Adapting to Change: The Temporal Persistence of Text Classifiers in the Context of Longitudinally Evolving Data

PhD thesis by **Rabab Ahmed Alkhalifa**
Primary Supervisor: Dr Arkaitz Zubiaga

Submitted in partial fulfilment of the requirement for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary University of London

July 2024

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply grateful for the opportunity to pursue my PhD studies in the **EECS** at Queen Mary University of London (QMUL). I would like to express my heartfelt appreciation to the **CCSIT** at Imam Abdulrahman bin Faisal University (IAU) for their support and belief in my academic journey. Their sponsorship, as part of my role as a faculty member at IAU University, has made this pursuit possible.

To my loving husband, who has been my pillar of strength throughout this challenging yet rewarding journey, I thank you for your unwavering encouragement and support. Your belief in my capabilities has been my constant motivation. To my wonderful daughter and son, your patience and understanding during the late nights and weekends spent studying have meant the world to me. Your resilience and love have kept me grounded and inspired. I am grateful to my parents and extended family for their continuous support and encouragement. Your belief in my potential has been a driving force behind my accomplishments.

I extend my sincere gratitude to my supervisor, Dr. Arkaitz Zubiaga, for his guidance, expertise, and unwavering dedication. Your insightful feedback and weekly meetings have been instrumental in shaping the direction of my research.

I express my gratitude to all those who have contributed to my research, with special acknowledgment to my second supervisor, Maria Liakata. Her invaluable insights and collaborative efforts have significantly enhanced the quality of my work. Additionally, I extend my appreciation to Elena Kochkina and Adam Tsakalidis as your support has played a pivotal role in elevating the depth and breadth of my research findings. I would like to acknowledge my colleagues, including Aiqi Jiang, Raneem Alharthi, and Wenjie Yin, who have stood by me throughout this academic journey. Your kindness, discussions, and shared experiences have made this pursuit memorable.

In the face of the global COVID-19 pandemic, our journey as students has been marked by unprecedented challenges. The strength we found in the Social Data Science lab in supporting each other, even from a distance in regular supervised meetings, is a testament to our resilience as a community. As I reflect on this journey, I am filled with gratitude for the people who have made it possible and could not mention their names. Thank you all for being a part of this chapter in my life.

Abstract

This thesis delves into the evolving landscape of NLP, particularly focusing on the temporal persistence of text classifiers amid the dynamic nature of language use. The primary objective is to understand how changes in language patterns over time impact the performance of text classification models and to develop methodologies for maintaining their effectiveness.

The research begins by establishing a theoretical foundation for text classification and temporal data analysis, highlighting the challenges posed by the evolving use of language and its implications for NLP models. A detailed exploration of various datasets, including the stance detection and sentiment analysis datasets, sets the stage for examining these dynamics. The characteristics of the datasets, such as linguistic variations and temporal vocabulary growth, are carefully examined to understand their influence on the performance of the text classifier.

A series of experiments are conducted to evaluate the performance of text classifiers across different temporal scenarios. The findings reveal a general trend of performance degradation over time, emphasizing the need for classifiers that can adapt to linguistic changes. The experiments assess models' ability to estimate past and future performance based on their current efficacy and linguistic dataset characteristics, leading to valuable insights into the factors influencing model longevity.

Innovative solutions are proposed to address the observed performance decline and adapt to temporal changes in language use over time. These include incorporating temporal information into word embeddings and comparing various methods across temporal gaps. The Incremental Temporal Alignment (ITA) method emerges as a significant contributor to enhancing classifier performance in same-period experiments, although it faces challenges in maintaining effectiveness over longer temporal gaps. Furthermore, the exploration of machine learning and statistical methods highlights their potential to maintain classifier accuracy in the face of longitudinally evolving data.

The thesis culminates in a shared task evaluation, where participant-submitted models are compared against baseline models to assess their classifiers' temporal persistence. This

comparison provides a comprehensive understanding of the short-term, long-term, and overall persistence of their models, providing valuable information to the field.

The research identifies several future directions, including interdisciplinary approaches that integrate linguistics and sociology, tracking textual shifts on online platforms, extending the analysis to other classification tasks, and investigating the ethical implications of evolving language in NLP applications.

