impact of Skin Conditions." *Journal of Investigative Dermatology* **134**, no. 6: 1527–1534. doi: 10.1038/jid.2013.446

Hubiche, T., L. Valério, F. Boralevi, E. Mahe, C. B. Skandalis, A. Phan, P. del Giudice. 2016. "Visualization of Patients' Skin Lesions on Their Smartphones." *JAMA Dermatology* **152**, no. 1: 95. https://jamanetwork.com/journals/jamadermatology/fullarticle/2453323. doi: 10.1001/jamadermatol.2015.2977

Hurt, M. A. 2012. "Weedon D. Weedon's Skin Pathology." *Dermatology Practical & Conceptual* **2**, no. 1 3rd ed. doi: 10.5826/dpc.0201a15

Introduction to Convolutional Neural Networks. 2022. "IBM Developer." Accessed February 11, 2022. https://developer.ibm.com/articles/introduction-to-convolutional-neural-networks/

Kiran, T. T. J. 2020. "Computer Vision Accuracy Analysis with Deep Learning Model Using Tensorflow." *International Journal of Innovative Research in Computer Science & Technology* **8**, no. 4: 319–325. doi: 10.21276/ijircst.2020.8.4.13

Li, H., Y. Pan, J. Zhao, and L. Zhang. 2021. "Skin Disease Diagnosis with Deep Learning: A Review." *Neurocomputing* **464**: 364–393. ACM Digital Library. doi: 10.1016/j.neucom.2021.08.096

Liu, J., F. Chao, C. M. Lin, C. Zhou, and C. Shang. 2021. "DK-CNNs: Dynamic Kernel Convolutional Neural Networks." *Neurocomputing* **422**: 95–108. doi: 10.1016/j.neucom.2020.09.005

Marchetti, M. A., N. C. F. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, N. Jaimes, A. A. Marghoob, E. Quigley, A. Scope, O. Yélamos, and A. C. Halpern. 2018. "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging Challenge." *Journal of the American Academy of Dermatology* **78**, no. 2: 270–277.e1. doi: 10.1016/j.jaad.2017.08.016

McDermott, J. 2022. "Convolutional Neural Networks: Image Classification with Keras–A Comprehensive Guide." Accessed February 11. https://www.learndatasci. com/tutorials/convolutional-neural-networks-image-classification/

Powell, K. 2019. "Searching by Grant Number: Comparison of Funding Acknowledgments in NIH RePORTER, PubMed, and Web of Science." *Journal of the Medical Library Association* **107**, no. 2: 172–178.

Saluja, A. 2022. “Softmax Function and Layers Using Tensorflow.” OpenGenus IQ: Computing Expertise & Legacy. Accessed February 11, 2022. https://iq.opengenus.org/softmax-tf/

Shinozaki, T. 2021. “Biologically Motivated Learning Method for Deep Neural Networks Using Hierarchical Competitive Learning.” *Neural Networks* **144**: 271–278. doi: 10.1016/j.neunet.2021.08.027

Skin Disease Images. 2023. “The Image Dataset With the Python Code Is Available in the GitHub Repository.” Accessed October 15, 2023. https://github.com/sharathkumarphd/Skin-diseases.git

Skin Lesions. 2022. “Newofmc.” Accessed February 22, 2022. https://www.ocalafamilymedicalcenter.com/skin-lesions

Takimoglu, A. 2021. “What is Data Augmentation? Techniques, Benefit & Examples.” AI Multiple. Accessed February 9, 2022. https://research.aimultiple.com/data-augmentation/

What is Deep Learning? 2022. “SAS.” Accessed February 4, 2022. https://www.sas.com/en_us/insights/analytics/deep-learning.html

Yang, K., Z. Sun, A. Wang, R. Liu, Q. Sun, and Y. Wang. 2018. “Deep Hashing Network for Material Defect Image Classification.” *IET Computer Vision* **12**, no. 8: 1112–1120. doi: 10.1049/iet-cvi.2018.5286

Yang, J., X. Wu, J. Liang, X. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang. 2020. “Self-Paced Balance Learning for Clinical Skin Disease Recognition.” *IEEE Transactions on Neural Networks and Learning Systems* **31**, no. 8: 2832–2846. doi: 10.1109/TNNLS.2019.2917524

Yu, L., H. Chen, Q. Dou, J. Qin, and P. A. Heng. 2017. “Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks.” *IEEE Transactions on Medical Imaging* **36**, no. 4: 994–1004. doi: 10.1109/TMI.2016.2642839

Journal of Big Data and Artificial Intelligence

WWW.JBDAI.ORG

ISSN: 2692-7977

JBDAI Vol. 2, No. 1, 2024 DOI: 10.54116/jbdai.v2i1.30

Are Emotions Conveyed Across Machine Translations? Establishing an Analytical Process for the Effectiveness of Multilingual Sentiment Analysis with Italian Text

Richard Anderson
Rutgers University
rick.anderson@rutgers.edu

Carmela Scala
Rutgers University
carmela.scala@rutgers.edu

Jim Samuel
Rutgers University
jim.samuel@rutgers.edu

Vivek Kumar
University of Cagliari
vivek.kumar@unica.it

Parth Jain
Rutgers University
pj269@rutgers.edu

Abstract

Natural language processing (NLP) is being widely used globally for a variety of value-creation tasks ranging from chat-bots and machine translations to sentiment and topic analysis and multilingual large language models (LLMs). However, most of the advances are initially implemented within the English language framework, and it takes time and resources to develop comparable resources in other languages. The advances in machine translations have enabled the rapid and effective conversion of content in global languages into English and vice versa. This creates potential opportunities to apply English language NLP methods and tools to other languages via machine translations. However, although this idea is powerful, it needs to be validated and processes and best practices need to be developed and kept updated. The present research is an effort to contribute to the development of best practices and an evaluation framework. We present a systematic and repeatable state-of-the-art process to evaluate the viability of applying English language sentiment analysis tools to Italian text by using multiple English language machine translation mechanisms such that it can be easily extended to other languages.

Keywords *natural language processing, natural language understanding, sentiment analysis, machine translation, italian, emotion.*

1. Introduction

Natural language processing (NLP) as a domain has been experiencing unprecedented breakthroughs and an exponential adoption growth rate by businesses, institutions, governments, and individual users, driven by an increasing interest in textual data and the analytical and generative potential it presents (Samuel et al. 2022b). The global NLP market is expected to grow to $49.4 billion (United States dollars) by 2027, and there have been many notable developments in NLP since late 2021 (Markets and Markets 2022): Apple will provide an open-source reference PyTorch implementation of the Transformer architecture for its products, enabling global developers to effortlessly run Transformer models. At the end of 2021, Baidu introduced PCL-BAIDU Wenxin (ERNIE 3.0 Titan), a state-of-the-art knowledge-enhanced 260 billion parameters-based large language model (LLM) for the Chinese language. This model outperformed its predecessors easily, and more recently, in March of 2023, the controlled launch of OpenAI's multimodal (accepts images and text as input) Generative Pre-trained Transformer 4 (GPT-4), speculated to have around two trillion parameters, via ChatGPT Plus has further demonstrated the power and expanding capabilities of LLMs (Liu et al. 2023).

Large language models have been developed with a primary focus on English, and a few other LLMs such as ERNIE 3.0 Titan in the Chinese language have also been developed (Wang et al. 2021; Nguyen et al. 2023). Google's 2021 MUM language model was trained across 75 languages and is an exception to mainly English-focused language models. Google's VP of Search declared that MUM as "1,000 times more powerful than BERT" and that it has "…the potential to transform how Google helps [users] with complex tasks" (Nayak 2021). From a resource availability and allocation perspective, it would be expensive and probably unfeasible to expect such models to be built and kept updated for every human language in the short term. It is clear that LLM performance "among under-represented languages fall behind due to pre-training data imbalance" (Nguyen et al. 2023).

