Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Predicting News Outdatedness Using a Probabilistic Logistic Decay Model

HWR Berlin (Berlin School of Economics and Law)
Business Intelligence and Process Management SS2025/2026

Supervisors:	Prof. Dr. Diana Hristova
	Prof. Dr. Roland M. Müller
Submission:	18.07.2025
Author:	Sheilla Mahardika Purwandiary
Student ID No:	77212008343
Word count:	13,166

Abstract

Detecting outdated news is as important as identifying fake news. While it widely recognized that we should be cautious of fake news, fewer people are equally mindful of the risks posed by outdated information.

This thesis focuses on developing probabilistic model to estimate when news articles become outdated using logistic decay approach.

This study develops an annotated dataset to be used in fitting a logistic decay function through the fine-tuning of a BERT model. A semi-automated annotation approach is applied, combining rule-based logic with manual annotation to create the training set. The resulting fine-tuned BERT model achieves an accuracy of 90.36% in classifying article outdatedness. Based on this annotated data, the logistic decay function is shown to effectively model the gradual decline in news relevance over time[1].

This study concludes by recommending future directions such as employing news article topic or category into the modelling, using TimeBERT instead of base BERT, adding user engagement and multi-class classification of outdatedness for more nuanced and accurate outdatedness modelling. This research provides a foundation for understanding the potential and challenges in news articles and temporal decay for further exploration in this field.

Keywords: News Article Outdatedness, Logistic Decay Function, Probabilistic Model, BERT, BERT Fine-tune, Logistic Function, Temporal Decay, Bernoulli Distribution, Text Classification, Binary Classification, Data Freshness, Data Relevance

All code and documents are publicly available at:
https://github.com/shellamp/Thesis_PredictingNewsOutdatedness_Logistic_Decay

[1] AI-assisted phrasing. See AI Directory 20

Acknowledgements

I would like to thank Prof. Dr. Diana Hristova and Prof. Dr. Roland Müller for the help during my thesis research at Berlin School of Economics and Law (HWR Berlin). I would also like to thank my classmates in our master program for the great collaboration throughout our degree and for creating such a great and fun working environment.
Finally, special shout-outs to Jco, Thobie Jovian, my family and friends for their continuous support during my studies and particularly during this last phase in my master's degree.

Declaration of Authorship

I hereby declare

- that I have written the submitted academic work entirely by myself without anyone else's assistance, and that I have not used any sources or aids other than those stated. Wherever I have drawn on literature or other sources, either in direct quotes, or in paraphrasing such material, I have given, in accordance with academic standards, the reference to the original author or authors and to the source where it appeared.
- that if I have used AI-based tools that were classified as reportable at the time of submission of my paper, I have fully listed them in the "AI directory" section with the product name, the source (e.g., URLs), the purpose of use, details of the parts of the academic work affected, as well as the inputs/prompts. In addition, I have indicated the use of AI in the respective sections in the main text.

I am aware

- that the use of quotations, or of close paraphrasing, from books, magazines, newspapers, the internet or other sources, which are not marked as such, or the use of AI tools without detailed documentation, will be considered as an attempt at misconduct, and that the academic work will be graded with a fail. I have notified the examiners and the board of examiners in the case that I have submitted the academic work, either in whole or in part, for other examination purposes.
- that AI tools do not constitute scientific sources and therefore cannot be cited as such.
- that if I have used AI-based tools to create this academic work, I am responsibility for

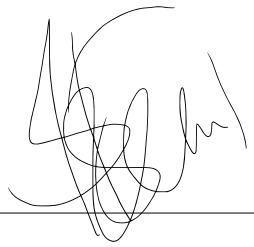
any incorrect or biased content, incorrect references, as well as violations of data protection, copyright, or plagiarism that may have been generated by the AI.

I hereby declare that I

☒ have used AI-based tools in my submitted academic work and documented their use in accordance with the above requirements.

☐ have not used any AI-based tools in my submitted academic work.

Berlin, 18.07.2025



Place, date

signature

Table of Contents

Abstract *2*
Acknowledgements *3*
Abbreviations *8*
List of Figures ..... *Error! Bookmark not defined.*
List of Tables *10*
1. Introduction *11*
1.1. Research Objective 12
1.2. Thesis Outline 12
2. Theoretical Foundations *13*
2.1. News Relevance and Outdatedness 13
2.2. Probabilistic Models 14
2.2.1. Bernoulli Distribution 14
2.2.2. Logistic Decay Function 15
2.3. Natural Language Processing (NLP) 16
2.3.1. Transformers 17
2.3.1.1. Transformer Architecture 18
2.3.1.2. Encoder and Decoder Structure 19
2.3.1.3. Positional Encoding 19
2.3.1.4. Self-Attention and Multi-Head Attention Mechanism 20
2.3.1.5. Feed Forward Network (FFN) 20
2.3.1.6. Normalization and Residual Connect 20
2.3.1.7. Bidirectional Encoder Representations from Transformers (BERT) 20
2.3.2. Fine-tuning BERT 21
2.3.3. Hyperparameter Tuning 22
2.3.4. Text Classification with Transformer 22
2.3.5. Named Entity Recognition (NER) 23
2.3.5.1. SpaCy 23
3. Related Work *23*
4. Methodology *25*
4.1. Data 25
4.1.1. Data Sourcing 25
4.1.2. Exploratory Data Analysis (EDA) 25
4.1.3. Data Preprocessing 28
4.1.3.1. Data Cleaning 28
4.1.3.2. Data Enrichment 30

4.1.4. Data Splitting 30
4.2. Modelling Methodology 31
4.2.1. Data Annotation Methodology for Outdatedness Detection 32
4.2.1.1. Training set 32
4.2.1.1.1. Rule-Based 32
4.2.1.1.1.1. Rule-Based Validation 33
4.2.1.1.2. Manual-based 34
4.2.1.1.2.1. Inter-annotator Agreement 35
4.2.1.1.3. Data Augmentation 36
4.2.2. BERT Fine Tuning 36
4.2.2.1. Hyperparameter Tuning 37
4.2.3. Probabilistic Model Development 39
4.2.3.1. Logistic Decay Fitting 39
4.3. BERT Fine-tuned Classification Model 40
4.4. Probability Model 41
5. Conclusion 43
5.1. Limitation and Further Research 44
References 46
Appendices 54
Appendix A – AI Directory 54
Appendix B – Sample raw articles 55
Appendix C – Sample annotated articles 55

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BT	Back-translation
EDA	Exploratory Data Analysis
EGL	Expected Gradient Length
FFN	Feed Forward Network
IAA	Inter-Annotator Agreement
LC	Least Confidence
NLP	Natural Language Processing
PE	Perceptron Ensemble
ReLU	Rectified Linear Unit
RQ	Research Question
TF-IDF	Term Frequency-Inverse Document Frequency

List of Figures

Figure 1: Transformer Architecture (Vaswani et al., 2017) 18
Figure 2: Encoder and Decoder (own illustration) 19
Figure 3 : Summary Length Distribution 26
Figure 4: Workflow (own illustration)..... 31
Figure 5 : Final Model training and validation loss and performance 40
Figure 6 : Decay and Residual Plot 42

List of Tables

Table 1: Classification report of rule-based annotation 33
Table 2: Summary of hyperparameter tuning 38
Table 3: Performance summary table on test set 40
Table 4: Exponential vs Logistic Performance 41

1. Introduction

One way to stay informed on the state of the world in this digital's era is to read information online. However, there are thousands if not hundreds of thousands of new pieces of content that are published every day. This massive flow of information makes it hard to always be vigilant of what is current and trustworthy.

This information overload reflects a trend that people today interact with a growing volume of unstructured data, particularly in text form.[2] This includes content from blogs, news website and social media. According to Gandomi and Haider (2015) unstructured data compose about 95% of today's data, comprise data from mobile apps, social networks, and sensors.

Among them, news remains popular. As technology evolved, news has become more easily accessible and widely distributed. A report by the Reuters Institute (2023) shows that digital media is now the most common ways people consume news globally.

However, the ease of access and the fast spread of news also brings major challenge, users are often being exposed to outdated news. This is often because they are unaware of how to detect them. While Fake News means that the news itself is deliberately false, outdated news means information that was once true or relevant but has since become no longer valid. As Ferranti et al. (2021) put it, outdated news is "news whose correctness is dependent on a specified time interval," which means whether a news article is still accurate depends on when you read it.

Commented [HD1]: Add source for taht definition

Data freshness or data relevance in information system, has been recognized as one of the most crucial data quality aspects (Bouzeghoub, 2004). Similarly, Ferranti et al. (2021) mention that it is critical to work with current information as outdated news can be misleading and can affect how people think long after it should have been dismissed. For instance, a study by Peterson, Christopher J et al. (2022) found that withdrawn COVID-19 articles continued to be shared on social media and cited in news coverage for months after removal. In fact, 44.2% of all mentions occurred after the articles had already been retracted. This continued visibility of outdated content causes confusion and misinformation during the pandemic.

[2] AI-assisted phrasing. See AI Directory 3

To address such risks, this study proposes building a probabilistic model that can estimate the chance that a news article is outdated. A probabilistic model is a type of model that doesn't just give a simple yes/no answer, it gives a probability instead. Probabilistic models are particularly useful for tasks involving uncertainty and time-dependent changes, as they provide a degree of belief rather than a rigid classification (Murphy, 2012). This is helpful because outdatedness is not always clear-cut. Some articles stay relevant longer than others, and their relevance fades gradually.

To do this, the study applies a logistic decay model, which is commonly used for binary outcomes (Murphy, 2012). This means that the model will estimate whether an article is outdated or not but the probability of outdatedness will increase over time in a curve that fits the way information usually loses relevance. Wang et al., (2020) also mentioned that due to its simple principle and efficient calculation, logistic is often used in regression fitting of time series data. Thus, this kind of model is useful because it reflects real life more closely than models that only give a fixed yes or no answer.

By doing so, this study aims to offer a practical solution to the challenge of outdated news in digital information environments, supporting better decision-making and reducing misinformation across multiple domains.

1.1. Research Objective

The thesis investigates the logistic decay model in predicting the probability of news outdatedness. The primary research question formulated as:

RQ: "*How can a probabilistic model to predict the outdatedness of a news data using logistic decay model be developed?"*

Since no publicly available news articles dataset annotated with their outdatedness exist, an additional objective of this thesis is to create annotated dataset to fit into the logistic decay function by fine-tuning BERT. This is because, Fine-tuning a pre-trained BERT models is a popular strategy for text classification due to its strong performance compare to traditional machine learning classifiers (Salih et al., 2025).

1.2. Thesis Outline

This thesis is structured into 6 chapters. Chapter 1 provides background motivation and research question of the thesis. Chapter 2 focuses on the theoretical foundations to provide a detailed understanding of the topic. Chapter 3 surveys prior

related research and literature on the thesis topic. Chapter 4 describes the methodology, outlining the approaches and procedures used in the research. Chapter 5 presents the study's results and offers a critical discussion of their implications. Finally, Chapter 6 summarizes the key findings, addresses the main research question, and highlights the study's limitations and potential directions for future research.