This thesis contributes to the NLP field by highlighting the importance of evaluating text classifiers' temporal persistence and offering methodologies to enhance their sustainability in dynamically evolving language environments. The findings and proposed approaches pave the way for future research, aiming at the development of more robust, reliable, and temporally persistent text classification models.

List of Abbreviations

ABRR	Amazon Books Rating Reviews
BERT	Bidirectional Encoder Representations from Transformers
CWRs	Contextual Word Representations
CNN	Convolutional Neural Network
DTE	Discrete Temporal Embedding
FT	FastText Embeddings
GESD	Gender Equality Stance Detection
GloVe	Twitter Global Vectors for Word Representation
GPT	Generative Pretrained Transformer 2
HCSD	Healthcare Stance Detection
HAN	Hierarchical Attention Network
ITE	Incremental Temporal Embedding
ITA	Incremental Temporal Alignment
JI	Jaccard Index
LE-TESA	LongEval Temporal English Sentiment Analysis
LinearSVC	Linear Support Vector Machine
LLML	Lifelong Machine Learning
LogisticRegression	Logistic Regression
LSTM	Long-short Term Memory network
ML	Machine Learning
MultinomialNB	Multinomial Naive Bayes
NLP	Natural Language Processing
PLMs	Pre-trained Language Models
RPD	Relative Performance Drop
RoBERTa	Robustly Optimized BERT Pretraining
RQ	Research Question
SD	Stance Detection
SWRs	Static Word Representations
TESA	Temporal English Sentiment Analysis
2TA	Source-Target Temporal Alignment
2TE	Source-Target Temporal Embedding

Chapter 1

Introduction

1.1 Motivation

State-of-the-art machine learning text classifiers, including deep learning models, are effective for identifying patterns in data to then apply to similar-looking data. However, these models tend to be supervised, i.e., reliant on previously labeled data. The process of labeling data tends to be both time-consuming and expensive, leading generally to a scarcity of labeled data, often not completely encompassing the patterns of the data the model may end up seeing during the application or testing phase of the model. This leads, among others, to the limited generalizability of models in highly dynamic environments where data evolves over time, leading to labeled datasets becoming obsolete after some time and limiting their applicability to new data. This limitation leads to labeled datasets becoming outdated, restricting the sustainability of models and adaptability when tested in new and emerging data sources in different fields, including image classification [Alzubaidi et al., 2023] and natural language processing (NLP) [Khurana et al., 2023]. While one may argue that the problem can be resolved by labeling new, more recent datasets, this thesis studies the scenario where one cannot afford such recurrent annotation of new datasets due to the cost and effort involved, and hence the adaptation needs to be achieved in the absence of new labels.

Recent developments in NLP leverage statistical patterns of context observed surrounding words [Volkova et al., 2016, D'Andrea et al., 2019, Lai et al., 2018, Lai et al., 2019], with

the intuition that words in close proximity to each other help determine and represent their meaning. One of the approaches building on this intuition are word embeddings [Mikolov et al., 2013a]. Word embeddings connect every word in vector representations based on its context, making them a powerful tool for enhancing text classifier performance. Recent state-of-the-art methods rely on static [Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017, Mikolov et al., 2018] and dynamic language model embeddings [Shibata et al., 1999, Devlin et al., 2019, Radford et al., 2019, Liu et al., 2019, Ha et al., 2020] to boost text classifier performance. In our context, to retain long-term classifier performance, we would want our text classifier to update vector representations based on incoming data and future data without the need for labeling additional data. Moreover, while a machine learning text classifier can fit any training data, it is more challenging to interpolate to new data points in a reasonable way without additional learning efforts, which consumes money and time. All of these weaknesses make our work a much-needed and challenging research area within an understudied task that requires establishing a benchmark for rigorous and consistent experimentation.

The evolving nature of language and social opinions adds an additional layer of complexity to the challenges faced by text classifiers. Language undergoes continuous changes, reflecting shifts in societal norms and opinions and the emergence of novel concepts and words. For instance, consider the evolution of public sentiments on climate change over the past two decades:

- **Sentence from 2000:** "Global warming is a theory that needs more proof; it's not urgent."
- **Sentence from 2010:** "Evidence for climate change is mounting, and we need to start taking action."
- **Sentence from 2020:** "Climate change is an undeniable crisis that requires immediate global action."