It is even more challenging for a large array of NLP tools, models, and methods available in Python, R and other languages to be readily extended to alternative and vernacular languages with the same level of effectiveness (Ranathunga and de Silva, 2022). While there have been recent localized efforts to develop NLP tools in other languages such as Welsh, Marathi, and Malayalam, it is evident that much work remains to be done (Cunliffe et al. 2022; Lahoti et al. 2023; Sebastian 2023). Given the growing importance of textual data analytics and NLP applications in a wide array of research, policy, socioeconomic, healthcare, business, and other domains, and in addressing global events such as the COVID-19 pandemic, it is important to address the challenge of multilingual data (Samuel et al. 2020a, 2020b; Rahman et al. 2021; Ali et al. 2021). This could be done using multiple approaches including the grassroots level development of local language NLP tools which would be time consuming and lag well behind English language tools, and also through the use of machine translations which could create opportunities for timely applications.

The key question therefore is: Given the NLP advances in one language such as English, can we extend the applications and benefits to other languages by machine-translating such languages into the language with advanced NLP models and tools and then draw implications back to the original languages effectively? Past research has shown that such an approach is feasible, and it is possible to use machine translations in conjunction with other NLP tools, including sentiment analysis with increasing effectiveness (Balahur and Turchi, 2012, 2014). However, there is a need to articulate a clear, updated, and repeatable process for applying NLP tools from one language to another with an evaluation mechanism to compare and gauge the effectiveness of such a process. To address this, we ran an experiment with a lab-developed Italian text corpus, using multiple machine translations and multiple sentiment analysis tools. In the next section, we conduct a literature review of relevant state-of-the-art NLP methods and tools, followed by a description of our dataset, process, evaluation methods, and analysis. We conclude with a discussion of our process and analysis, notes on limitations, future research, and concluding thoughts.

2. Literature Review

Extant research has emphasized the paucity of NLP tools for many languages, and past studies have experimented with the use of machine translations-based sentiment analysis for languages such as Arabic, researchers affirmed the usefulness of machine translations in spite of the lack of high levels of accuracy (Mohammad et al. 2016; Oueslati et al. 2020). Steering away from translating the text corpus, past multilingual sentiment analysis research has also obtained fair results using "automated translation of the dictionary" for legislative bills (Proksch et al. 2019). A recent study using French, Spanish, and Japanese machine translations analyzed the impact of indirect (pivot, using a mediating language) machine translations on automated sentiment analysis and highlighted weaknesses of sentiment classifiers when working with translated texts while also affirming the usefulness of machine translations-based analysis (Poncelas et al. 2020). Going further, recent research has also posited that with certain languages, machine translations based on sentiment analysis using English language tools yielded better results than the language-specific tools used for sentiment analysis (Araújo et al. 2020).

More recently, Kumar et al. (2023) used “a zero-shot learning-based cross-lingual sentiment analysis (CLSA)” to demonstrate the viability of using machine translations-based sentiment analysis for the Sanskrit language. So also machine translation has been shown to work well with classifier performance for the Bengali language (Sazzed and Jayarathna, 2019; Sazzed 2020). Berard et al. (2019) applied sentiment analysis and focused on the benefits of improving the quality of machine translation using French language user-generated content. This is useful because extant research has highlighted numerous challenges with machine translation-based approaches including “sparseness and noise in the data” and the failure of translation mechanisms to “translate essential parts of a text, which can cause serious problems, possibly reducing well-formed sentences to fragments” (Dashtipour et al. 2016). A number of NLP-based studies in the Italian language have used sentiment analysis, such as performing sentiment analysis on Italian Twitter data, the use of cross-lingual transfer learning for analyzing the sentiment of Italian TripAdvisor review data, application of sentiment analysis and text mining for generating insights from YouTube Italian videos on vaccination and a comparison of lexicon-based and Bert-based methods (Basile and Nissim, 2013; Porreca et al. 2020; Catelli et al. 2022a, 2022b). Similarly, there has been a fair amount of research on NLP and machine translations of the Italian language, including basic translation automation effort and more advanced applied research (Russo et al. 2012; Wiesmann 2019; Bawden et al. 2020; Modzelewski et al. 2023). However, in spite of numerous multilingual studies in Italian, we did not find any comparable combination of NLP tools and machine translations-based studies for the Italian language.

3. Data and Method

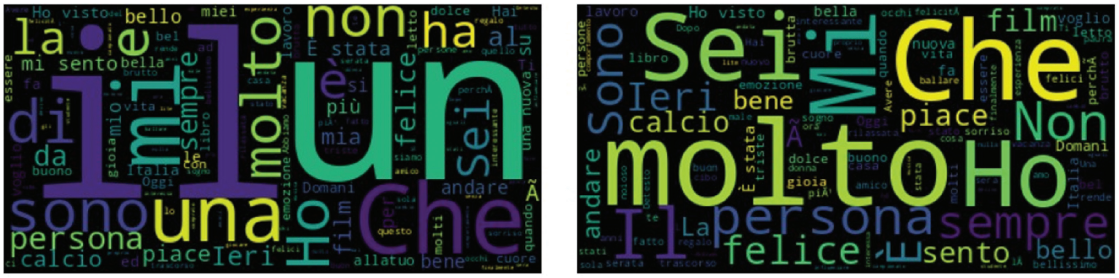
In this section, we describe the development of the Italian dataset and the ’gold standard’ human expert-assigned sentiment classifications and visualize a few key features as shown in Figures 1a and 1b. We then explain the machine translation process and report on the two translation models we applied (Figures 2a and 2b). We present our analysis of the accuracy and nuances of the machine translations from Italian to English, then report our findings from applying sentiment analysis to the English translations. We compare the sentiment assigned to the English translations to the original Italian language gold standard sentiment classes and present our findings.

3.1 Data

The unique dataset of sentences from Author 2 were human-generated Italian Sentences and not from public sources. These sentences were created to have a clear positive, neutral, or negative sentiment. This information was recorded in the dataset. For this experiment, we used two translation methods, one was the web tool for Google Translate (Han 2022). We used that as a common and popular source of translations. To get the same results as the web tool for Google Translate method, we used the googletrans Python library. This library provides a convenient interface to Google Translate, allowing for consistent translation operations within Python scripts. Next, we chose the Marian Machine Translation Transformer-based technique for translations (Junczys-Dowmunt et al. 2018). It had a documented method of translation that could be reproduced.

We used the Marian Neural Machine Translation (Marian MT) and model to translate Italian to English using the Marian MT method:

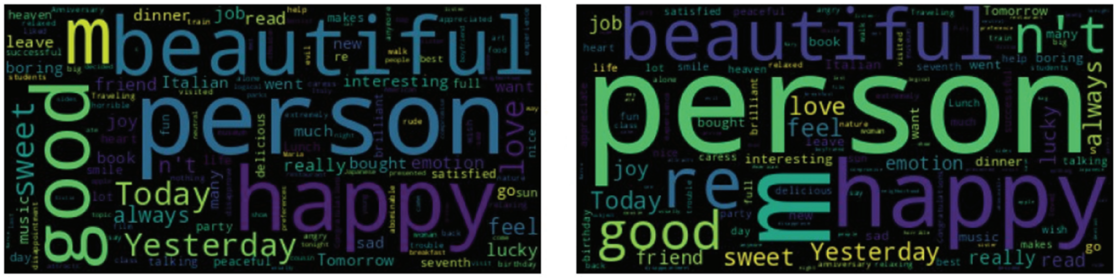
opus-mt-tc-big-it-en Neural machine translation model for translating from Italian (it) to English (en). This model is part of the **OPUS-MT project**, an effort to make neural machine translation models widely available and accessible for many world languages (Tiedemann 2020; Tiedemann and Thottingal, 2020). All models are originally trained using the framework of **Marian NMT**, an efficient MT implementation written in pure C++ (Junczys-Dowmunt et al. 2018). The models have been converted to pyTorch using the transformers library by Huggingface. Training data is taken from **OPUS**, and training pipelines use the procedures of **OPUS-MT-train**.