2. Theoretical Foundations

This chapter provides an overview of the theoretical foundations, theories and methods necessary for understanding the topic. First, news outdatedness is defined, followed by the explanation of probability model which includes Bernoulli distribution and logistics decay function. NLP, Transformers and its architecture are explained next, which is the building base of BERT. Further, expanded view of BERT with its Fine-Tune and its application in text classification. Lastly, the overview NER and Spacy are elaborated.

2.1. News Relevance and Outdatedness

Bouzeghoub (2004) defines data freshness by combining two ideas. One is Segev & Fang's, (1990) emphasis on currency, which characterizes how outdated the data is in relation to its source, and the other is R. Y. Wang & Strong's, (1996) emphasis on timeliness, which explains how old the data is. This definition is also what is adapted in this study, where news article is considered outdated once the time has passed long enough that the fact become obsolete and not relevant anymore.

This definition highlights that there are two factors in determining data freshness:

1. Content of the data
2. The time of publication of the data

Additionally, Jatowt et al. (2024) introduces the concept of contemporary relevance where the relationship of the data contents to the present times is assessed. The closer the data to the current times context, the more attractive and informative it become to the readers. This shows that data has a limited information lifespan, where it values, or relevance decreases as time passes. As Finger & Da Silva, (1998) puts it, "the confidence on the validity of data decays with time". This decline in relevance or validity of the data can be described as temporal decay. It is

especially true in domain like news, where hundreds if not thousands of new information are published every day. According to Ocaña et al. (2021) news organizations risk losing money and viewers if they don't publish new emerging events. Because of this, there is a continuous influx of new news contents every day, as a result this speeds up the process of news articles becoming obsolete. Barkemeyer et al. (2020) even observed that traditional news media cycles have accelerated with the rise of internet and social media, leading to faster decay of news relevance than in earlier decades. In other words, information freshness disappears more quickly in the modern news ecosystem.

2.2. Probabilistic Models

Probabilistic models are based on probability theory and rely on the idea that observed variables can be treated as random processes, which follow a known or estimated probability distribution[3] (Long et al., 2020). In binary classification tasks, to predict whether a news article is outdated or not, there are only two possible outcomes. Applying a simple linear model directly to such binary data is not ideal, Sainani, (2014) explained this is because a line may not capture the actual distribution of results. Instead, logistic regression changes the binary outcome using the logit function, which maps probabilities to a continuous scale of log odds. This transformation allows a linear relationship to be modeled more effectively. The logistic function then converts the logit back into a probability between 0 and 1. Under this framework, the outcome variable is assumed to follow a Bernoulli distribution, making logistic regression a natural choice for binary classification problems[4]. Since this study is modeling a decay, logistic decay function is a natural choice to model the distribution of the outcomes.

2.2.1. Bernoulli Distribution

The Bernoulli distribution is a probability distribution that models a random variable with binary outcomes, 1 or 0. A singe parameter p, represents the probability of success or 1 and while 0 or the probability of failure is $1 - p$. Thus, Bernoulli

[3] AI-assisted phrasing. See AI Directory 17
[4] AI-assisted phrasing. See AI Directory 18

distribution can be expressed as follows, where X is the random variable of the outcomes (Goodfellow et al., 2016):

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p$$

2.2.2. Logistic Decay Function

The logistic function originated from 19th century studies of population growth by Pierre François Verhulst (Sharma et al., 2022). He was building on Malthusian exponential growth model, by proposing the following logistic equation (P. Wang et al., 2020):

$$\frac{dQ}{dt} = rQ(1 - \frac{Q}{K})$$

Here, Q represents the population size, r is intrinsic growth, and K indicate the maximum population size that the environment can support. The growth of the population is represented by $\frac{dQ}{dt}$. The value of Q changes over time to produce an S-shaped curve in the logistic equation, while r and K are constants number. When analyzing real-world data, logistic models work better than exponential models. This is because, despite exponential models being helpful in the short term, the growth rate of exponential growth remains the same regardless of population size (Sharma et al., 2022).

To describe this S-shaped growth more generally, the logistic function can also be written as:

$$P(t) = \frac{c}{1 + \alpha e^{-bt)}}$$

In this equation P is the population at time t, c is the carrying capacity, α is related to the initial population size and b control the steepness of the curve. If P is much smaller than K, the growth rate is positive and roughly proportional to P. As P nears K, the growth rate slows and eventually stabilizes. When P exceeds K, the growth

rate can become negative, illustrating how a population may decline once it surpasses the sustainable limit (Sharma et al., 2022). [5]

To model decay, the logistic decay model inverts the curve by negating the slope parameter b. The switches represent a decreasing process instead of growth, making it useful for describing phenomena, like resource depletion or declining relevance (Sharma et al., 2022).

The logistic decay function thus can be shown by the following equation:

$$P(t) = \frac{c}{1 + \alpha e^{-(-bt)}}$$

In news articles outdatedness case, logistic decay is characterized by a slow decrease of the "datedness" or relevance of the news until reaching the fastest decay point. From that point, it decays quickly until stabilization is reached. Once it is reached, the article is considered outdated or no longer relevant for people due to the time pass since its publication date, making the fact obsolete.

The formula is as follows:

$$P(X = 1) = \frac{1}{1 + e^{\lambda(t - t_0)}}$$

where:

- λ = Decay rate (how fast outdatedness occurs) and $\lambda > 0$
- t_0 = Half-life (time when outdatedness probability reaches 50%).
- t = Time since publication.

2.3. Natural Language Processing (NLP)

NLP is the foundation of modern Artificial Intelligence (AI) (Gupta & Choubisa, 2024) and has been one of the fastest growing field in recent years (Ayorinde et al., 2025). NLP's goal is to have computers, the power to process the complex and nuances of human language automatically, either in written or spoken form for various purposes (Dahl, 2010; Gupta & Choubisa, 2024). By doing so, it helps bridging the gap between human and computer interactions, making it easier for people to 'communicate' with machine. Gupta & Choubisa (2024) elaborates the many tasks in NLP including sentiment analysis, language translation and chatbot interactions. This

[5] AI Assisted phrases. See AI Directory 19

can be achieved by using computational algorithms, linguistic theories and AI. However, before the rise of AI and deep learning, Ayorinde et al., (2025) describes that early approaches were mainly rule-based and heavily dependant on predefined grammar rules. This means the system could only handle simple tasks, for example word-to-word translation. The significant breakthrough happens in 2010 when Word2Vec and GloVe revolutionized the understanding of word meanings. They are able to comprehend and capture semantic relationships in vast amounts of text data. Since then, deep learning models like Transformers have become the pillars in modern NLP systems. (Ayorinde et al., 2025).

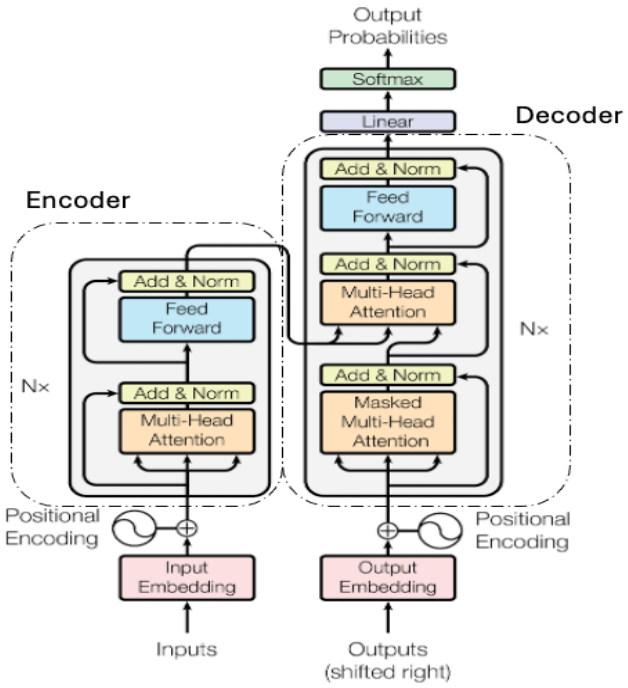
2.3.1. Transformers

The Transformers model, first introduced by Vaswani et al. (2017), has revolutionized the field of NLP, due to its ability to excel from the previous model like CNNs and RNNs in language comprehension and generating (Gardazi et al. 2025). Unlike RNNs/LSTMs, this architecture doesn't use recurrence at all, which mean it can process all parts of the data at once. It also uses attention mechanism to understand relationship between words. The attention mechanism is the key innovation that differentiate transformer models than the more traditional models (Vaswani et al., 2017). The transformer is thus boosting the perfomance speed of pretraining on large dataset, which significantly enhance the accurancy and efficiency of complex language tasks such as text classification and machine translation. Gardazi et al., (2025) highlight that transformer "has propelled NLP forward, enabling more robust and nuanced language models capable of deeper language understanding and analysis".

2.3.1.1. Transformer Architecture

Originally, the transformer was developed for sequence transduction tasks, such as machine translation (Werner, 2023). Most advanced sequence transduction models have an encoder-decoder structure, which the transformer also have as illustrated in Figure 1. In its simplest form, the transformer turns inputs (e.g., French “Comment ça va ?”) into output (e.g., English *“How are you?”*). To achieve this, there are two key innovations in transfomer’s setup, which are stacked self-attention and point-wise. This combination allows transformer to handle long-range dependencies parallely (Vaswani et al., 2017).

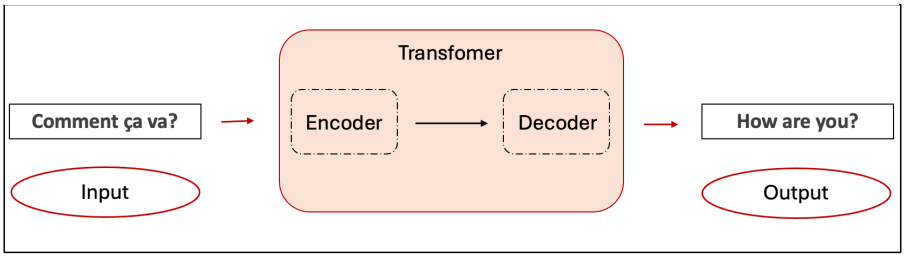
Figure 1: Transformer Architecture (Vaswani et al., 2017)



2.3.1.2. Encoder and Decoder Structure

The transformer's architecture follows the Enconder and Decoder structure, where encoder takes in and process the input text while the decoder produce the output text as shown in Figure 2.

Figure 2: Encoder and Decoder (own illustration)



The encoder and decoder in the transformer are not single layers, rather they consist of multiple stacked layers. Importantly, the encoder and decoder always have the same number of layers. In the original architecture proposed by Vaswani et al., (2017), this number was set to 6 layers each, but the design is flexible: the model can be scaled up or down with N layers depending on the task's requirements[6].
All encoder layers have the same structure and the input text goes through each encoder layer one after another, with each layer processing the output of the previous one. Similarly, all decoder layers have the same structure. However, decoder would receives two inputs, which are the final encoder output and its own output from the previous decoder.