The context over two decades in the above example shows that language and urgency surrounding climate change have evolved from skepticism to an accepted crisis [Alkhalifa et al., 2024]. Models not updated with recent discussions and policy changes might fail to accurately capture the critical tone and terminology used in current dialogues about the environment. Similarly, the rapid emergence of new vocabulary, as witnessed with terms like COVID-19, highlights the dynamic nature of language, presenting unique challenges for text classifiers [Alkhalifa et al., 2020b].

In this thesis, our primary focus and interest lie in user-generated content on social media, which serves as a rich and dynamic source of linguistic expressions. User-generated content captures real-time opinions, diverse language use, and evolving trends, making it a suitable domain for studying the temporal dynamics of language and opinion. The inherent challenges of adapting models to this dynamic environment encouraged us to explore innovative approaches to assess persistence and enhance the adaptability of text classifiers.

Our work adopts novel approaches, focusing on automatically quantifying classifier temporal drops and adapting to the temporal dynamics in textual expressions. This approach aims to achieve cross-temporal textual classification, enabling text classifiers to effectively navigate the evolving linguistic landscape over time. Our investigation specifically highlights the **study of the impact of time on text classifiers** in general, employing state-of-the-art prediction algorithms across diverse textual datasets. Additionally, we **assess the characteristics of datasets** that make them more prone to performance drops over time, **propose temporal adaptation methods**, with a special focus on stance detection task where language and social attitudes towards different events impact language use in social media, contributing valuable insights to the new field of temporal text classification.

1.2 Research Questions and Objectives

Our hypothesis is grounded in the idea that textual classification data, and the linguistic content underlying in such texts, spanning diverse topics or events, undergoes temporal fluctuations, resulting in the emergence of historically evolving contexts. This historical context signifies the evolution of language use, shedding light on how it has been generated and its influence on our contemporary understanding, as compared to past and future perspectives. In essence, our overarching research inquiry seeks to investigate:

RQ1. **How does the dynamics of language use impact text classifiers?** or, in other words, can we capture the temporal dynamics of language use so as to adapt text classifiers? Our exploration began with a comprehensive review of the theoretical foundations of dynamic of textual data in Chapter 2. This laid the groundwork for understanding the temporal dynamics of language use. The subsequent chapters, 3 and 4, delve into methodological approaches and practical assessments, providing insights into the challenges posed by dynamic language use on text classifiers.

RQ2. **How can we assess the influence of time on text classifier performance using a longitudinal dataset?** We studied the influence of time on text classifier performance through different longitudinal datasets. Chapters 3, 4 and 5 laid the foundation for

understanding the challenges associated with temporal persistence of text classifiers over different longitudinal datasets.

RQ3. **What machine learning and statistical methods can facilitate the maintenance of classifier accuracy in the context of evolving language use?** Chapters 4 and 6 address RQ3 by evaluating various machine learning and statistical methods to enhance classifier accuracy amid evolving language use. The assessment encompasses a shared task that enabled the scientific community to compare different models in terms of short-term, long-term, and overall temporal persistence. These chapters contribute insights into the methods that can mitigate challenges posed by dynamic language use and sustain classifier accuracy.

RQ4. **How can we maintain our classifiers' performance by circumventing the impact of evolving language and patterns?** RQ4 is addressed in Chapter 5, where we propose a novel solution to enhance text classifier performance by incorporating temporal information into word embeddings. This chapter sheds light on the challenges posed by evolving language and patterns, and performs a comparison of methods to improve classifier accuracy in this context.

Hypothesis. Our hypothesis is that ***textual classification data on a variety of issues fluctuates over time, creating historically shifting contexts which will in turn impact text classification model performance over time***. Through the following chapters, we introduce detailed experiments to investigate the above research questions, delving into specific NLP methods such as the effects of various word representations, the efficacy of deep learning text-based classifiers, the detection and inspection of shifting features, and the potential utility of frequency-based methods for text classifier adaptation of temporal changes. Our aim is not only to enhance our comprehension of the persistence of text classifier performance over time but also to provide pragmatic insights into sustaining text classifier performance while alleviating the challenges posed by evolving language and knowledge. These inquiries serve as the foundation for our comprehensive investigation into the temporal dynamics of text classifiers' performance.