(a) Complete Italian no stop words.

(b) Complete Italian stop words removed.

Figure 1: Word clouds for complete Italian text with and without stop words.



(a) Word Cloud for Google Translate Text where both trans- both translations are true.

(b) Word Cloud for Marian MT Italian to English Text where lations are true.

Figure 2: Word clouds for Google Translate text and Marian MT Italian to English text where both translations are true.

Then we ran VADER sentiment evaluation on each of the translated sentences. Starting from a total of 167 sentences. Sixty-five sentences are correct in translation. Thirty-nine percent of the sentences were accurate from both datasets. The official source data frame will include the data where sentences are accurate in both translations. These "both true" sentences will be used in the remaining analysis. That way, we are measuring the good translations from here on out.

Google Translate had 91 good translations, while Opus Translate had 119. This indicates that Opus Translate performed slightly better in terms of translation quality in this specific dataset. There were 65 instances where both Google Translate and Opus Translate had good translations. This suggests some overlap in the quality of translations between the two engines.

BLEU and chrF scores are commonly used to measure the quality of a corpus and how well it adheres to accepted translations for the corpus. We used both techniques on our dataset to determine whether either would give us an automated method of evaluating the translated sentences. Our dataset includes the true sentiment and the "correct" translation. Either method might have biases based on their method of calculation. When we compared, the data BLEU and chrF scores on our dataset matched human the approved gold standard sentences. The BLEU and chrF metrics vary for how far off a translation is from what is expected, but both agree that the same sentences the expert has said are good translations. Either metric is good at confirming the human-chosen good translations for our dataset.

3.2 Machine Translations EVALUATION (- RICK to DESCRIBE MT and EVAL METRICS)

The process of analysis was automated in the following Colab Notebook: https://github.com/rianders/mtnlpxlmsentiment/blob/main/SentimentAnalysisAll.ipynb.

We used this dataset:

https://github.com/rianders/mtnlpxlmsentiment/blob/main/data/SentinmentALL-20230508.csv.

Evaluation tools: Marian MT OPUS Italian Dataset Google Translate VADER BLEU chrF

The notebook fetches the source data created by Dr. Scala and Dr. Samuel. These data include the true and gold standard sentiment, source sentence, and official translation sentence information. The next step cleans the data, runs the BLEU and chrF comparisons, and adds that information to the dataset. Then the translation and machine translation quality checks and graphs are created to confirm quality and accuracy. These quality checks and graphs include word frequency, sentence length, BLUE, and chrF scores. Translation comparisons are performed between Google Translated and Marian MT using OPUS Italian Data.

Now that the quality of translation has been determined, a review of sentiment distribution is shown. We use the VADER method where −1 is negative, 0 is neutral, and 1 is positive.

Then a new data frame is created that only contains the "correct" translations. Then word clouds are generated from that data frame. We calculate the confusion matrix, word frequency, and sentence length for each translation method. Then identify and show the outliers.

This process can be repeated with the same or updated dataset. The evaluation process will be the same and can give accurate feedback.

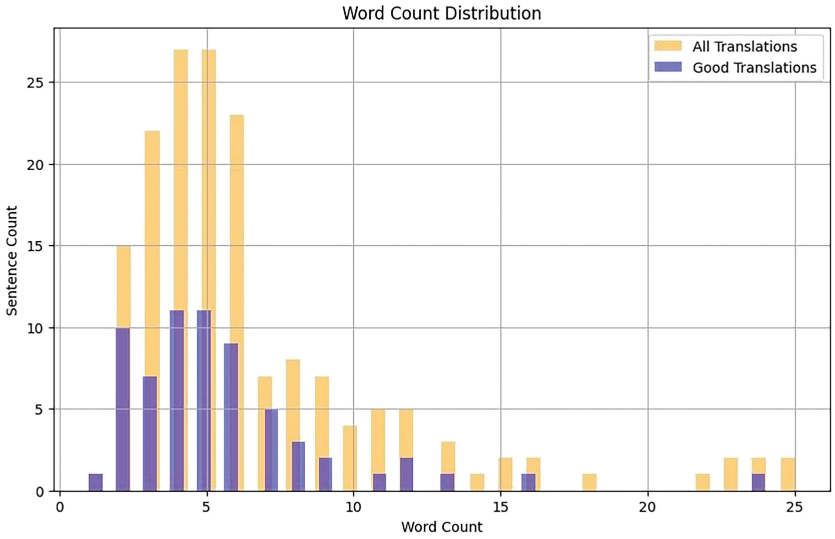


Figure 3: Histogram of the count of sentences by number of words.

3.3 Machine Translations—Observations

These word clouds show us something exciting about languages and language translations. Generally speaking, both translation software performed much better with short sentences and had some problems with longer ones (Figure 3). If we look closely at them, the Italian one does not have many words in common with the various English images, which seem very similar. We have a predominance of nouns, adjectives, and adverbs in the English clouds. In the Italian one, we have primarily articles, conjunctions (che in the specific), prepositions, and verbs. This discrepancy between the Italian word clouds and the English ones is easy to explain whether we consider the typical sentence structure of the Italian language. Italian uses articles, prepositions, and conjunctions (especially che) much more than English, and the word clouds captured this difference perfectly.

Generally speaking, both translation software performed much better with short sentences and had some problems with longer ones. In the sentence below, for example, in the second part, Google assumed the ‘subject’ was “Tom Cruise,” thus translating “mi ha emozionato” with “he excited me” when it should have been “it excited me.” In the original sentence, the subject was the movie, not the actor. Google translations were also more “literal”; hence they did not always produce sensible sentences in English. On the contrary, Opus’s second set of translations was more accurate from an “idiomatic” point of view. In fact, Opus could better identify the idiomatic peculiarities of the sentences. In the sentence below, for example, in the second part, as noted above, Google Translate assumed the “subject” was “Tom Cruise,” thus translating “mi ha emozionato” with “he excited me” when it should have been “it excited me.” In fact, the subject was the movie and not the actor. Opus, instead, provided the perfect translation, identifying the right subject, “it.”:

Italian: *Ho visto il nuovo film di Tom Cruise, “Maverick”, e devo dire che mi ha emozionato percȟe mi ha riportato alla mia gioventù.*

English: *I saw Tom Cruise’s new film, “Maverick,” and I must say that he excited me because he brought me back to my youth.*

Also, it is essential to point out that some translations would have been correct in British English but are considered incorrect or inaccurate in USA English. Here are some examples:

1. Giovanna ha scelto di giocare a calcio: Giovanna chose to play football.
2. Ieri siamo andati allo stadio a vedere una partita di calcio: Yesterday we went to the stadium to watch a football match.

   Football is accurate in British English but in the United States, “calcio” is referred to as soccer.
3. Abbiamo deciso di cambiare casa: We decided to change homes.
4. E stata una vacanza da sogno: It was a dream vacation! OR It was a dream holiday! “We decided to change homes” and “It was a dream holiday” could be accepted in British English, but they sound wrong in United States English.

3.4 Data Analysis

The original dataset containing all translations had a range of sentiments that did not show coherence between translation methods. When we reviewed the sentiment for sentences considered accurate, VADER generated linear agreement across the negative, neutral, and positive categories. This is observed in Figures 4 and 5. Figure 4 shows the range of sentiments that include inaccurate translations. Figure 5 shows the VADER sentiment of accepted translations and that they stay along a linear path across the sentiment values.