2.3.1.3. Positional Encoding

Transformer use positional encodings to provide information about the position of each token in the sequence (Werner, 2023). This needs to be done because unlike RNNs, the model processes all words in a sentence at the same time. Vaswani et al., (2017) proposed using sinusoidal functions (sine/cosine) to generate these encodings, ensuring each position in the sequence has a unique and interpretable representation[7].

[6] AI-assisted phrasing. See AI Directory 4

[7] AI-assisted phrasing. See AI Directory 5

2.3.1.4. Self-Attention and Multi-Head Attention Mechanism

One of the components that differentiate this model compared to the traditional models is the self-attention mechanism. Unlike traditional approaches, it introduced:

- Self-Attention: It enables the model to relate each with word in the input text with another word and parallelly process on different parts of the input when processing each word (Vaswani et al., 2017).
- Multi-Head Attention: the model run self-attention not only once, but multiple times in parallel. Each 'head' learn different type of relationships, this allows the model to jointly attend different patterns (Vaswani et al., 2017).

2.3.1.5. Feed Forward Network (FFN)

Inside each layer of the encoder and decoder contains a fully connected feed-forward network (Werner, 2023). This consists of two linear transformations layers with ReLu (Rectified Linear Unit) activation, applied identically independently to each token's representation one after the other (Vaswani et al., 2017). This helps words develop deeper and context-aware at each layer.

2.3.1.6. Normalization and Residual Connect

ReLu outputs can vary wildly, causing unstable shifts in the next layer's inputs. To address this:

- Layer Normalization (Ba et al., 2016) normalizes activations within each layer.
- Residual Connections (He et al., 2015) adds the sub-layer's input to its output.

This combination helps in mitigating vanishing gradients problem, which allow in creating deeper models.

2.3.1.7. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a further development of the transformer model (Pham, 2024). González-Carvajal & Garrido-Merchán (2023) mentioned that BERT is capable in handling NLP tasks such as supervised text classification. It has gained popularity in academia and industry due to its flexibility and effectiveness in delivering great results across different corpora.

BERT uses the multilayer bidirectional transformer explained in 2017 paper's by Vaswani (Koroteev, 2021). A large Wikipedia dataset is trained for BERT using two unsupervised tasks: masked language modeling and next sentence prediction. This enable BERT learn deeply the context of words based on the full sentence from both direction of the sentence rather than just from left-to-right or right-to-left, which leads to better performance in NLP tasks (Pham, 2024). As a result, BERT has superior performance in question answering and language inference, without modifying task-specific architecture (Devlin et al., 2019). Conceptually BERT is simple and empirically powerful (Devlin et al., 2019). In addition to that, it has superior results in many automated word processing tasks compare to the other model such as word2vec, which as a result is preferred by many people and thus is becoming the industry standard (Koroteev, 2021).

2.3.2. Fine-tuning BERT

BERT is a powerful model, however because it is not specifically designed for a particular downstream task, it needs to be fine-tuned in order to perform well for a specific NLP task (Chun et al., 2023). Salih et al. (2025) demonstrated that fine-tuned BERT surpassed not only classical machine learning algorithms but also not fine-tuned BERT in performing text classification specifically news classification in their study.

According to Chun et al. (2023), the concept of fine-tune was introduced by Radford et al. (2018) in their paper called "Improving Language Understanding by Generative Pre-Training". This paper shows the potential of fine-tuning pre-trained models for downstream tasks such as text classification and question answering.

Fine-tuning can be described as a process to train pre-trained model (e.g. BERT) for a few times on a supervised dataset (Mosbach et al., 2020a). This allows the model to improve the performance on target tasks without requiring the computational cost of training the model from scratch. Devlin et al. (2019) explained that fine-tuning BERT is a straightforward process since the self-attention mechanism in the transformer allows BERT to model many specific tasks (single text or text pairs) just by changing inputs/outputs.

2.3.3. Hyperparameter Tuning

Hyperparameter refers to parameter values that can be set by user before training an algorithm (Montesinos-López et al., 2022). It impacts the behavior of the algorithm, which in the end also effect the performance of the model due to its sensitivity to the values of the hyperparameters (Wong et al., 2019). Thus, Hyperparameter tuning can be defined as the process in choosing the best value in machine learning model to optimize the performance of the model (Montesinos-López et al., 2022). The process is usually done by trying out different hyperparameters and compare the outcomes with the performance measures such as F1 score and accuracy (Rahmi et al., 2024).

Shekhar et al. (2022) explained that classification, clustering and other machine learning algorithms for the learning tasks are linked with parameters and hyperparameters. Parameters in algorithm are the ones that can be learned through optimization of a loss function or through the gradients. While hyperparameters, unlike parameters, control the learning process and cannot be inferred during the model fitting (Shekhar et al., 2022).

2.3.4. Text Classification with Transformer

Text classification, or text categorization, is the automated process of assigning known categories to documents based on their content (Sebastiani, 2005). Alam et al. (2020) mentioned that text classification has been one the earliest problem in Natural Language Processing (NLP). Similarly, Aliwy & Ameer (2017) considered it as one of the most important field in NLP due to its wide applications into many real world's scenarios such as spam filtering and information retrieval.

Algorithm such as Decision Tree, Support Vector Machine, KNearest Neighbors are among the commonly used classical techniques in text classifications. However, (Alam et al., 2020) stated that over time the complexity of text classification increased as the applications' scope evolved. This growing difficulty has created the need to switch from the traditional machine learning approaches to deep learning algorithms. Transformers have emerged as one of the more recent and popular techniques in deep neural network architectures (Alam et al., 2020). Among these BERT has become one of the highly researched and a common model used for text classification task (Zaman-Khan et al., 2024).

2.3.5. Named Entity Recognition (NER)

Named Entity Recognition (NER) is widely regarded as a fundamental component of Natural Language Processing (NLP). Chavan & Patil (2024) consider NER the cornerstone of NLP, highlighting its role in automating the extraction and categorization of named entities from text document. Similarly, Chan et al. (2025) emphasize that NER is a basic building block of NLP systems, as it enables the identification and classification of entities in unstructured text. Supporting this view, Abilio & Coelho (2024) define NER more simply as a technique for extracting information from textual documents.[8] Typical entities that can be extracted are people, names, organizations, locations, dates, and more (Pakhale, 2023; Salah et al., 2024).

2.3.5.1. SpaCy

SpaCy is a free, open-sourced Python library designed for various NLP tasks (Akhtar et al., 2024). It offers integrated NER models to recognize entities such as names, organizations, times, and locations (Satheesh et al., 2020).
SpaCy may has lower performance compared to transformer-based models on complex tasks. Regardless, SpaCy is lightweight and fast and offers explainable and ready-to-use framework, which is ideal for many NER applications (Chan et al., 2025).

3. Related Work

Detecting outdated information in news and other text data has gained attention as a distinct research problem in recent years. (Almquist & Jatowt, 2019) pointed out that humans often struggle to judge whether information remains valid over time. They proposed a machine learning approach to forecast the outdatedness of sentences using only linguistic features.
Furthermore, Ferranti et al. (2021) introduced outdated news detection as a special case of fact-checking, defined by news content “whose correctness is dependent on a specified time interval”. Their work introduced a knowledge graph-based workflow

[8] AI-assisted phrasing. See AI Directory 16

that compares facts in news articles to current data from sources like Wikidata and DBpedia. However, one challenge is apparent, if the KG data is not updated with real-world changes, the system may miss outdated facts or, worse, reinforce outdated information. For example, Ferranti , Krickl and Nissl (2021) observed that Wikidata and DBpedia sometimes contained conflicting or stale facts for the same entity. Other related work mostly focuses on general misinformation detection and only indirectly addresses outdatedness. Studies in this area use traditional machine learning classifiers like SVMs or decision trees, as well as deep learning models such as LSTMs (Shu et al., 2017). More recent approaches also explore fact-checking using large language models (LLMs). However, these models often treat facts as static, it ignores the reality that information can become outdated over time. Chen et al., (2024) show that the performance of LLM-based fact checkers decreases over time as facts shift or evolve.

To understand the current state of research and define the gap, a literature search was conducted using Google Scholar, ACM digital, HoWeR, and Research Gate. The search included keywords such as: ("outdated news" OR "outdated data" OR "data freshness") AND ("probabilistic model" OR "logistic decay" OR "logistic function" OR "Bernoulli model") AND ("detection" OR "classification") AND (“temporal text analysis”) AND (“fake news”). The search identified 24 relevant sources. To broaden the search, backward citation searching was used to explore the references in the key works and forwards citation searching was applied through Google Scholar, ACM digital, and Research Gate. While earlier studies have addressed the topic of outdatedness, none of them have used a probabilistic approach. Instead, most rely on binary classification. There is also no application of Bernoulli based probabilistic models using logistic decay model to calculate the likelihood of outdatedness over time.

This study fills this gap by suggesting a new method that estimates how likely a news article is to be outdated, based on its age. Using a Bernoulli framework with logistic decay model, this approach contributes to the literature by combining theoretical concepts from probability modeling with practical method for improving the detection of outdated information in digital news content. This way, it helps handle the uncertainty, when facts might become outdated and calculate the chances of it happening.

Commented [HD2]: Why could that generate added-value?

4. Methodology

The chapter aims to explain the used datasets, choices taken during data fine-tuning and logistics decay modelling.

4.1. Data

This section explains the dataset used for this study, including the sources of the data, its pre-processing strategy, and the explorative analysis process.
Since no annotated dataset with outdatedness label is available online, this study creates its annotated dataset by using several approaches that are explained in section 4.2.1.[9]

4.1.1. Data Sourcing

A new dataset is created by scrapping the news articles using an API from a paid service to access and collect the necessary dataset. Mediastack API was the chosen service, due to their affordable cost to access their API. They offer a free tier of their API though with very limited historical articles. Thus, to be able to scrap older news articles, their cheapest tier is chosen.
Mediastack is a REST Api interface collecting news and blog delivering it in JSON format. News articles are from more than 7,500 popular international sources such as CNN (www.cnn.com), The Guardian (www.theguardian.com), and local news sources such as Brisbane Times (www.brisbanetimes.com). Furthermore, they are also available across a variety of categories, including business, sports, education.