In pursuit of these overarching questions, we have defined the following specific **research objectives**:

- **O1.** To design and formalize a robust methodology to enable temporal evaluation of text classification models, including the identification and collection of suitable datasets, as well as a definition of the evaluation methodology.
- **O2.** To gain a deep understanding of the impact of different datasets when applied

to longitudinal social media data and their influence on text classifier performance.

- **O3.** To assess the effectiveness of various classification models and representation approaches in achieving temporal persistence in text classifiers.
- **O4.** To propose computational methods that leverage older annotated datasets while minimizing the drop in performance, thereby facilitating the adaptation of text classifiers to changing language use dynamics.

1.3 Thesis Roadmap

This thesis is structured with an introductory chapter, a background chapter, a dataset chapter, three analytical chapters, and a conclusive chapter. Chapters 1 and 2 establish the research's motivation, questions, objectives, and background. Chapter 3 delineates the datasets employed in the subsequent analysis. Chapters 4 to 6 delve into a comprehensive examination of the temporal dynamics in the performance of text classifiers. The conclusive Chapter 7 wraps up the study and set future research directions.

Specifically, the thesis is organized as follows:

Chapter 2 Background on Temporal Persistence of Text Classifiers This chapter provides the background and related work relevant to understanding the rest of the thesis. It explores three key dimensions: a detailed examination of challenges related to studying temporally-evolving data, an introduction to methods and techniques for the temporal persistence of text classifiers, and a comprehensive review of existing literature in text classification with a focus on its connection to semantic change literature. Additionally, it introduces a novel theoretical framework tailored to capturing stance dynamics in social media through an in-depth review of prior research on language use and temporal changes, specifically in the context of stance detection and other opinion change literature, state-of-the-art stance detection datasets, and other longitudinal datasets. Finally, the chapter identifies both core challenges within the domain of stance detection and general gaps in temporal text classification, offering a robust foundation for understanding the issues associated with temporally-evolving data and maintaining temporally persistent classifiers.

Chapter 3 Temporal Analysis Methodology, Evaluation Metrics, and Longitudinal Datasets. This chapter introduces the methodology we proposed to evaluate any longitudinal datasets employed to assess the persistence of text classifiers. It also introduces the datasets we use throughout the thesis for our experiments, including

two new datasets that we create and one that we borrow from previous work. The chapter presents the details of these datasets and performs an analysis of their characteristics by looking at their changes over time. Additionally, it outlines all the steps taken to preprocess and sample the data used in the thesis, ensuring its quality and representativeness for the experiments covered in the following chapter.

Chapter 4 Assessing the Temporal Persistence of Text Classifiers. Here, we delve into the methodology and findings of the temporal persistence of text classifiers showing a novel experiential set up to understand the deficiency of text classifier performance when using dynamic longitudinal datasets containing different characteristics and linguistic features.

Chapter 5 Implementation of Temporally Adaptive Classification Methods. This chapter presents a novel approach to implement a persistent text classifier by addressing the problem of performance drop over time with using two novel stance detection datasets. It focuses on incorporating temporal information into word embeddings and presents findings from experiments comparing different methods across various temporal gaps.

Chapter 6 Comparative Analysis of Classifier Temporal Persistence. This chapter discusses our efforts in organizing a shared task on longitudinal evaluation of text classification models, LongEval, where we encouraged other researchers to develop their own methods by following the guidelines and datasets we published. The shared task was organized as one of the labs of the CLEF forum. Its main aim was to attract a wider community of NLP researchers to engage with the task, which in turn gave us the opportunity to broaden our analysis and findings by incorporating the approaches tested by others.

Chapter 7 Conclusions and Further Work. This chapter summarizes the achievements and findings of the thesis, proposing future research avenues. It highlights the importance of evaluating text classifiers' temporal persistence and how the thesis helps pave the way to researchers undertaking this challenging task in the future.