When we categorize the remaining sentiment after removing the inaccurate translations, Figure 6 shows that 49.2% of the remaining sentences are positive. The Negative is 23.1%, and the Neutral is 27.7%.

The values for VADER sentiment are shown for Google Translate in Figure 7 and Marion MT in Figure 8. The VADER sentiment was the same for both translation techniques in terms of percentages. The way to tell the difference was to observe the outliers.

Figure 9 shows the VADER sentiment Confusion Matrix for Google Translate and Figure 10 shows the Confusion Matrix for the Marion MT translations. These two graphs show that the misclassifications were similar and that to determine the type of outlier, is to look at those misclassification cases. Those outliers are listed in listings 1 through 3. Figures 11 and 12 show the outliers and that neutral values did not have a clear cluster towards neutral. That one value in the negative was in the positive range.

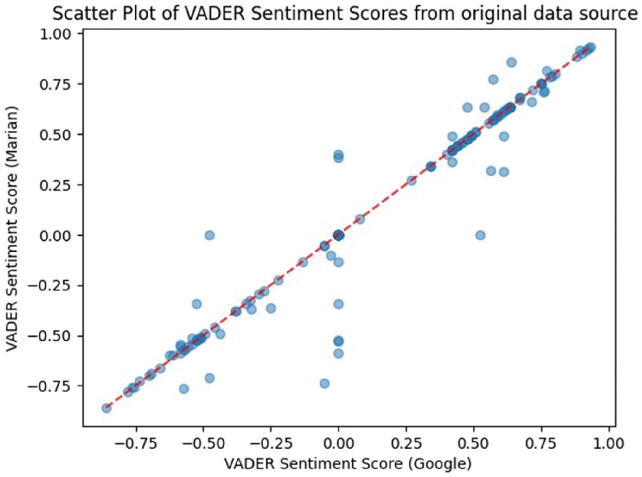


Figure 4: VADER sentiment for Marion MT and Google.

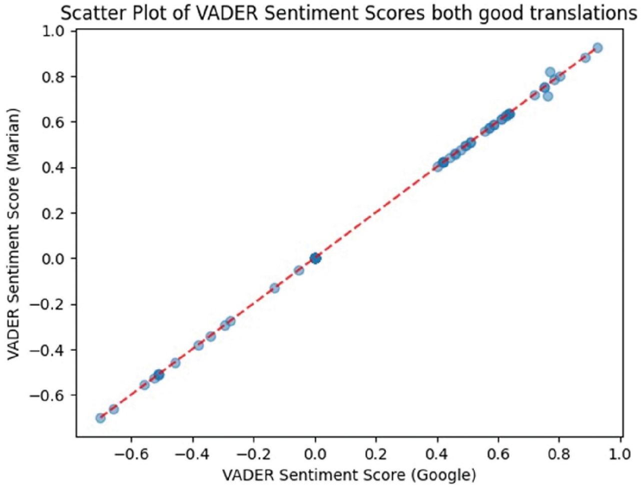


Figure 5: VADER sentiment for Marion MT and Google Translate including original bad translations translate accepted translations.

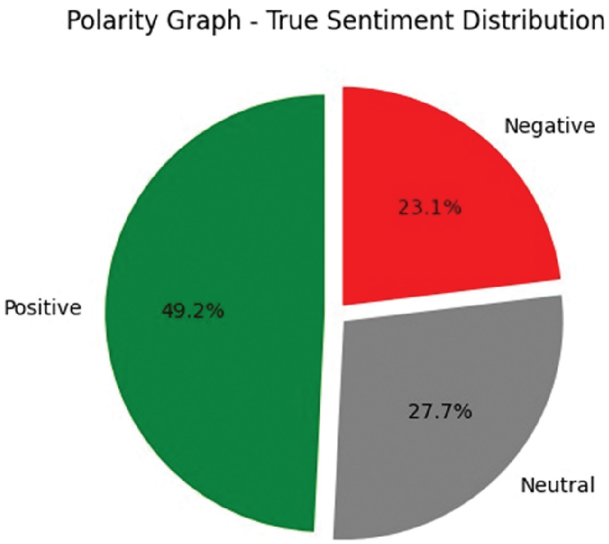


Figure 6: The true sentiment distribution for accurate translations.

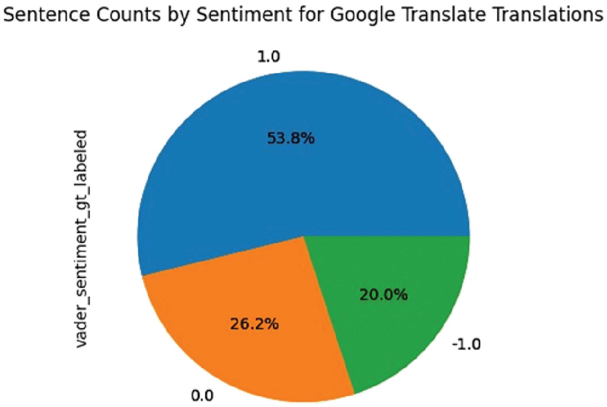


Figure 7: VADER sentiment for Google Translate by percentage number of sentences in category.

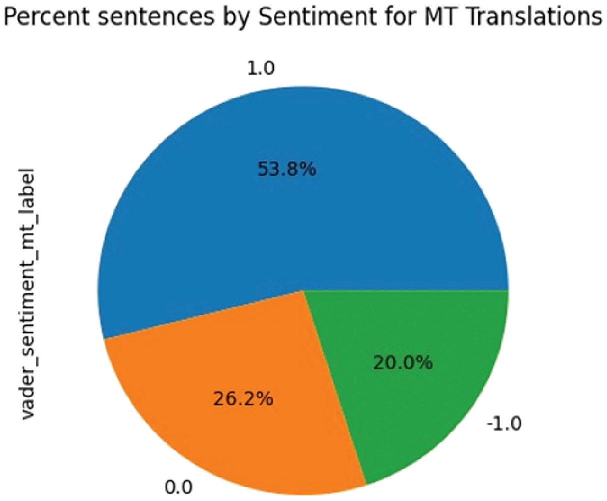


Figure 8: VADER sentiment for Marian Machine Translation by percentage number of sentences in category.

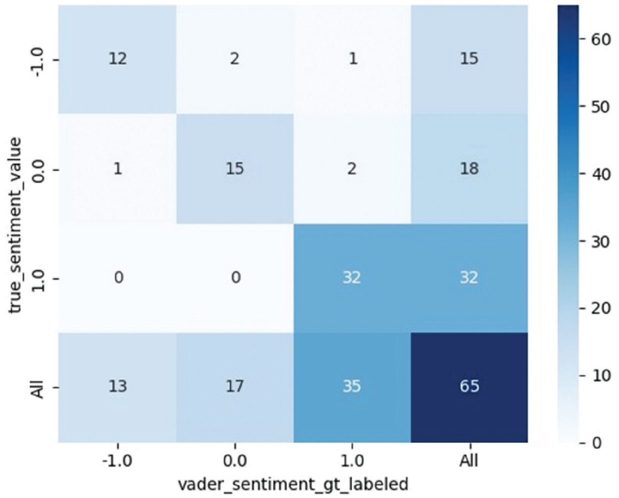


Figure 9: VADER sentiment confusion matrix for Google Translate.