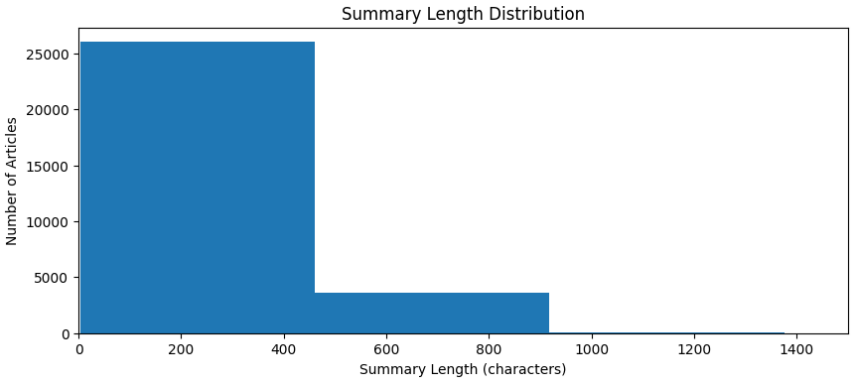
4.1.2. Exploratory Data Analysis (EDA)

Machine is only as accurate as the data it receives, or as people says Garbage in, Garbage out. To avoid feeding 'garbage' into the machine, it's very important to explore the dataset first because it uncovers anomalies and patterns in the dataset. The weakness and the strength of dataset will be showed, which later can be resolved or notified during the preprocess.

[9] AI-assisted phrasing. See AI Directory 8

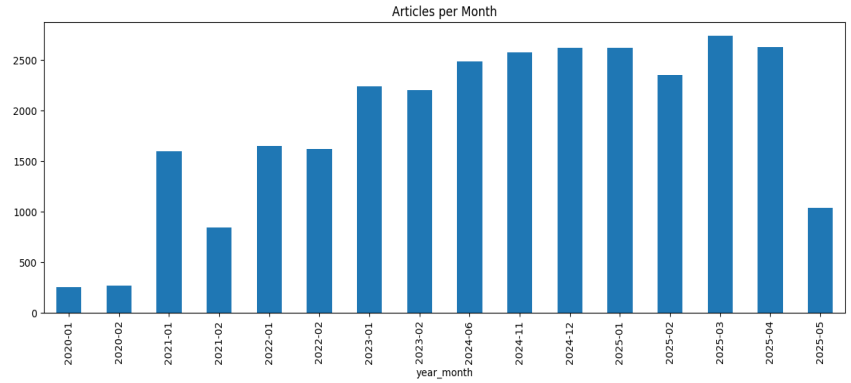
The initial dataset consists of 29734 articles and 12 columns with 761 unique news sources. 95.68% articles are in English and 4.31 % are non-English articles such as French and Indonesian. Each entry contains important information about the article, such as publication date, title, body (full text), summary, category and hyperlinks, however it also includes noisy fields that are not needed such as image_url and time. In this study, instead of using body field as the input text, title + summary are considered more appropriate. It ensures efficiency as these fields typically contains the article's core meaning, which usually enough for binary classification task. While this approach may ignore deeper context in the body, it provides clean and compact article representation of the article's main content.

Figure 3 : Summary Length Distribution



Therefore, Figure 3 shows the distribution of summary length of the dataset not the body length, this examines articles with 0 or short number of characters, those are articles that are scrapped incorrectly or cannot be scrapped due to paywall which will be removed during the preprocess. It indicates that there are unreasonable number of articles with 0 summary length which can be a problem, since it cannot be used for input text.

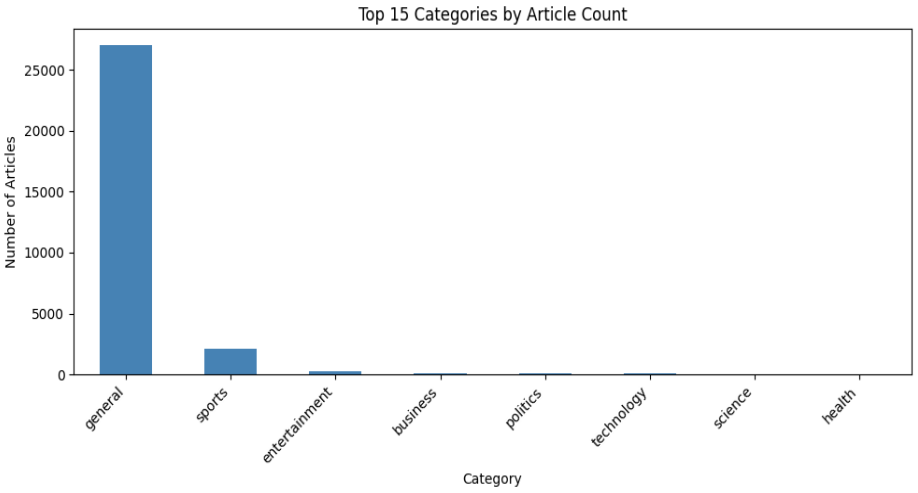
Figure 4: Publication Date Distribution



Next, the distribution of the publication date can also be seen in Figure 4, the publication date range of the articles are between 2020 and 2025. From 2020 to 2023, only months of January and February were scrapped. While in 2024, more months were included to ensure good representation of recent articles and which more likely not be outdated.

Figure 4 also reveals that there are some inconsistencies in data representation across the timeline. Data in the year 2020 show significantly lower article counts, in contrast with the other month and year, this variation in articles count is due to limitation of the Mediastack API.

Figure 5: Category Distribution



Category “General”, can be seen in Figure 5 to dominate the dataset with 90.78%, which is an issue as it makes it highly unbalanced, because it appears more frequently than others. However, since this study is more about is about detecting outdatedness and its temporal awareness, this class imbalance can be ignored. Furthermore, exact duplicates and near duplicates or articles that has the same content or title but published by different sources, can be found in the dataset. Near duplicate is common as news articles are often repeated with minor wording changes for example, when local news agencies would republish the same articles from big agencies like AP in their own site. Additionally, while scanning the dataset, boilerplates phrases (e.g. subscribe to our site) can be found in title, body and summary columns.

Finally, the EDA process shows that the dataset from Mediastack is noisy, containing non-english articles, boilerplates, missing values, unbalanced labels, duplicate entries, and outliers. All these noises could affect the training and thus make it learn incorrect pattern or skew the Logistic Decay parameter. This shows that data preprocess is an important step after the EDA.

4.1.3. Data Preprocessing

Data preprocessing helps clean and transform the dataset into a suitable format for model input (Panagides et al., 2024). In this study, the process is split into two stages: First, data cleaning and second, data enrichment. In the first stage, cleaning of the data includes duplicate, unnecessary fields and outlier removal. Furthermore, to standardize the raw text, a minimal text normalization is also applied, for example removal of irrelevant characters, whitespaces & special symbols. While the second stage is focus on adding important information to the dataset, such as adding fields with values derived from the existent fields.

4.1.3.1. Data Cleaning

To start with the data cleaning, non-english articles are deleted. Afterwards, all the articles with summary length below the 25th percentile or less than 118 characters are removed, this ensures only articles that have empty or summary length shorter than the 25th percentile are not included during the training. Then, fields that are not

needed are deleted so that the dataset is less noisy. These fields include "image_url", "source_type", "time".

Next, the duplicates are removed, this to ensure unique articles are in the dataset. There are two types of duplicates here, the exact duplicate and near duplicate. By creating a temporary normalized versions of the body and summary, which includes converting them to lowercase, stripping whitespace, and removing formatting inconsistencies. Then, the normalized version of body + summary + date is combined as the input, it identifies articles that are exact duplicates or effectively the same news item posted more than once. Here, only the first occurrence was retained. Furthermore, according to Wu et al. (2010), Term Frequency-Inverse Document Frequency (TF-IDF) is widely used in text-mining for representing documents in similarity-based tasks. Since news articles can be considered as document in this case, thus TF-IDF vectorization and cosine similarity is used here to detect both near-duplicate and exact-duplicate. After the body and summary is normalized, it transformed the text into weighted vectors. TF-IDF assigned higher importance to words that frequent in an article but rare across the entire dataset. Then cosine similarity was generated between all pairs of vectors. The similarity score ranges from 0 to 1, where values closer to 1 indicate higher similarity. A pair of articles was annotated as near-duplicates if their cosine similarity exceeded a threshold of 0.99, indicating extremely high textual similarity. 0.99 is chosen because only articles that are near exact duplicates need to be removed, while it is okay to have articles with some similarity for example when the context of the articles is the same, but the article text is rephrased when published in different news sources. From each group of near duplicates, only the first occurrence was retained, and the others were removed from the dataset. This procedure helped ensure that the final dataset was free of overtly redundant articles while preserving content diversity, which is essential for robust modeling in the later stages of analysis.

Lastly, the dataset removed boilerplates phrasing such as “image 4 of 4”, HTML, Unicode Artifacts, Line Breaks. However, the removal of the boilerplates was not optimal because there are phrases left from the scrapping in the dataset that couldn’t be automatically removed, thus this becomes the limitation of this dataset.

4.1.3.2. Data Enrichment

Once data is cleaned, new fields that are derived from other fields to support semantic enrichment are created. First, a minimal text normalization is done to the title and summary of each entry, including removal of irrelevant characters (e.g. emojis or symbols), but it preserves the standard punctuation marks. To ensure consistent spacing in the summary, whitespaces was normalized. No lower spacing is done here as case-sensitivity is important in news articles to differentiate between apple the fruit and Apple the company. The original title and summary were not replaced by the new normalized results, two new fields called “title_normalized” and “summary_normalized” were created. This approach ensures that both the original and cleaned versions are retained in the dataset. Body texts were not normalized as it won’t be used during the model training or probability modelling due to its higher noise level. Thus, from now on “title_normalized” and “summary_normalized” together are referred as input text in this study.

Afterwards, “t” column is created, it is the age of article and used to determine time difference (in days) between the article's publication date and a reference date, which is 09.02.2025. By using a fixed date rather than calculating the t based on the actual date of ‘today’, it simplifies training because dataset doesn't need to be scrapped and preprocessed each time. 09.02.2025 was chosen as the reference date, after systematic evaluation of all possible reference dates in the dataset were performed. The goal was to identify the date that maximizes the number of articles for bin with the lowest articles count in “t_bin” field which is fresh bin. There are five bins that article can be put into based on their t values: 0-9 “fresh”, 10-29, “recent”, 30-59, “mid-age”, 60-89, “old”, >=90, “very old”. From all possible cases, using 09.02.2025 would produce the maximum number of articles in fresh bin which is 580 articles.