1.4 Contributions

This thesis makes the following contributions:

- Shedding light on the importance of expanding the time frame covered in classification datasets, this thesis underscores how this extension significantly enhances machine

learning model performance evaluation beyond a single performance score (Chapter 2 and 4).

- Proposing a practical methodology for assessing the persistence of text classifiers over time using time gaps, widely applicable to various classification tasks (Chapter 3, 4 and 5).
- Emphasizing and analyzing how language use evolves in longitudinal text, highlighting the challenge of temporal dynamics of changing vocabulary for text classifiers (Chapter 3).
- Demonstrating when, why, and how a text classifier's performance changes and writing insights about dataset complexity through a comparison of word familiarity between training and testing data (Chapter 4).
- Investigating the influence of temporal dynamics on text classifiers' performance across diverse models and datasets, revealing patterns of performance decline over time (Chapter 4 and 5).
- Introducing adaptation approaches for evaluating model generalization in real scenarios by incorporating knowledge derived from unlabeled data (Chapter 5).
- Introducing the concept of evaluating text classifiers over time as a collaborative task for the NLP research community (Chapter 6).

1.5 Associated Publications

The detailed work in this thesis has been presented in national and international scholarly publications. The **journal** venue names are highlighted in bold, and the *conference* names are in italic.

Journal Publications

- **Chapter 2:** Sections 2.3 and 2.4 accepted for publication in the **International Journal of Digital Humanities**[1] as **Capturing stance dynamics in social media: open challenges and research directions** ([Alkhalifa and Zubiaga, 2022]).
- **Chapter 4:** Research from this chapter has been accepted for publication in the **Information Processing & Management**[2] as **Building for tomorrow: Assessing the temporal persistence of text classifiers** ([Alkhalifa et al., 2023]).

Conference Publications

- **Chapter 5:** Presented at the *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks (OASIS '21)*, co-located with *ACM Hypertext 2021*[3] as **Opinions are made to be changed: Temporally adaptive stance classification** ([Alkhalifa et al., 2021]).
- **Chapter 6:** Further development of the Temporal Multilingual Sentiment Analysis (TMSA) tweets dataset, initially shared publicly in [Yin et al., 2021], led to the creation of its English-only version, LE-TESA[4]. This refined dataset, specifically cleaned and human-annotated, was utilized as a novel longitudinal baseline dataset for the LongEval shared task. It provided researchers with a robust tool for investigating longitudinal sentiment analysis. LE-TESA's development and its application in the task were initially presented as part of a publication named **LongEval: Longitudinal Evaluation of Model Performance** ([Alkhalifa et al., 2023b]) at the *Conference and Labs of the Evaluation Forum 2023 (CLEF 2023)*. An extended overview of the *CLEF-2023 LongEval lab* and its participants' results were further detailed in **Extended overview of the CLEF-2023 LongEval lab on longitudinal evaluation of model performance** ([Alkhalifa et al., 2023a]).

[1] https://www.sciencedirect.com/journal/information-processing-and-management
[2] https://www.sciencedirect.com/journal/information-processing-and-management
[3] https://dl.acm.org/doi/proceedings/10.1145/3472720
[4] https://clef-longeval.github.io/data/

Preprints

- **Chapter 3: The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset** ([Yin et al., 2021]) including a dataset were publicly shared as an archival dataset paper.

Other Publications

Additional work carried out during the PhD, though not directly related to the thesis, has been published in workshops, including:

- **QMUL-SDS at CheckThat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions** ([Alkhalifa et al., 2020b]), presented at *CLEF CheckThat! 2020* [5].
- **QMUL-SDS @ SardiStance2020: Leveraging network interactions to boost performance on stance detection using knowledge graphs** ([Alkhalifa and Zubiaga, 2020]), presented at *EVALITA 2020* [6].
- **QMUL-SDS @ DIACR-ITA evaluating unsupervised diachronic lexical semantics classification in Italian** ([Alkhalifa et al., 2020a]), also presented at *EVALITA 2020*.