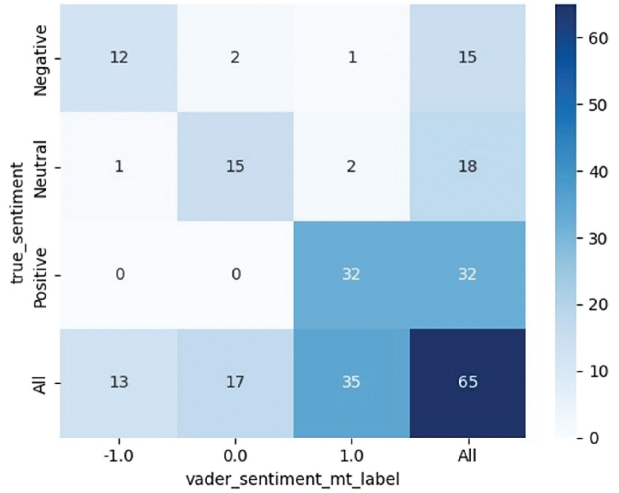


Figure 10: VADER sentiment confusion matrix for Marian Machine Translate.

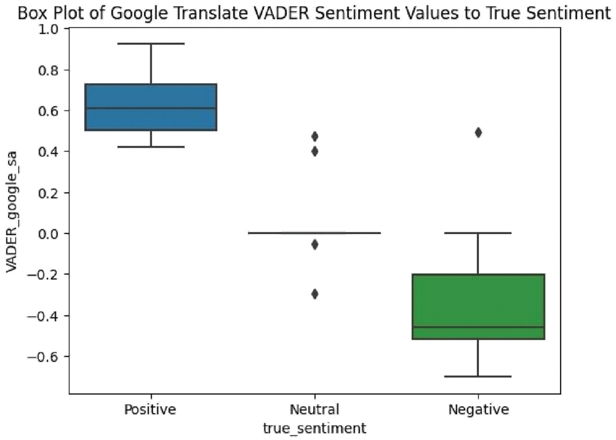


Figure 11: VADER sentiment box plot for Google Translate.

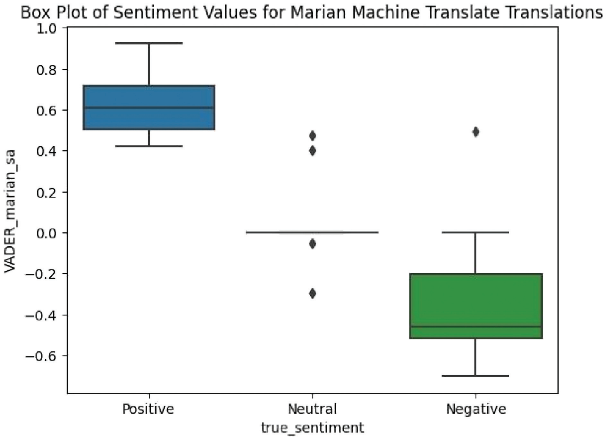


Figure 12: VADER sentiment box plot for Marian Machine Translate.

When examined, the outliers are the same; both calculated VADER values for either Google Translate or Marian MT are the same so all outliers are the same. This would be an area where more overall data would be useful.

4. Discussion

For both Google Translate and Marian MT we compared the true values with the rated values and looked for patterns in the outliers. What were there any after we removed the bad translations. We will review the outliers by Google Translate Positive, Neutral, and Negative. Then do the same for Marian MT. In this subsection, we review outliers for Google Translate output. There were no Google Translate positive outliers, implying no true positive sentiment statements falsely classified as neutral or negative after being translated into English.

The Google Translate sentence had no positive outliers.

Listing 1: Outliers for Neutral

```
Outliers for Neutral:

Sentence: La musica americana attrae sempre molti giovani italiani

Google Translation: American music always attracts many young Italians

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment: 0.4019
```

```
Sentence: I miei amici sono venuti a farmi visita

Google Translation: My friends came to visit me

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment: 0.4767
```

```
Sentence: Non ho preferenze su cosa fare stasera

Google Translation: I have no preferences on what to do tonight

True Sentiment: Neutral

True Sentiment Value: 0.0

VADER Sentiment: —0.296
```

```
Sentence: Domani partiamo per andare in Italia
Google Translation: Tomorrow we leave to go to Italy
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment: −0.0516
```

Outliers Marian Machine Translation

Using Marian MT, we had no positive outliers. Listing 3 and 4 show the neutral and negative outliers.

Listing 2: Marian MT Outliers for Neutral

```
Outliers for Neutral:
Sentence: La musica american attrae sempre molti giovani italiani
Marian MT Translation: American music always attracts many young Italians
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment (MT): 0.4019
```

```
Sentence: I miei amici sono venuti a farmi visita
Marian MT Translation: My friends came to visit me
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment (MT): 0.4767
```

```
Sentence: Non ho preferenze su cosa fare stasera
Marian MT Translation: I have no preference on what to do tonight
True Sentiment: Neutral
True Sentiment Value: 0.0
VADER Sentiment (MT): −0.296
```

Listing 3: Marian MT Outliers for Negative

```
Outliers for Negative:
Sentence: Sei un essere abominevole.
Marian MT Translation: You are an abominable being.
True Sentiment: Negative
True Sentiment Value: −1.0
VADER Sentiment (MT): 0.0
```

```
Sentence: Disapprovo la tua scelta!
Marian MT Translation: I disapprove of your choice!
True Sentiment: Negative
True Sentiment Value: −1.0
VADER Sentiment (MT): 0.0
```

```
Sentence: Buono a nulla!
Marian MT Translation: Good for nothing!
True Sentiment: Negative
True Sentiment Value: −1.0
VADER Sentiment (MT): 0.4926
```

All the sentences in the original Italian have no connotation of positive or negative sentiment; hence, they are considered neutral. Indeed, they do not express an opinion but simply state facts. So, why are they perceived differently in translation? In the case of the second and third sentences, we can assume that the sentiment was perceived as negative due to the presence of some negative signifiers, such as "no" (in "I have no preference…") and the verb "leave" (in "Tomorrow we leave…"). "Leave" implies a separation, and that is why it was probably perceived as negative. As for the first sentence, which was neutral but perceived as positive, we can infer that the presence of verbs indicating "company," such as "came" and "visit," triggered the positive interpretation. As mentioned above, these misinterpretations are also due to the absence of context and "voice inflection."

4.1 Why Was Polarized Italian Made Neutral?

The neutral score assigned to these sentences was a bit of a surprise. Let's analyze the original sentiments. Here are possible explanations for each of them: "Sei un essere abominevole" (translated correctly as "You are an abominable being") clearly has a negative sentiment. Even in the absence of overtly negative signifiers (such as "no," "not," "never," etc.), the word "abominable" sets the tone for a negative interpretation. However, since NLP models are primarily trained to identify the presence of positive and negative words to determine the sentiment of a sentence, in really short sentences where the rest of the signifiers are neither positive nor negative, the NLP model might make a "decision" to assign a neutral sentiment. In this case, that might be why it was perceived as neutral.

As for the second sentence, "Disapprovo la tua scelta" (translated as "I disapprove of your choice"), "disapprove" is the only overtly negative word, making the sentiment clearly negative.

Regarding the last sentence, "Buono a nulla" (translated as "Good for nothing"), it is possible that the NLP model was confused by the equal presence of positive and negative signifiers: "buono" (good) is positive, while "nulla" (nothing) is negative. Consequently, it might have perceived the conflicting sentiments as canceling each other out and opted for a neutral score.

In summary, the different interpretations of the original sentiments in translation could be attributed to the NLP model's reliance on identifying positive and negative words to determine sentiment and the specific words present in each sentence that contribute to the overall sentiment.

Sentiment Analysis General observations:

The sentiment analysis presented some interesting "challenges," more so in dealing with neutral sentences.

Let us look at some examples.

1. "L' esperienza studio in Italia è stata unica. Mi ha cambiato letteralmente la vita e mi ha aperto gli occhi su una nuova realtà." (Original positive) [The study experience in Italy was unique. It literally changed my life and opened my eyes on a new reality.]