4.1.4. Data Splitting

After the preprocess, there were 12477 articles in total. As this study involves fine-tuning BERT for text classification and fitting a probability decay function, two separate subset of the data is needed. One subset is used to train and fine tune BERT, and the other is reserved for fitting the temporal decay function. Furthermore, to ensure a strong result in BERT fine-tuning a balanced dataset is needed. Due to

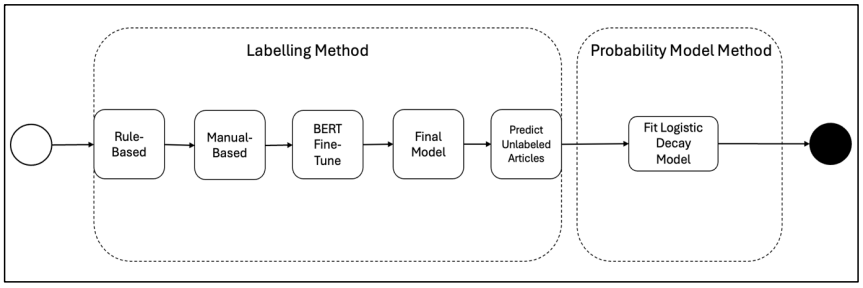
temporal feature being an important feature, it was important to represent the full distribution of t values in the training data, stratified sampling was applied instead of random sampling to achieve balanced and well-represented t distribution. However, since raw “t” is a continuous date variable and too widely spread, “t_bin” is used as the base for a balanced dataset. As mentioned in the 4.1.3.2, the lowest articles count is in the fresh bin with 580 articles, this means to have a balance sampling, each bin can only have max of 580 articles, which in total, there are 2900 articles in the stratified dataset. The first objective is to fill each bin with all the t unique values in order to meet the quota of 580 articles, if this is not feasible, then random t values are used.

While the remaining articles that were not chosen for training become the dataset to fit into the logistic decay, with 9577 articles.

4.2. Modelling Methodology

To answer the research question, the following methods are done. First, since outdatedness labels in the dataset are not available, a semi-automated labelling approach is developed to generate labels for seed dataset. After fine-tuning BERT with the training data and the best model are created, the model is used to classify label in the remaining unannotated dataset. Lastly, the annotated probability dataset is fitted into a logistic decay model.

Figure 4: Workflow (own illustration)



4.2.1. Data Annotation Methodology for Outdatedness Detection

Semi-automated methods are applied to classify news articles as relevance (1) or outdated (0). The first method applied is a is rule-based, and the second method is manual labelling. The annotated dataset generated are used as the training set for fine-tuning BERT.

4.2.1.1. Training set

To fine-tune BERT, a training set is needed to feed it into the BERT model. Two methods are chosen to help label the training set Rule-Based and Manual-Based.

4.2.1.1.1. Rule-Based

A Rule-Based labelling was applied to the stratified dataset, the goal was to create as much high-confidence labels as possible. Thus, this rule-based approach prioritized precision over coverage to ensure that only articles with strong outdatedness signals were annotated. This approach also reduced the manual effort required in the later hand-labeling stage.

The labelling logic included a mixed of temporal- and content-based features in the article. Here, the fields used are title_normalized or summary_normalized instead of raw title and summary.

The structure of the logic is as follows:

- The label articles titles were analysed using spaCy's English language model (en_core_web_sm) to detect part-of-speech tags and filter out grammatically ambiguous entries.
- Titles were examined for verb tense to detect ongoing, past, or future-oriented phrasing based on POS tags. This allowed temporal intent in the title to be inferred.
- A set of conservative rules was applied:
    - Articles with $t \leq 3$ days were annotated as relevance or 1.
    - Titles with ongoing phrasing and $t \leq 10$ days were also annotated 1.
    - Titles with future-oriented phrasing and $t > 180$ days were annotated as outdated or 0.

- If the article summary contained the word *"today"* and t > 1, the article was annotated 0.
- Articles with past event keywords in title and t > 30, it is also annotated as 0
- All other cases were left unannotated for manual review.

- Each label was stored along with a “label_comment” field that documented which rule was applied or whether the article was skipped due to ambiguity.

A total of 695 articles are annotated, where 322 articles were annotated 1, and 373 articles were annotated 0. While the rest is left unannotated due to ambiguity that the title has, including title that has no verb and title with less than 5 words.

4.2.1.1.1.1. Rule-Based Validation

To evaluate the quality of the rule-based labels, a validation set was created by the label attribute. 60 articles from each class (label = 0 and 1) from the stratified dataset. This ensured a fair and interpretable assessment across both class of outdated and non-outdated content. Once validated by a human, a classification report is used to compare the rule-based label values against human-reviewed “reviewed_label” values. The report includes precision, recall, and F1 scores for each class (outdated vs. current). The result can be seen in Table 1.

Table 1: Classification report of rule-based annotation

	Precision	Recall	F1-score	Support
0	1.00	0.98	0.99	61.00
1	0.98	1.00	0.99	59.00
Accuracy			0.99	0.99
Macro avg	0.99	0.99	0.99	120.00
Weighted avg	0.99	0.99	0.99	120.00

The model’s performance is evaluated using standard metrics commonly used in information retrieval and natural language processing: precision, recall, and most importantly, the F1 score. These metrics are based on the following definitions:

- True Positives (TP): Instances correctly labeled as positive by the model

- False Positives (FP): Instances incorrectly labeled as positive when they are actually negative.
- True Negatives (TN): Instances correctly labeled as negative.
- False Negatives (FN): Instances that are actually positive but were missed by the model.

From these values, the calculation of the metric is a s follows:

- Precision = = $\frac{TP}{TP+FP}$ How many predicted positives are correct
- Recall = $\frac{TP}{TP+FN}$ How many actual positives are correctly predicted
- F1 Score = $2 \times \frac{Precision \times Recall}{Precision + Recall}$ Harmonic mean of Precision & Recall
- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ How many correctly predicted instances predictions across all classifications.

Generally, the F1 score is particularly important, as it helps assess how well the rule-based approach balances between identifying correct positives and avoiding incorrect ones.

Here, the rule-based labelling approach achieved high agreement with human-reviewed labels, with an overall accuracy of 99.9%. Both classes, “relevant” and “outdated”, were classified with a high F1 scores of 99.9%, this means the model has a robust rule performance. Precision and recall values were both above 98% for both classes, showing the effectiveness of the rule-based strategy as a high-confidence pre-labelling method. This strong performance justifies its use as a foundation for bootstrapping further machine learning models and active learning strategies. [10]

4.2.1.1.2. Manual-based

The remaining 2205 articles with missing annotated were annotated by the author manually. The parameters to label the articles are based on the news relevance and outdatedness definition in section **2.1** and the structure in the rule-based logic in section **4.2.1.1.1** in case it got skipped accidentally. Similarly, with Rule-Based

[10] AI assisted phrases. See AI directory 24

labelling method, the fields used to extract the article context are from title_normalized or summary_normalized.

The article is label 1 if any the following criteria is fulfilled by the article:

- If the event mentioned in the article are still relevant or has not changed to the reference date, below are the examples:
    - Discovery of new scientific or health
    - New regulations from government
    - Tips

And it is label 0, when the article had any of the following characteristics:

- If t field is more than 365 days, and the article belongs to sport, business, or politics category.
- If in the input text, it refers to people, events or organizations that no longer relevant as of reference date (e.g. President Biden, Covid, Pandemic)
- If temporal expressions such as this week, this month, next month etc. are mentioned but the context of the event in the article is no longer relevant to reference date.

4.2.1.1.2.1. Inter-annotator Agreement

Manual-based labels were evaluated using Inter-Annotator Agreement (IAA). For this, a second annotator was involved and provided with definition of news outdatedness from section **2.1** and the structure in the rule-based and manual-based logic in section **4.2.1.1.1** and **4.2.1.1.1.1** respectively. The second annotator task is to label the same dataset that first annotator already did. After the annotator completed the labeling, IAA was calculated to assess the level of agreement between them (Grasso et al., 2024). Having high-degree of IAA helps show that the metric is reliable and fair (Lommel et al., 2014).

Deleger et al. (2012) explained that Cohen's Kappa is one of the more commonly used technique to measure IAA. It is calculated based on two components: the observed agreement (Ao) and the agreement expected by chance (Ae). The observed agreement (Ao) refers to the count of instances where both annotators assigned the same label, calculated by dividing the number of agreements by the total number of instances. The expected agreement (Ae) estimates how often the annotators would agree purely by chance. This is done by multiplying the probabilities of each annotator assigning an instance to a particular category, then

summing those products across all categories. These probabilities are estimated from the observed distribution of labels assigned by each annotator[11] . The formula for Cohen's Kappa is as follows:

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

The manual-based classification task resulted in a score of 0.700, which is as Grasso et al. (2024) put it in accordance with the a binary task due to its simpleness which is to only classify two class, which in this case is 1 and 0. Higher score indicates higher agreement, this means 0.700 is a high score for IAA. In conclusion, based on the score both annotators have quite a high agreement which makes the labels reliable and thus acceptable for BERT fine-tuning.

4.2.1.1.3. Data Augmentation

The annotated training set contained less than 31% of label 1 (relevant) articles. To ensure a balanced dataset for BERT fine-tuning, data augmentation was applied only to the label 1 class. Additionally, Ivarsson (2019) mentioned in her study that data augmentation is a common regularization method, which help in avoiding overfitting. Thus, it's important to augment the data.

Since the meaning of the article contents is important, back-translation or BT is chosen as the data augmentation method. The method is done by translating the text to another language, which in this case German and translate it back to English. 498 synthetic articles were generated by using BT, creating paraphrased versions of the same content which enrich the dataset while keeping core meaning of the article.

4.2.2. BERT Fine Tuning

The model used in this study is bert-base-cased, a pretrained transformer-based language model provided by Hugging Face. Bert-base-cased is a case sensitive model, which means it can detect the difference between apple (fruit) or Apple (company). This is important because news content have case-sensitive distinctions, for example "BREAKING: New Policy Announced", the capitalized "breaking" is a time expression which could help in classifying outdatedness.

Early stopping is applied during the training given the small training dataset. This helps reduced not only overfitting but also computation resources. The early

[11] AI-assisted phrasing. See AI Directory 19

stopping is based on the main performance metric, which is the macro F1 score. If the metric is not improving over 2 epochs, then the training will stop automatically. After stopping, the checkpoint which resulted the best results on the metric is loaded and used for evaluation on the test set.

During fine-tuning, the model was adapted to the binary classification task using the Hugging Face Trainer API. Each input text consisted of a text string that combined a relative time statement from column t (e.g., “This article was published 25 days ago”) with the normalized title and summary of the article, joined by a special [TIME] token. This design allowed the model to incorporate both temporal and content information effectively as done by Han et al. (2025) in their study about Time-Specifier Model Merging (TSM). They fine-tune separate models on query subsets containing explicit time specifiers.

Furthermore, to increase training efficiency and speed up fine-tuning process, tokenization was performed using the BertTokenizer with a maximum length of 160 tokens. This limit was chosen because after combining the time statement, title and summary into one input text, 95th percentile of the input text contains sequences of fewer than 160 tokens, which is enough to capture important content and temporal expressions that is usually found at the beginning of the summary text. This approach minimizes the loss of context while still retaining the essential information.

The model was trained in Python 3.12.0 and the PyTorch framework through the Hugging Face’s transformers.Trainer API. Finetuning was performed in Google Colab on a Tesla T4 GPU with 16GB of memory.

4.2.2.1. Hyperparameter Tuning

To maximize the model performance, hyperparameter tuning was conducted to find the best settings. The goal was to find the best combination of training parameters that enhance the model’s performance, specifically, the macro F1-score on the validation set.