[5] https://sites.google.com/view/clef2020-checkthat/
[6] https://books.openedition.org/aaccademia/6732

Chapter 2

Background on Temporal Persistence of Text Classifiers

Social media platforms provide invaluable insights for analyzing public opinion on various societal issues, making them crucial for tasks such as stance detection, fact-checking, hate speech detection, and sentiment analysis. In the realm of text classification, whether for sentiment or stance detection tasks, temporal persistence involves categorizing individual social media posts, considering the evolving nature of textual expressions toward a given issue due to shifts in natural language use over time. While previous research has predominantly focused on short-term datasets, there is a growing interest in studying longitudinal datasets to understand evolving dynamics over time for more temporally persistent classifiers. The ever-changing linguistic and behavioral patterns in new data necessitate adaptive text classifier models to accommodate these shifts.

In the upcoming sections, we delve into the theoretical foundations of textual classification, conducting a thorough review of existing literature on temporal text classification. This review will explore state-of-the-art temporal text adaptation methods and the understanding of semantic shifts in text data and how they are connected to temporally persistent classification. Additionally, we will introduce a theoretical framework that serves as the cornerstone for comprehending the temporal dynamics of stance detection classifiers, which is further investigated through novel methods for temporal adaptation (Chapter 5).

This comprehensive background chapter is designed to set the stage for our subsequent

chapters. Our aim is to shed light on the temporal dynamic factors that impact text classifiers, providing valuable insights into their performance over time. Furthermore, we seek to explore their applicability to various classification tasks, with a particular focus on stance detection. While addressing research gaps and motivations, we will carefully navigate both core and general challenges within this domain.

2.1 Overview of Text Classifier

Most NLP research predominantly concentrates on exploring methods, algorithms, data structures, and the reliability of predictive techniques, primarily machine learning, to comprehend representative textual data. This encompasses the utilization of lexicon cues, supervised, weakly supervised, and unsupervised machine learning tools, as well as deep learning and word embedding models, which have proven highly effective (more in Section 2.2.2). These cutting-edge approaches contribute significantly to various NLP tasks, including stance detection and sentiment analysis. However, it is crucial to note that the roots of textual classification extend beyond the realm of computer science to encompass linguistics, communication research, and real-life contexts.

Traditional text classification Traditional text classification methods include various supervised learning algorithms that have been widely used due to their simplicity and effectiveness. Linear Support Vector Machines (LinearSVC) are an algorithm that searches for a hyperplane (decision boundary) to linearly separate classes, widely used in numerous NLP classification problems. It can generalize to high-dimensional and sparse feature spaces without requiring extensive feature engineering. Logistic Regression is typically used for binary classification, utilizing a probability curve to predict classes. It fits a single line to split the space into two using a sigmoid function and can interpret model coefficients to determine feature importance. Multinomial Naive Bayes (MultinomialNB) is a probabilistic learning algorithm that assumes feature independence, robust against infrequent features, and effective in text classification tasks.

Deep learning models Deep learning models have revolutionized text classification by leveraging large datasets and complex architectures to capture intricate patterns in text. Convolutional Neural Networks (CNNs) can retain the order of feature elements and process text by sliding over it with a convolutional filter, learning various levels of n-gram co-occurrences. This shift-invariant classifier is highly effective in extracting features from text sequences. Long-Short Term Memory (LSTM) Networks are capable of learning long-term dependencies and semantic changes due to their recurrent units and temporal backpropagation. They model sequences word by word, making them suitable for

capturing latent syntactic and semantic signals. Hierarchical Attention Networks (HANs) combine GRU units with an attention mechanism to generalize predictions based on word and sentence relationships. They use bidirectional GRU layers to capture context and an attention mechanism to highlight important parts of the text.

***Large Language Models (LLMs) *** Large Language Models (LLMs) have set new benchmarks in NLP by utilizing extensive pretraining on large corpora and fine-tuning for specific tasks. Notable LLMs include BERT (Bidirectional Encoder Representations from Transformers), which generates contextual word representations by processing all tokens in a text bidirectionally. It uses a static masked language modeling (MLM) loss objective to fine-tune on downstream tasks, making it effective in understanding context. RoBERTa (Robustly Optimized BERT Pretraining Approach) enhances BERT's capabilities with improved pretraining efficiency and a dynamic MLM loss objective, making it more effective in capturing contextual meanings across a larger vocabulary. GPT-2 (Generative Pretrained Transformer 2) is an auto-regressive model that generates text iteratively, predicting tokens based on left-to-right context. It is primarily used for text generation tasks, employing a causal language modeling (CLM) loss objective.