   All three engines, Google, Opus, and NLTK assigned a score of 0, neutral, to this sentence. The original sentence bears a clear positive message: the study abroad experience was mindblowing, and it changed the student's life forever (it is implicit that it changed it in a positive way.) Yet this positivity did not translate into the English version even though the sentence's translation was correct for both Google and Opus.

   I believe the problem here was the word "unique," which can have positive, negative, and neutral connotations according to the context. If unique is intended as being "peculiar," it has a negative meaning; if it is used to point out that something or someone is just "different," the word carries a neutral connotation. If it is used to indicate that something or someone has "no equal," then it is positive. It is possible that the neutral scores are justified by the fact that the engines perceived the study abroad experience as being simply "different." Furthermore, reflecting on the second part of the original sentence, "mi ha cambiato letteralmente la vita" (it changes my life completely), one could argue that a change in life is not always a positive event. It depends on the context and the person's perception of the events. From a linguistic point of view, the expression "L' esperienza studio in Italia è stata unica" in Italian is undoubtedly positive. In Italian, something that is unique to you is "positive." You would never use this expression to talk about something that was indifferent to you or negative. If an event is perceived as neutral, one would say, for example, "è stata un'esperienza normale," or "è stata un'esperienza come un'altra" ("it was a normal/ uneventful experience," or "it was an experience like any other." If it is negative, then one would say: "E' stata una brutta esperienza" or "Un'esperienza negativa." ("It was a bad experience" or "It was a negative experience.")

2. "Sei raggiante!" (original positive) [You are glowing! (Opus); You are radiant! (Google)] Google sentiment score: 0.5255 (positive) Opus Sentiment score: 0 (neutral) NLTK Sentiment score:1 (neutral)

   This is another interesting case. The original sentence is clearly positive. To tell someone they are glowing in Italian is to compliment them. Yet Opus and also NLTK assigned it a neutral score. Opus's score is even more interesting because the translation it provided is more accurate than the one provided by Google. The most sensible explanation for this mistake in sentiment analysis would be that the word "glowing" in English is used in a variety of expressions that also carry negative feelings.

   "glow with something. 1. Lit. [for something] to put out light, usually because of high heat. The embers glowed with the remains of the fire. The last of the coals still glowed with fire. 2. Fig. [for someone's face, eyes, etc.] to display some quality, such as pride, pleasure, rage, health. Her healthy face glowed with pride. Her eyes glowed with a towering rage." "[https://idioms.thefreedictionary.com/glowing:. . . . .: text=

   I believe that this different use of the word "glowing" contributed to the final calculation of the sentiment score and justified the neutral rating assigned by Opus. The NLTK score averages out the sentiment scores provided by Google and Opus and leans towards the neutral sentiment. However, it also provides a 0.629 score for positive sentiment, thus recognizing the intent of the original.

3. È stata una vacanza da sogno! (Original positive) [It was a dream vacation! (Opus); It was a dream holiday! (Google)]

   Google Sentiment score: 0.6114 (positive) Opus Sentiment score: 0.3164 NLTK: positive score 0.433; neutral score: 0.567

   This one is worth discussing for the disparity among the different scores. Although all of the "datasets" assigned a positive score, there was a significant difference between Google and the score provided by Opus and NLTK. Google's assignment of the score appeared to be much more confident; Opus and NLTK provided a positive score with a lower level of confidence (in fact, NLTK also assigned a higher neutral score to this sentence) How do we explain this? The translations are both good (even though the one provided by Google seems more proper in British English). A plausible explanation could be the fact that "a dream vacation" is something different for everyone, thus subjectivity plays a role in determining the positivity or neutrality of the sentiment.

4. Mi hai delusa! (Original negative) [You let me down! (Opus); You disappointed me! (Google)] Google sentiment score: −0.4767 Opus sentiment score: 0 (neutral) NLTK Neutral score: 1 NLTK Negative score: 0 TextblobSentimentpolarity: −0.1555556

   The translations provided are both good, with a slight preference for Opus, which is more exact from an idiomatic point of view. Google's score is perfect as it identifies the original sentiment. This is interesting because, as mentioned above, Google does not provide a better translation but still is on point with the sentiment score. However, despite providing a better translation, Opus read the sentence as neutral, and NLTK also assigned the sentence a neutral score. The sentiment polarity was a low negative.

   There is no doubt that the original sentence has a negative connotation, "mi hai delusa" is a sentence that expresses sadness, anger, and disillusionment. I would imagine that "You let me down!" works the same way. That is why the neutral score was a surprise and needs further investigation. In fact, as of now, there is no plausible explanation for the mistake.

5. Sei un inetto! (Original negative) [You're inept! (Opus); You're an inept! (Google)] Opus Google score: 0 (neutral) NLTK Neutral score: 1 NLTK Negative score: 0

Again, as in the case above, the original leaves no room for misunderstanding. In Italian culture, calling someone "inetto" is certainly an offense; hence the expression carries a negative sentiment. It is possible, however, that the sentence was read as a "personal opinion," which is obviously not universal and open to personal interpretation. This would justify the neutral score assigned.

6. Buono a nulla! (Original negative) [Good for nothing! (Opus & Google)] Google score: 0.4926 Opus score: 0.4926 NLTK Positive score: 0.615 NLTK Negative score: 0 NLTK Neutral score: 0.385

   In Italian, "Buono a nulla!" is another way to say "inept" and just like the sentence above expresses a negative sentiment. The error in sentiment analysis can be justified by the presence of the word "good," which is usually positive. The three datasets picked up the sentiment score carried by the word 'good' and consequently read the sentence as positive.

7. "In questo momento mi sento piuttosto calma non provo emozioni forti. (Original Neutral) [Right now, I feel rather calm; I don't feel strong emotions. (Opus); "At this moment, I feel quite calm I do not feel strong emotions. (Google)] Google score: −0.0281 Opus score: −0.1032 NLTK Positive score: 0.192 NLTK Negative score: 0.225 NLTK Neutral score: 0.583

   "Alla fine dei conti puoi fare quello che desideri, a me non interessa molto. (Neutral) [At the end of the accounts, you can do what you want. I don't care much. (Opus Google)] Google score: −0.3244 Opus score: 0.3705 NLTK Positive score: 0.076 NLTK Negative score: 0.166 NLTK Neutral score: 0.758

   "Non ho preferenze su cosa fare stasera." [I have no preference on what to do tonight.(Opus Google) Google score: −0.296 Opus score: −0.296 NLTK Positive score: 0 NLTK Negative score: 0.239 NLTK Neutral score: 0.787

   "Non ho mai favorito nessuno studente, per me sono tutti uguali." (neutral) [I have never favored any student, for me, they are all the same. (Opus Google)] Google score: −0.3252 Opus score: −0.3252 NLTK Positive score: 0 NLTK Negative score: 0.189 NLTK Neutral score: 0.811

   These sentences were presented as neutral in the original because they do not express positive or negative feelings or attitudes. Indeed, the "subjects" of the sentences are neither upset nor happy; neither in favor nor against a particular situation, they are simply "emotions/opinions free"; thus, the sentences are neutral. However, they were rated as negative by both Google and Opus, while the NLTK scores were more on point.

   Here are some possible explanations:

   The presence of the negative words "Non ho/I do not"; "Senza/Without" might have led the "analysis" in the wrong direction.

   Also, the absence of context might have had a role in leading to the wrong score.

   Indeed, some of these sentences could sound negative if pronounced with an upset tone. This is true, especially for these two sentences:

   "Alla fine dei conti puoi fare quello che desideri, a me non interessa molto." "A me non interessa" (I don't care much) can be negative if pronounced with an altered/upset tone. It can communicate a lack of "interest" and "feelings." However, if the same sentence (at least in Italian) is pronounced with a flat tone, then it just communicates "neutrality." "Non ho preferenze su cosa fare stasera." In this case, if the sentence is pronounced within the context of an argument, hence with an altered tone of voice, then it can have a negative feeling. But if a person says it just to express that "s/he would go with the flow," it is entirely neutral.