The hyperparameter tuning tool chosen is Optuna. It is a next generation optimization frameworks because it can dynamically construct the search space, availability of user-customization and its easy-to-setup and versatile architecture (Akiba et al., 2019).

Below are the hyperparameters that were tuned.

- Learning Rate: Determines the step size during gradient descent. A log-uniform range from 1e-6 to 5e-5 was used; this allows more flexibility in exploring optimal convergence.
- Batch Size: Controls how many samples are processed before the model's internal parameters are updated. Values of 16 and 32 were tested, as used in the original BERT paper.
- Weight Decay: A regularization term to prevent overfitting by penalizing large weights. This was sampled from a uniform range between 0.0 and 0.3. While not included in the original BERT fine-tuning, weight decay has been shown in later research to stabilize training.
- Warmup Ratio: Specifies the proportion of training steps during which the learning rate increases linearly from zero to the initial rate. A range of 0.0 to 0.3 was tested.

All models were trained for up to 5 epochs with early stopping to prevent overfitting. The best hyperparameter configuration was selected based on the highest F1-score on the validation set.

Table 2: Summary of hyperparameter tuning

Trial	F1	Learning Rate	Batch Size	Weight	Warm up Ratio
0	**91.861**	**4.33E-06**	**16**	**0.180**	**0.047**
1	54.830	1.84E-06	32	0.180	0.212
2	54.616	1.08E-06	16	0.064	0.055
3	56.386	2.05E-06	32	0.130	0.087
4	**91.861**	**1.10E-05**	**32**	**0.110**	**0.137**
5	90.638	2.16E-05	32	0.178	0.014
• **6**	**92.765**	**1.08E-05**	**16**	**0.285**	**0.290**
7	91.252	2.36E-05	16	0.205	0.132
8	58.087	1.61E-06	16	0.273	0.078
9	90.961	1.34E-05	32	0.164	0.928

Note. Performance reported on validation set. The reported result is the maximum F1 score reached by the model. Bold indicates highest performance and trial with this symbol + , is the chosen hyperparameters.

Table 2 shows the combinations of hyperparameters tried during tuning. Trials 1,2,3,8 have lower learning rates, which means the model learn slow and thus underfit and have low F1 score. Runs with mid to higher learning rate have achieved

generally better outcomes. However, when it is too high, it could also cause overshooting. Thus, here mid-range is more optimal. Smaller batch of 16 appears in top 3 trials, having lower helps to better generalizes. Given the result from hyperparameter tuning, the final model was trained with a batch size of 16, a learning rate of 1.08E-05, weight decay of 0.285 and warm up ratio of 0.290.

4.2.3. Probabilistic Model Development

Once classification model from finetuning BERT were created, it needs to predict the remaining unannotated data so that it can be used to fit into the logistic decay function. In order to compare the performance of the logistic function, the dataset is also fit into exponential function.

4.2.3.1. Logistic Decay Fitting

Before fitting the annotated dataset to the logistic function, the remaining unannotated dataset, as described in Section **4.1.4**, must first be annotated using the best-performing model from the BERT fine-tuning stage. However, this unannotated dataset, which is reserved for training the probability model, lacks articles in the "fresh" time bin. This is because those articles were used in the training set to improve the performance of the fine-tuned BERT model. During training, the dataset was intentionally stratified and split into separate training, validation, and test sets. To address the missing representation of recent articles, the test set was intentionally merged back into the unannotated pool to enrich the probability dataset, especially since article decay tends to occur shortly after publication rather than much later.

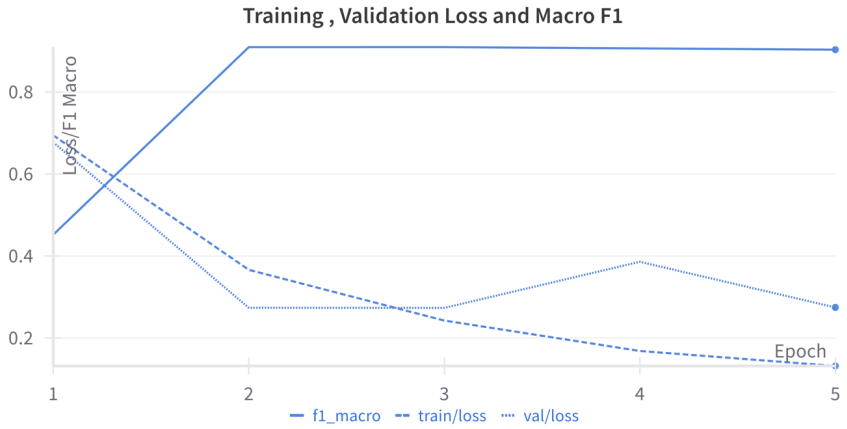
Once all articles in the dataset were annotated, the data was used to fit the logistic decay function. In addition to logistic decay, an exponential decay function was also applied to the predicted relevance values over time (t). This allowed for a comparative evaluation of model performance by assessing which function better fit the predicted data.[12]

[12] AI Assisted Phrases. See AI Directory 22

4.3. BERT Fine-tuned Classification Model

During the fine-tuning phase, the training and validation loss were monitored to assess model stability and generalization. The training loss consistently decreased across epochs, indicating that the model was learning from the training data. However, the validation loss reached its minimum at epoch 2 and began to increase afterward. This divergence between training and validation loss suggests that the model started to overfit the training data beyond epoch 2. Although the accuracy and F1 scores remained high in subsequent epochs, the rising validation loss reflects decreased generalization to unseen data. Such behaviour is a sign of training instability, particularly common in small datasets when epoch is short (Mosbach et al., 2020b). To mitigate overfitting and maintain a balance between learning and generalization, model from epoch 2 were selected as the final checkpoint, as it provided the most stable and reliable performance on the validation set.

Figure 5 : Final Model training and validation loss and performance



After selecting the best model, predictions were run on the held-out test set, which had not been used during training or validation. The metrics used are the standard metrics such as F1 score and Accuracy. The following results were obtained:

Table 3: Performance summary table on test set

	Accuracy	Precision	Recall	F1
Model with Best Hyperparam	90.36	90.84	90.36	90.33

The model managed to achieve a score of 90.33 F1 score. These scores demonstrate the model's ability to classify articles as outdated or relevant, with generalization validated through the test set. The classification model successfully predicted article outdatedness using a fine-tuned BERT architecture. By selecting the best model at epoch 2 and validating on a separate test set, the evaluation process provided a reliable measure of model performance. These predictions were later used for decay modeling in the next stage of the thesis.

4.4. Probability Model

The evaluation of both Exponential and Logistic Decay functions was performed using several statistical and classification metrics: Logloss, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), MSE (Mean Squared Error), and AUC (Area Under the ROC Curve).

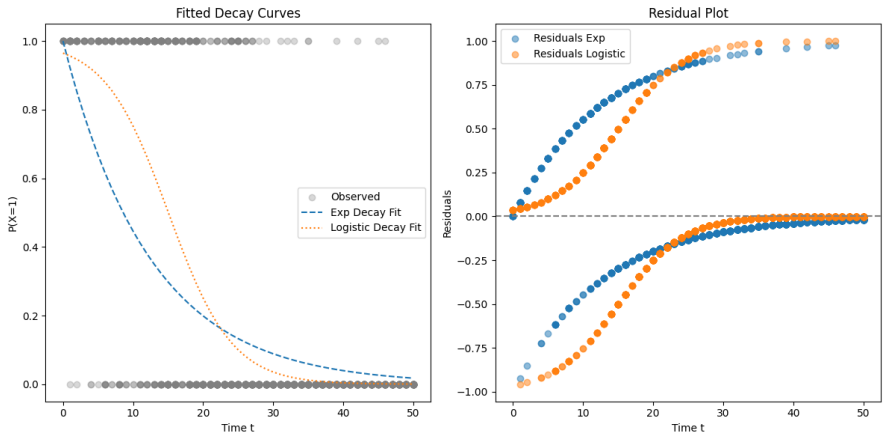
Table 4: Exponential vs Logistic Performance

Metric	Exponential	Logistic
Logloss	0.2767	0.2366
AIC	962.5	823.44
BIC	973.41	834.35
MSE	0.854	0.07
AUC	0.9372	0.9372

- Log Loss: measures predictive probability quality (lower is better).
- AUC (Area Under the ROC Curve): measures the model’s ability to distinguish between outdated and non-outdated.
- Mean Squared Error (MSE): quantifies the average squared difference between predicted and actual labels.
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): penalize model complexity and help compare model fit.

This multi-metric evaluation allows us to compare which function (exponential or logistic) better captures the decay trend in the data.

Figure 6 : Decay and Residual Plot



In the beginning, there is around 9000 articles reserved for the purpose of fitting them into the logistic functions. However, the distribution of t values in the dataset is too wide, which affecting the plotting and making it hard to see the shape of the line. Thus, to be able to see the shaped better, the dataset fitted here is limited to t max 50 days.

The decay rate of the model 0.2134, it measures how sharply news relevance drops over time. Here, the values imply a moderate decay speed (not to sharp, not too gradual). While for t_0 or the tipping point where relevant decay accelerates rapidly occurs around the point 14.94 days or 15 days to round it up. It can be concluded that the life cycle of news article is around 2 weeks, and that it starts to lose it traction after 15 days.

To see how well the dataset fit into the logistic functions. Two plots were produced as seen in Figure 6:

- The Fitted Decay Curves plot shows the logistic decay curve aligning more closely with the observed data points than the exponential curve, especially in the mid-range of t, where the decay happens gradually.

- The Residual Plot reveals that the logistic model produces residuals much closer to zero across time. In contrast, the exponential model shows systematic underestimation in the early t period and overestimation later, indicating poorer fit and bias.

Based on the results, the Logistic Decay Model demonstrates superior performance across almost all evaluation metrics. It provides better fit (lower AIC/BIC), more accurate predictions (lower MSE and Logloss), and a well-calibrated decay pattern that matches the observed probabilities more closely. Therefore, the logistic decay function is selected as the final model to represent the temporal decay of news article relevance in this study.

5. Conclusion

As shown in the evaluation section, two models were created. One is the classification model by finetuning BERT and the other is the probability model by fitting the annotated dataset to logistic decay function. This answers the main research question of this research, which was ‘How can a probabilistic model to predict the outdatedness of a news data using logistic decay model be developed?’. Generating annotated dataset is the first step to achieve it. Annotating small portion of the unannotated dataset using a mix rule-based and manual-based, showed to be a great approach in creating the final BERT fine-tuned model. The model achieved a 90.3 in F1 score on its task to classify outdatedness. It allows further annotation for the remaining unannotated dataset. Furthermore, the logistic decay model proved to be a good fit for modeling the relevance decline over time. Compared to an exponential model, the logistic function showed promising results, which is a better fit to the data by having lower logloss, MSE, AIC, and BIC score, and plotting a more realistic modeling of gradual decline. The logistic model’s S-shaped curve captures the hypothesis that articles remain relevant for a short period before experiencing a sharper drop-off in relevance, after which they stabilize into long-term irrelevance. The parameters λ and t0 provide insight into how quickly relevance fades and when an article typically becomes outdated. These insights can support applications such as news archiving, content recommendation, and information retrieval, where understanding the lifespan of information is essential.