For a text classifier to be more generalized, robust, and able to persist over time, the traditional method of gathering consistent data, assuming static words distributed uniformly across time frames, proves inadequate. This limitation is particularly evident in highly dynamic systems featuring evolving topics and emerging vocabulary. Social media platforms, being hubs of new subjects and real-world events, foster constantly changing vocabularies and dynamic shifts within short and long timeframes. While evolutionary classification, as introduced by [He et al., 2018], presents a solution to manage short-term fluctuations and long-term word changes, it involves a trade-off, demanding more annotated data to effectively address the impact of language evolution. Additionally, it requires more up-to-date language models that reflect a more timely and representative vocabulary along with contextual meaning. This differs from the objectives set forth in this thesis of achieving temporal persistence in scenarios where continually labeling new datasets over time becomes impractical or unaffordable.

In this thesis, GPT embeddings, along with BERT and RoBERTa, were used to enhance text classification tasks by providing rich contextual word representations. GPT-2 embeddings were particularly beneficial in tasks requiring human-like text generation and understanding nuanced contextual information [Radford et al., 2019]. In our work, the embeddings are integrated into the classification pipeline using TensorFlow and the Hugging Face Transformers library. The models were trained with custom architectures, incorpo-

rating features like hierarchical attention networks and convolutional layers, to optimize performance on various datasets. The embeddings were fine-tuned to align with specific dataset characteristics, ensuring that the models could effectively handle the dynamic and evolving nature of real-world textual data. For more details on the classifiers used throughout this thesis, refer to Chapter 4, where we provide an in-depth analysis of their implementation and performance.

2.2 Temporal Perspectives in Text Classification Research

In the rapidly evolving landscape of text classification, the temporal dimension plays a crucial role in shaping model performance. This section delves into the intricate relationship between time and text classifiers, exploring how the dynamics of language evolution impact the effectiveness of these models over different time periods. We review research methodologies, approaches, and empirical investigations that shed light on the challenges and opportunities posed by temporal shifts in textual data.

2.2.1 Adaptation to Temporal Changes

Exploring Temporal Shifts in Data

Related areas of research addressing temporal aspects of model performance have focused on evolutionary learning [He et al., 2018, Pustokhina et al., 2021] and lifelong machine learning (LLML) [Nguyen et al., 2020, Ha et al., 2018, Xu et al., 2020]. These works, however, assume that longitudinally annotated data is available, which can be incorporated into the model for continuous training.

This line of research can be complementary to ours, however, it is not applicable in our scenario where the data available for training is temporally restricted, i.e., the model needs to make the most of labelled data for a particular period of time, which will then be applied to test data distant in time.

In this scenario, our own work in Alkhalifa and Zubiaga [2022], as discussed in Section 2.3, contributes a theoretical perspective on the influence of time on classification models. This study delves into several challenges encountered by text classification models when dealing with data from time periods distinct from those used for training. Specifically focusing on the stance detection task in the realm of social media, the authors, through a comprehensive literature analysis, identified three primary factors attributing to the temporal impact on stance detection models: stance utterance, stance context, and stance influence.

Other works have studied temporal aspects of model performance through empirical experi-

ments. The link between words and classes evolves over time as a result of terms emerging, disappearing, and exhibiting varying dimensions, e.g., word polarity drift in sentiment classifiers, as demonstrated by Rocha et al. [2008]. Lukes and Søgaard [2018] investigated polarity rank drift over time using a logistic regression classifier. They illustrated how the polarity of words can impact model performance. Another line of research that focuses on concept drift looks into the impact of changes in label distribution over time [Nishida et al., 2012], such as hashtag usage change. Continuing with text classifiers, our own work in Alkhalifa et al. [2021] which will be discussed in Chapter 5, and Röttger and Pierrehumbert [2021] conducted preliminary experiments that resulted in findings about the impact of time on model performance and dataset usability on upstream and downstream tasks.