   Last but not least, two more sentences are worthy of attention:

8. "La musica americana attrae sempre molti giovani italiani." (Original neutral) ["American music always attracts many young Italians." (Opus Google)] Google score: 0.4019 Opus score: 0.4019 NLTK Positive score: 0.31 NLTK Negative score: 0 NLTK Neutral score: 0.69

   "Domani partiamo per andare in Italia." (Original neutral) ["Tomorrow we leave to go Italy." (Opus Google)] Google score: −0.0516 Opus score: −0.0516 NLTK Positive score: 0 NLTK Negative score: 0.167 NLTK Neutral score: 0.833

   These two sentences in Italian are plain statements. They express simple facts: American music is popular, and "tomorrow" we are going to Italy. Yet the first was rated with a positive score (only NLTK proposed a neutral score). The presence of the word "attracts," which intrinsically has a positive meaning, possibly led the datasets to identify this sentence as positive.

As for the second one, it is plausible that the word 'leave' which indicates "separation," might have led to the negative score.

5. Limitations

Our research in machine translations-based NLP solutions is presented as a lead study to establish a robust process at the intersection of state-of-the-art machine translations, English language NLP tools and languages other than English. For the purposes of this study, we use apply sentiment analysis, and two machine translations of an original Italian language dataset. There are a few limitations and these serve as areas of future research. We initiated the pilot project with a set of expert-created Italian language sentences. First, the dataset is specifically created for this study and not "real world" data in the sense of it being secondary data from external sources such as social media posts or news articles or blogs. Secondly, it is a small dataset, especially in the context of the LLMs which uses large quantities of data. Hence the findings may have limited external validity. However, since this is a lead study aimed at establishing a research process, the use of a custom dataset with limited size is justified given the rigorous and thorough analytical framework which is critical for a sustainable process and this enhances internal validity.

The third limitation is that we have tested only two machine translation models and fourth, only three sentiment analysis methods were applied. However, this is sufficient because this approach meets the goals of process centricity and process validations for this study. The use of two translations and three sentiment analysis methods ensures a simplified but rigorous approach for process validation, ensuring external validity. Therefore, in spite of the aforementioned limitations, the research accomplishes the main objective of the lead study, which is to establish a transparent and repeatable process for further and extensive analysis of the reliability of machine translation-based NLP solutions.

6. Future Research

Our lead study using English translations of Italian text has conceptually illustrated the usefulness of such an approach for extending the use of English language NLP tools to Italian text. Our future research will include additional languages, expand the size of the data analyzed, increase the number of machine translations applied, and explore the use of additional sentiment analysis methods. Incorporating open data into future research will be useful to facilitate public benefit and greater application potential (Samuel et al. 2023). We will also include the validation of additional NLP solutions such as identifications of topics and named entity recognition (NER). There is a significant need to establish best practices for machine translation-driven application of NLP solutions and future research should aim to address this need. In spite of recent calls to slow down NLP research and development, there is sufficient reason to believe that we will see rapid developments in this domain over the next few years (Samuel 2023a). Furthermore, within the broader context of artificial intelligence (AI), defined as the ability of machines to "mimic the functions and expressions of human intelligence, specifically cognition, and logic," it will be valuable to explore combining machine translations and multimodal approaches, including the recognition of images and handwritten text (Samuel 2021; Jain et al. 2023; Liu et al. 2023).

7. Conclusion

Our research affirms past studies that have illustrated the viability of using English translations of native texts with machine translation mechanisms for applications of sense-making methods and tools such as sentiment analysis (Balahur and Turchi, 2012, 2014). Furthermore, our study has created a new Italian dataset and a simple, repeatable, and effective process for testing and validating the use of English translations for NLP applications—this will enable us and other researchers to quickly validate many global languages for machine translations-based NLP solutions. Despite recent concerns over risks and ethics, NLP, generative, and adaptive AI technologies are expected to grow exponentially over the next few decades and have a significant societal impact (Samuel et al. 2022a; Samuel 2023b). In this context, we anticipate it will become increasingly important to use machine translations in conjunction with other NLP tools and AI technologies to address complex individual, community, and societal problems effectively. We anticipate an increased emphasis on machine translation-based NLP solutions to address issues of public importance and expect our novel process contribution to help applied NLP researchers develop solutions with greater efficiency.

References

Ali, G. M. N., M. M. Rahman, M. A. Hossain, M. S. Rahman, K. C. Paul, J. C. Thill, and J. Samuel. 2021. "Public Perceptions of Covid-19 Vaccines: Policy Implications from US Spatiotemporal Sentiment Analytics." *Healthcare* **9**, no. 9: 1110. doi: 10.3390/healthcare9091110

Araújo, M., A. Pereira, and F. Benevenuto. 2020. "A Comparative Study of Machine Translation for Multilingual Sentence-Level Sentiment Analysis." *Information Sciences* **512**: 1078–1102. doi: 10.1016/j.ins.2019.10.031

Balahur, A., and M. Turchi. 2012. "Multilingual Sentiment Analysis Using Machine Translation?" *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, Jeju, Republic of Korea, The Association for Computer Linguistics, July 12.

Balahur, A., and M. Turchi. 2014. "Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis." *Computer Speech & Language* **28**, no. 1: 56–75. doi: 10.1016/j.csl.2013.03.004

Basile, V., and M. Nissim. 2013. "Sentiment Analysis on Italian Tweets," *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, Association for Computational Linguistics, June 14.

Bawden, R., G. M. Di Nunzio, C. Grozea, I. J. Unanue, A. J. Yepes, N. Mah, D. Martinez, A. Névéol, M. Neves, M. Oronoz, O. Perez-de-Viñaspre, M. Piccardi, R. Roller, A. Siu, P. Thomas, F. Vezzani, M. V. Navarro, D. Wiemann, and L. Yeganova. 2020. "Findings of the Wmt 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages." *Proceedings of the Fifth Conference on Machine Translation*, Online, Association for Computational Linguistics, November 19.

Berard, A., I. Calapodescu, M. Dymetman, C. Roux, J. L. Meunier, and V. Nikoulina. 2019. "Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness." Preprint, submitted October 31. https://doi.org/10.48550/arXiv.1910.14589

Catelli, R., L. Bevilacqua, N. Mariniello, V. S. di Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito. 2022a. "Cross Lingual Transfer Learning for Sentiment Analysis of Italian Tripadvisor Reviews." *Expert Systems with Applications* **209**: 118246. doi: 10.1016/j.eswa.2022.118246

Catelli, R., S. Pelosi, and M. Esposito. 2022b. "Lexicon-Based vs. Bert-Based Sentiment Analysis: A Comparative Study in Italian." *Electronics* **11**, no. 3: 374. doi: 10.3390/electronics11030374

Cunliffe, D., A. Vlachidis, D. Williams, and D. Tudhope. 2022. "Natural Language Processing for under-Resourced Languages: Developing a Welsh Natural Language Toolkit." *Computer Speech & Language* **72**: 101311. doi: 10.1016/j.csl.2021.101311

Dashtipour, K., S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou. 2016. "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques." *Cognitive Computation* **8**, no. 4: 757–771. doi: 10.1007/s12559-016-9415-7

Han, S. 2022. "googletrans: Free and Unlimited Google Translate API for Python." Accessed February 27, 2023. https://github.com/ssut/py-googletrans