In summary, the approach provides a good and explainable way to quantify news outdatedness over time. It lays a foundation for future research that could extend the model with richer temporal signals (e.g., event dates), dynamic decay patterns, content topics/category or user engagement data to capture relevance from a wider contextual perspective.[13]

5.1. Limitation and Further Research

This study has several limitations that affect how broadly the findings can be applied and how deeply they can be interpreted.

First, the dataset is limited and unevenly distributed over time. Although the initial dataset included almost 30,000 news articles, only about 3,000 were usable for training after data cleaning, stratification, and label filtering. This significantly reduced the number of articles available across different time intervals (t values). In particular, there were not enough newer articles (small t values or more recent articles) to properly capture how relevance begins to decline shortly after publication. This is especially important given that several studies report that news article decay can begin within 36 hours, highlighting the need for finer granularity in *t* values, potentially even tracking time by the hour.

Second, there is a strong class imbalance in news categories. Each article in the dataset includes a category label (such as business, politics, or technology), but about 90% of the data falls into the "general" category. This heavy imbalance limits the ability to compare how outdatedness differs between topics or domains. As a result, category-based analysis was excluded from this study.

Third, the dataset lacks true event dates. The study used the article's publication date as a main feature for the time expression during fine-tuning. This is because it’s not easy to extract the true event time of a news articles, even though it represents of the articles better. Furthermore, having true event time would help in outdatedness classification better as we can easily derived outdatedness with it.

Lastly, the annotation process did not include factual verification. The study is mainly dependent on the annotator knowledge and understanding to classify the article outdatedness. While time is a quite a straightforward metric that could be quantified

[13] AI Assisted Phrase. See AI Directory 2

easily, detection of obsolete fact is not as simple. Thus, having a fact checker to filter out the old content with old fact, would make the annotation more robust faster.

To improve on these limitations and further explore how news becomes outdated over time, future research could consider the following directions:
First, by using Time-Aware Language Models rather than just base BERT. Models like TimeBERT or other temporally sensitive versions of BERT could help in capturing how time influences the meaning and relevance of content better.
Next, including Knowledge Graphs to help with outdatedness classification. This can be done by linking articles to external databases such as DBpedia or Wikidata. It can help check whether content, entities and events are still current, improving the accuracy of labels and understanding of outdatedness.
Another suggestion is to find dataset that include user engagement data. Metrics like article views, social media shares can be used as real-world signals of what people still find relevant, which could improve how outdatedness is measured. Multi-label classification is also a good future research recommendation, so instead of using binary classification and simply annotating articles as "outdated" or "current," future models could use multiple classification or relevance scores to reflect different levels of outdatedness for example as fresh, slightly outdated, or fully outdated.
Additionally, having a more balanced and richer dataset better would help in fine-tuning and fitting into the logistic decay better. This means finding and collecting more articles from under-represented time periods and categories would lead to better model performance.
Lastly, inclusion of tags for evergreen content, which is those that remain relevant over a longer period. This could be an interesting topic to see how differ time-sensitive vs evergreen content decay.[14]

[14] AI assisted phrasing. See AI directory 23

References

Abilio, R., & Coelho, G. P. (2024). Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset. *Applied Soft Computing*, *166*, 112158. https://doi.org/10.1016/j.asoc.2024.112158

Akhtar, Z., Arshad, A., Zubair, M., Qasim, A., Ullah, F., Zamir, M. T., Ahmad, M., Sidorov, G., & Gelbukh, A. (2024). NAMED ENTITY RECOGNITION TOOLS AND TECHNIQUES IN CUSTOM NATURAL LANGUAGE PROCESSING MODELS. *Journal of Population Therapeutics and Clinical Pharmacology*, *31*(10), Article 10. https://doi.org/10.53555/e8b2km73

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2623–2631. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3292500.3330701

Alam, T., Khan, A., & Alam, F. (2020). Bangla Text Classification using Transformers. *ArXiv*. https://www.semanticscholar.org/paper/1c04c17b0f6545b832d05fd6a09b9eb2d49ce0da

Aliwy, A. H., & Ameer, E. H. A. (2017). *Comparative Study of Five Text Classification Algorithms with their Improvements*. *12*(14).

Almquist, A., & Jatowt, A. (2019). *Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences* (pp. 86–101). https://doi.org/10.1007/978-3-030-15712-8_6

Ayorinde, F., Ajeleke, G., & Samuel, A. (2025). *An Overview of Natural Language Processing*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization* (No. arXiv:1607.06450). arXiv. https://doi.org/10.48550/arXiv.1607.06450

Barkemeyer, R., Faugère, C., Gergaud, O., & Preuss, L. (2020). Media attention to large-scale corporate scandals: Hype and boredom in the age of social media. *Journal of Business Research*, *109*, 385–398. https://doi.org/10.1016/j.jbusres.2019.12.011

Bouzeghoub, M. (2004). A framework for analysis of data freshness. *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*, 59–67. https://doi.org/10.1145/1012453.1012464

Chan, L., Ford, P., & Ogunrinde, V. (2025). *Named Entity Recognition (NER) with spaCy and Transformers*.

Chavan, T., & Patil, S. (2024). NAMED ENTITY RECOGNITION (NER) FOR NEWS ARTICLES. *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY*, *2*(1). https://doi.org/10.34218/ijaird.2.1.2024.10

Chen, M., Wei, L., Cao, H., Zhou, W., & Hu, S. (2024). *Explore the Potential of LLMs in Misinformation Detection: An Empirical Study* (No. arXiv:2311.12699). arXiv. https://doi.org/10.48550/arXiv.2311.12699

Chun, A., Hsu, E., & Nguyen, E. (2023). *A Comprehensive Analysis of Fine-Tuning Strategies for BERT*.

Dahl, V. (2010). An Introduction to Natural Language Processing: The Main Problems. *Triangle*, *1*, Article 1. https://doi.org/10.17345/triangle1.65-78

Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K., & Solti, I. (2012). Building gold standard corpora for medical

natural language processing tasks. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2012*, 144–153.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Digital News Report 2023 | Reuters Institute for the Study of Journalism. (n.d.). Retrieved March 24, 2025, from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023

Ferranti, N., Krickl, A., & Nissl, M. (2021). *Knowledge Graphs: Detection of Outdated News*.

Finger, M., & Da Silva, F. S. (1998). *Temporal data obsolescence: Modelling problems*. 45–50. Proceedings. Fifth International Workshop on Temporal Representation and Reasoning (Cat. No.98EX157). https://doi.org/10.1109/time.1998.674130

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: A review. *Artificial Intelligence Review*, *58*(6), 166. https://doi.org/10.1007/s10462-025-11162-5

González-Carvajal, S., & Garrido-Merchán, E. C. (2023). Comparing BERT against traditional machine learning text classification. *Journal of Computational and Cognitive Engineering*, *2*(4), 352–356. https://doi.org/10.47852/bonviewJCCE3202838

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org

Grasso, F., Locci, S., Siragusa, G., & Di Caro, L. (2024). EcoVerse: An Annotated Twitter Dataset for Eco-Relevance Classification, Environmental Impact Analysis, and Stance Detection. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 5461–5472). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.485/

Gupta, A., & Choubisa, M. (2024). Natural Language Processing (NLP): Enabling Machines to Understand and Process. *International Journal of Food and Nutritional Sciences*, *09*(03). https://doi.org/10.48047/ijfans/09/03/29

Han, S., Hwang, T., Cho, S., Jeong, S., Song, H., Lee, H., & Park, J. C. (2025). *Temporal Information Retrieval via Time-Specifier Model Merging* (No. arXiv:2507.06782; Version 1). arXiv. https://doi.org/10.48550/arXiv.2507.06782

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (No. arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Ivarsson. (2019). *Racing Bib Number Recognition Using Neural Networks*. HWR Berlin.

Jatowt, A., Sato, M., Draxl, S., Duan, Y., Campos, R., & Yoshikawa, M. (2024). Is this news article still relevant? Ranking by contemporary relevance in archival search. *International Journal on Digital Libraries*, *25*(2), 197–216. https://doi.org/10.1007/s00799-023-00377-y

Koroteev, M. V. (2021). *BERT: A Review of Applications in Natural Language Processing and Understanding* (No. arXiv:2103.11943). arXiv. https://doi.org/10.48550/arXiv.2103.11943

Lommel, A., Popović, M., & Burchardt, A. (2014). *Assessing Inter-Annotator Agreement for Translation Error Annotation*.

Long, J., Marshall, A., & Feng, Z. (2020). *Probability Models*. 33–39. https://doi.org/10.1016/B978-0-08-102295-5.10445-7

Montesinos-López, O., Montesinos, A., & Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. https://doi.org/10.1007/978-3-030-89010-0

Mosbach, M., Andriushchenko, M., & Klakow, D. (2020a). On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *ArXiv*. https://www.semanticscholar.org/paper/8b9d77d5e52a70af37451d3db3d32781b83ea054

Mosbach, M., Andriushchenko, M., & Klakow, D. (2020b). On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *ArXiv*. https://www.semanticscholar.org/paper/8b9d77d5e52a70af37451d3db3d32781b83ea054

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Ocaña, M. G., Opdahl, A. L., & Dang-Nguyen, D.-T. (2021). *Emerging News task: Detecting emerging events from social media and news feeds*.

Pakhale, K. (2023). *Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges* (No. arXiv:2309.14084). arXiv. https://doi.org/10.48550/arXiv.2309.14084

Panagides, R. K., Fu, S. H., Jung, S. H., Singh, A., Eluvathingal Muttikkal, R. T., Broad, R. M., Meakem, T. D., & Hamilton, R. A. (2024). Enhancing Literature Review Efficiency: A Case Study on Using Fine-Tuned BERT for Classifying Focused Ultrasound-Related Articles. *AI*, *5*(3), Article 3. https://doi.org/10.3390/ai5030081

Peterson, C. J., Anderson, C., & Nugent, K. (2022). Continued Visibility of COVID-19 Article Removals. *Southern Medical Journal*, *115*(6), 371–373. https://doi.org/10.14423/SMJ.0000000000001397

Pham, T. V. A. (2024, January 24). *Semi-Automated Enhancement of Knowledge Graphs with Large Language Models*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Rahmi, N. A., Defit, S., & Okfalisa, -. (2024). The Use of Hyperparameter Tuning in Model Classification: A Scientific Work Area Identification. *JOIV : International Journal on Informatics Visualization*, *8*(4), 2181. https://doi.org/10.62527/joiv.8.4.3092

Sainani, K. L. (2014). Logistic Regression. *PM&R*, *6*(12), 1157–1162. https://doi.org/10.1016/j.pmrj.2014.10.006

Salah, R., Mukred, M., Zakaria, L. Q., & Al-Yarimi, F. (2024). A Machine Learning Approach for Named Entity Recognition in Classical Arabic Natural Language Processing. *KSII Transactions on Internet and Information Systems*, *18*, 2895–2919. https://doi.org/10.3837/tiis.2024.10.005

Salih, M. I., Mohammed, S. M., Ibrahim, A. K., Ahmed, O. M., & Haji, L. M. (2025). Fine-Tuning BERT for Automated News Classification. *Engineering,*

Technology & Applied Science Research, *15*(3), Article 3. https://doi.org/10.48084/etasr.10625

Satheesh, D. K., Jahnavi, A., Iswarya, L., Ayesha, K., Bhanusekhar, G., & Hanisha, K. (2020). *Resume Ranking based on Job Description using SpaCy NER model*. *07*(05).

Sebastiani, F. (2005). *Text Categorization*. 683–687. https://doi.org/10.4018/978-1-59140-560-3.ch112

Segev, A., & Weiping, F. (1990). *Currency-Based Updates to Distributed Materialized Views.* IEEE Computer Society.

Sharma, B. A., Owens, G., Asthana, G., & Mishra, P. (2022, November 5). *Time Series Forecasting of Southern Hemisphere's Sea Ice Extent Using the Logistic Model by Bhavna Ajit Sharma, Gavriel Owens, Gargi Asthana, Prakhar Mishra: SSRN*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4103583

Shekhar, S., Bansode, A., & Salim, A. (2022). *A Comparative study of Hyper-Parameter Optimization Tools* (No. arXiv:2201.06433). arXiv. https://doi.org/10.48550/arXiv.2201.06433

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake News Detection on Social Media: A Data Mining Perspective* (No. arXiv:1708.01967). arXiv. https://doi.org/10.48550/arXiv.1708.01967

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons, and Fractals*, *139*, 110058. https://doi.org/10.1016/j.chaos.2020.110058

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Werner, S. V. (2023). *Extracting Political Relations from News Articles Using Transformers*. HWR Berlin.

Wong, J., Manderson, T., Abrahamowicz, M., Buckeridge, D. L., & Tamblyn, R. (2019). Can Hyperparameter Tuning Improve the Performance of a Super Learner?: A Case Study. *Epidemiology*, *30*(4), 521. https://doi.org/10.1097/EDE.0000000000001027

Wu, X., Ide, I., & Satoh, S. (2010). PageRank with Text Similarity and Video Near-Duplicate Constraints for News Story Re-ranking. In *Advances in Multimedia Modeling* (p. 544). https://doi.org/10.1007/978-3-642-11301-7_53

Zaman-Khan, H., Naeem, M., Guarasci, R., Bint-Khalid, U., Esposito, M., & Gargiulo, F. (2024). Enhancing Text Classification Using BERT: A Transfer Learning Approach. *Computación y Sistemas*, *28*(4), Article 4. https://doi.org/10.13053/cys-28-4-5290

Appendices

Appendix A – AI Directory

No.	AI Tool	Purpose of Use	Affected Thesis Sections	Remarks/Prompt
1	ChatGPT	Check paragraph for grammar and clarity	Abstract	Improve grammar and clarity
2	ChatGPT	Check paragraph for grammar and clarity	Conclusion	Polish text for formality
3	ChatGPT	Help in making two paragraph cohesive	Introduction	Enhance text cohesion and flow
4	Deepseek	Help in making text more cohesive	Theoretical Foundations	Improve grammar and clarity
5	Deepseek	Help in making text more cohesive	Theoretical Foundations	Improve grammar and clarity
6	ChatGPT	Help to be more organized	Coding	Suggest improvements
7	ChatGPT	Help in creating script to combine json files	Coding	Request to create script
8	Deepseek	Improve text consiceness & directness	Methodology (Data)	Suggest improvements
9	Deepseek	Understanding of research text	Methodology	Help with concept
10	ChatGPT	EDA	Coding	Request for script
11	ChatGPT	Data Prep	Coding	Request for script
12	ChatGPT	Preprocess	Coding	Request for script
13	Deepseek	Data Enrichment	Coding	Request for script
14	ChatGPT	Help in describing the process	Methodology (Data)	Describe concept
15	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
16	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
17	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
18	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
19	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
20	ChatGPT	Help in paraphrasing text	Theoretical Foundations	Improve grammar and clarity
21	ChatGPT	Data Cleaning	Coding	Request for script
22	ChatGPT	Help in paraphrasing text	Methodology	Improve grammar and clarity
23	ChatGPT	Help in paraphrasing text	Methodology	Improve grammar and clarity
24	ChatGPT	Help in paraphrasing text	Methodology	Improve grammar and clarity
25	Deepseek	Clarification of data preprocessing	Methodology	Ask about text clarity
26	ChatGPT	Data Augment	Coding	Request for script
28	ChatGPT	Hyperparameter Tuning	Coding	Request for script
29	ChatGPT	BERT finetuning	Coding	Request for script
30	ChatGPT	Fit logistic	Coding	Request for script

31	ChatGPT	Predict unlablled data	Coding	Request for script

Appendix B – Sample raw articles

Below are 4 examples of raw articles, though not the full content as it's not possible to include it due to text lengths:


```
  },
  "259": {
    "source": "kvia",
    "url": "https://kvia.com/news/top-stories/2024/12/20/house-passes-bill-to-avert-shutdown-hours-ahead-of-deadline/",
    "date": "2024-12-20",
    "time": "23:39:32",
    "title": "House passes bill to avert shutdown hours ahead of deadline",
    "body": "By Clare Foran, Haley Talbot and Sarah Ferris, CNN\n\n(CNN) — The House voted on Friday to pass a stopgap funding bill in a bid to avert a shutdown just
    "summary": "By Clare Foran, Haley Talbot and Sarah Ferris, CNN (CNN) — The House voted on Friday to pass a stopgap funding bill in a bid to avert a shutdown just
    "keywords": "['warner', 'hours', 'deadline', 'suspension', 'shutdown', 'pass', 'bill', 'avert', 'republicans', 'way', 'funding', 'house', 'trump', 'debt', 'ahead'
    "image_url": NaN,
    "category": "general",
    "source_type": "mediastack_scrapper"
  },
  "260": {
    "source": "nytimes",
    "url": "https://www.nytimes.com/2024/12/20/world/americas/justin-trudeau-canada-popularity.html",
    "date": "2024-12-20",
    "time": "23:39:11",
    "title": "Justin Trudeau Is Stepping Down: What Comes Next?",
    "body": "Justin Trudeau said on Monday he would step down as prime minister of Canada and as leader of the Liberal Party, after months of mounting pressure.\n\nMr
    "summary": "Allies and opponents alike are calling for him to resign as prime minister.",
    "keywords": "['comes', 'justin', 'trudeau', 'prime', 'step', 'trudeaus', 'liberal', 'unpopularity', 'months', 'party', 'successor', 'stepping', 'minister']",
    "image_url": "https://static01.nyt.com/images/2024/12/21/multimedia/21canada-trudeau-explainer-fvmk/21canada-trudeau-explainer-fvmk-mediumSquareAt3X.jpg",
    "category": "general",
    "source_type": "mediastack_scrapper"
  },
  "338": {
    "source": "nesn",
    "url": "https://nesn.com/2024/11/bruins-brad-park-assesses-team-personal-legacy-during-centennial-celebrations/",
    "date": "2024-11-26",
    "time": "23:43:38",
    "title": "Bruins Great Assesses Team, Personal Legacy During Centennial Celebrations",
    "body": "BOSTON — Brad Park played eight seasons with the Bruins and settled in after a trade when he found comfort living in Lynnfield, Mass.\n\nHe scored 100 go
    "summary": "BOSTON &#8212; Brad Park played eight seasons with the Bruins and settled in after a trade when he found comfort living in Lynnfield, Mass. He scored
    "keywords": "['seasons', 'winning', 'assesses', 'dec', 'boston', 'bruins', 'garden', 'wins', 'tradition', 'celebrations', 'great', 'personal', 'team', 'legacy', '
    "image_url": NaN,
    "category": "general",
    "source_type": "mediastack_scrapper"
  },
  "339": {
    "source": "wral",
    "url": "https://www.wral.com/weather/ai-pros-and-cons-in-weather-forecasting-nc-november-2024/",
    "date": "2024-11-26",
    "time": "23:43:22",
    "title": "AI can produce weather forecasts quickly but faces challenges",
    "body": "Artificial intelligence, a technology that seems to be infiltrating every aspect of our lives, is making significant strides in the realm of weather fore
    "summary": "Artificial intelligence is enhancing the accuracy of weather forecasts by leveraging decades of historical weather data to generate predictions rapidl
    "keywords": "['faces', 'weather', 'ai', 'ability', 'technology', 'predict', 'quickly', 'produce', 'vast', 'forecasts', 'data', 'forecasting', 'models', 'challenge
    "image_url": "//www.wral.com/",
    "category": "general",
    "source_type": "mediastack_scrapper"
  },
```


Appendix C – Sample annotated articles

Below are 2 examples of annotated articles, though not the full content as it's not possible to include it due to text lengths:

```
{
  "193": {
    "source": "TechCrunch",
    "url": "https://techcrunch.com/video/figure-02-performing-tasks-autonomously/",
    "date": "2024-11-26",
    "title": "Figure 02 performing tasks autonomously",
    "body": "Experts from OpenAI, Anthropic, Cohere deliver exclusive insights across a must-attend industry even
    "summary": "Robotics startup Figure released a video update showcasing their Figure 02 humanoid robot autonom
    "category": "technology",
    "title_normalized": "Figure 02 performing tasks autonomously",
    "summary_normalized": "Robotics startup Figure released a video update showcasing their Figure 02 humanoid ro
    "t": 75,
    "t_bin": "old",
    "label": 0,
    "label_comment": "manual",
    "text": "This article was published 75 days ago. [TIME] Figure 02 performing tasks autonomously Robotics star
  },
  "2385": {
    "source": "Yardbarker.com",
    "url": "https://www.yardbarker.com/nba/articles/nba_cup_preview_dallas_mavericks_at_oklahoma_city_thunder/s1_
    "date": "2024-12-10",
    "title": "NBA Cup preview: Dallas Mavericks at Oklahoma City Thunder",
    "body": "The knockout round of the NBA Cup begins Tuesday night, with the final eight teams competing for a c
    "summary": "The knockout round of the NBA Cup begins Tuesday night, with the final eight teams competing for
    "category": "general",
    "title_normalized": "NBA Cup preview: Dallas Mavericks at Oklahoma City Thunder",
    "summary_normalized": "The knockout round of the NBA Cup begins Tuesday night, with the final eight teams com
    "t": 61,
    "t_bin": "old",
    "label": 0,
    "label_comment": "manual",
    "text": "This article was published 61 days ago. [TIME] NBA Cup preview: Dallas Mavericks at Oklahoma City Th
  },
```