Jain, P. H., V. Kumar, J. Samuel, S. Singh, A. Mannepalli, and R. Anderson. 2023. "Artificially Intelligent Readers: An Adaptive Framework for Original Handwritten Numerical Digits Recognition with OCR Methods." *Information* **14**, no. 6: 305. doi: 10.3390/info14060305

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, *et al.* 2018. "Marian: Fast Neural Machine Translation in C++." *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, Association for Computational Linguistics, April 14. http://www.aclweb.org/anthology/P18-4020

Kumar, P., K. Pathania, and B. Raman. 2023. "Zero-Shot Learning Based Cross-Lingual Sentiment Analysis for Sanskrit Text with Insufficient Labeled Data." *Applied Intelligence* **53**, no. 9: 10096–10113. doi: 10.1007/s10489-022-04046-6

Lahoti, P., N. Mittal, and G. Singh. 2023. "A Survey on NLP Resources, Tools, and Techniques for Marathi Language Processing." *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**, no. 2: 1–34. doi: 10.1145/3548457

Liu, H., R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang. 2023. "Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4." Preprint, submitted May 5. https://doi.org/10.48550/arXiv.2304.03439

Markets and Markets. 2022. "Natural Language Processing Market." Accessed September 5, 2022. https://www.marketsandmarkets.com/Market-Reports/natural-language-processing-nlp

Modzelewski, A., W. Sosnowski, M. Wilczynska, and A. Wierzbicki. 2023. "Dshacker at Semeval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation." *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada, Association for Computational Linguistics, July 13.

Mohammad, S. M., M. Salameh, and S. Kiritchenko. 2016. "How Translation Alters Sentiment." *Journal of Artificial Intelligence Research* **55**: 95–130. doi: 10.1613/jair.4787

Nayak, P. 2021. “MUM: A New AI Milestone for Understanding Information.” Google. Accessed March 2023. https://blog.google/products/search/introducing-mum/

Nguyen, X. P., S. M. Aljunied, S. Joty, and L. Bing. 2023. “Democratizing LLMs for Low-Resource Languages by Leveraging Their English Dominant Abilities with Linguistically-Diverse Prompts.” Preprint, submitted June 20. https://doi.org/10.48550/arXiv.2306.11372

Oueslati, O.,. E. Cambria, M. B. HajHmida, and H. Ounelli. 2020. “A Review of Sentiment Analysis Research in Arabic Language.” *Future Generation Computer Systems* **112**: 408–430. doi: 10.1016/j.future.2020.05.034

Poncelas, A., P. Lohar, A. Way, and J. Hadley. 2020. “The Impact of Indirect Machine Translation on Sentiment Classification.” Preprint, submitted August 25. https://doi.org/10.48550/arXiv.2008.11257

Porreca, A., F. Scozzari, and M. Di Nicola. 2020. “Using Text Mining and Sentiment Analysis to Analyse Youtube Italian Videos Concerning Vaccination.” *BMC Public Health* **20**, no. 1: 259. doi: 10.1186/s12889-020-8342-4

Proksch, S. O., W. Lowe, J. Wäckerle, and S. Soroka. 2019. “Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches.” *Legislative Studies Quarterly* **44**, no. 1: 97–131. doi: 10.1111/lsq.12218

Rahman, M. M., G. M. N. Ali, X. J. Li, J. Samuel, K. C. Paul, P. H. Chong, and M. Yakubov. 2021. “Socioeconomic Factors Analysis for COVID-19 US Reopening Sentiment with Twitter and Census Data.” *Heliyon* **7**, no. 2: e06200. doi: 10.1016/j.heliyon.2021.e06200

Ranathunga, S., and N. de Silva. 2022. “Some Languages Are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World.” Preprint, submitted October 20. https://doi.org/10.48550/arXiv.2210.08523

Russo, L., S. Loáiciga, and A. Gulati. 2012. “Improving Machine Translation of Null Subjects in Italian and Spanish.” *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, Association for Computational Linguistics, April 16.

Samuel, J. 2021. *A Call for Proactive Policies for Informatics and Artificial Intelligence Technologies*. Boston, MA: Scholars Strategy Network.

Samuel, J. 2023a. “Response to the March 2023 ‘Pause Giant Ai Experiments: An Open Letter’ by Yoshua Bengio, Signed by Stuart Russell, Elon Musk, Steve Wozniak, Yuval Noah Harari and Others.” Preprint, submitted March 29. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4412516

Samuel, J. 2023b. “The Critical Need for Transparency and Regulation amidst the Rise of Powerful Artificial Intelligence Models,” accessed August 2, 2023, https://scholars.org/contribution/critical-need-transparency-and-regulation

Samuel, J., G. Ali, M. Rahman, E. Esawi, Y. Samuel. 2020a. “COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification.” *Information* **11**, no. 6: 314. doi: 10.3390/info11060314

Samuel, J., M. Brennan, M. Pfeiffer, C. Andrews, and M. Hale. 2023. “Garden State Open Data Index for Public Informatics.” NJSPL Report. https://policylab.rutgers.edu/report-release-garden-state-open-data-index/#_ftn1

Samuel, J., R. Kashyap, Y. Samuel, and A. Pelaez. 2022a. “Adaptive Cognitive Fit: Artificial Intelligence Augmented Management of Information Facets and Representations.” *International Journal of Information Management* **65**: 102505. doi: 10.1016/j.ijinfomgt.2022.102505

Samuel, J., M. M. Rahman, G. M. N. Ali, Y. Samuel, A. Pelaez, P. H. J. Chong, and M. Yakubov. 2020b. “Feeling Positive about Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics.” *IEEE Access* **8**: 142173–142190. doi: 10.1109/ACCESS.2020.3013933

Samuel, J., R. Palle, and E. Soares. 2022b. “Textual Data Distributions: Kullback Leibler Textual Distributions Contrasts on GPT-2 Generated Texts with Supervised, Unsupervised Learning on Vaccine & Market Topics & Sentiment.” *Journal of Big Data: Theory and Practice* **1**, no. 1. doi: 10.54116/jbdtp.v1i1.20

Sazzed, S. 2020. “Cross-Lingual Sentiment Classification in Low-Resource Bengali Language.” *Proceedings of the Sixth Workshop on Noisy User-Generated Text (W-NUT 2020)*, Online, Association for Computational Linguistics, November 19. http://dx.doi.org/10.18653/v1/2020.wnut-1.8

Sazzed, S., and S. Jayarathna. 2019. “A Sentiment Classification in Bengali and Machine Translated English Corpus.” *Proceedings of the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Los Angeles, CA, IEEE, July 30. https://doi.org/10.1109/IRI.2019.00029

Sebastian, M. P. 2023. “Malayalam Natural Language Processing: Challenges in Building a Phrase-Based Statistical Machine Translation System.” *ACM Transactions on Asian and Low-Resource Language Information Processing* **22**, no. 4: 1–51. doi: 10.1145/3579163

Tiedemann, J. 2020. “The Tatoeba Translation Challenge–Realistic Data Sets for Low Resource and Multilingual MT.” *Proceedings of the Fifth Conference on Machine Translation*, Online, Association for Computational Linguistics, November. https://aclanthology.org/2020.wmt-1.139

Tiedemann, J., and S. Thottingal. 2020. "OPUS-MT–Building Open Translation Services for the World." *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, European Association for Machine Translation, November 3–5. https://aclanthology.org/2020.eamt-1.61

Wang, S., Y. Sun, Y. Xiang, Z. Wu, S. Ding, W. Gong, S. Feng, *et al.* 2021. "Ernie 3.0 Titan: Exploring Larger-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation." Preprint, submitted December 23. https://doi.org/10.48550/arXiv.2112.12731

Wiesmann, E. 2019. "Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German." *Comparative Legilinguistics* **37**, no. 1: 117–153. doi: 10.14746/cl.2019.37.4