In our thesis, we introduced a methodology for temporal splitting, dividing the dataset into intervals to capture a range of temporal dynamics (discussed in detail in Chapter 3). The data is split into temporal gaps such as 0, ± 1, ± 2 years, representing both past and future intervals relative to the training period. This approach aids in understanding model performance over different time spans and the impact of lexical and contextual changes. By ensuring dataset consistency and relevance across time periods, we can observe performance degradation in models trained on specific intervals and evaluated on subsequent ones.

Additionally, we employ a lighter version for simpler analysis, as discussed in Chapter 6 using sentiment analysis data with human annotated testing splits. This version involves basic model training and evaluation on a smaller subset of data focusing on key aspects of temporal variations, such as major shifts in language use or trending topics.

Empirical Investigations in Temporal Text Classification

There is a body of research looking into the impact of time on classification performance, which we summarise in Table 2.1. Previous work investigated **supervised temporal adaptive approaches** that study the impact of the size of the annotated dataset on the persistence of performance over time [Rocha et al., 2008, Nishida et al., 2012, He et al., 2018, Lukes and Søgaard, 2018, Florio et al., 2020] assuming that the annotated data is available for training and testing periods. Moreover, [Florio et al., 2020] demonstrated that when the **amount of annotated data is progressively augmented** from the target year, some models tend to be more sensitive than others. This differs from our main objective, where we consider scenarios where the annotated data for model training only covers a limited time period. Using temporal gaps between training and test data, we restrict our model's ability to acquire additional annotated data. Our research is unique in that we aim to measure model persistence by assessing design models' temporal generalisability across a sizable number of dynamically evolving training and test sets.

Ref.	Task	Temporal Granularity	Dataset (time-span)
Rocha et al. [2008]	Document classification	Yearly	ACM-DL (1980-2001) and MedLine (1980-2001)
Nishida et al. [2012]	Document classification	Daily and hourly	Hashtags as different topics
Preoţiuc-Pietro and Cohn [2013]	Hashtag classification	Monthly	Twitter public Gardenhose stream (2011)
He et al. [2018]	Topic classification	Monthly and yearly	NYTimes (1996 to 1997), RCV1 (Jan 01, 1987 to June 19, 2007)
Lukes and Søgaard [2018]	Review raing prediction	Across group of years	ARR (2001 to 2014)
Florio et al. [2020]	Hate speech detection	Monthly	unbalanced Haspeede (Oct 2016 to 25 April 2017), TWITA (2012-2017)
Murayama et al. [2021]	Fake news detection	Across group of years	MultiFC, Horne17, Celebrity, Constraint (2007–2015, 2016, 2016-2017, 2020)
Allein et al. [2021]	Fact-checking	Daily	MultiFC
Röttger and Pierrehumbert [2021]	Document classification	Monthly	Reddit Time Corpus (RTC) (March 2017 and February 2020), Political Subreddit Prediction (PSP)
Alkhalifa et al. [2021]	Stance detection	Yearly	GESD (2014-2019)
Lazaridou et al. [2021]	Question answering	Yearly	WMT (2007-2019), ARXIV (1986-2019) and CUSTOMNEWS (1969-2019)
Alkhalifa et al. [2023]	Stance detection, sentiment analysis and review rating prediction	Yearly	GESD (2014-2019), TESA (2013-2020) and ABRR (2000-2018)
Chuang et al. [2023]	Stance detection	Across group of years	COVID (2020-2022), WTWT (2015-2018), SCI-ERC(1980-2016), PUBCLS (Not specified)

Table 2.1: Summary of related works addressing data evolution and model performance over time.

Several existing works introduced **unsupervised temporal adaptive approaches** also to improve the temporal persistence of text classifiers, either through **incremental static embedding training** [Alkhalifa et al., 2021, He et al., 2018] or through **continuous pretraining** using transformers [Röttger and Pierrehumbert, 2021, Lazaridou et al., 2021, Chuang et al., 2023]. Their objective, however, differs from ours, as we focus instead on drawing an in-depth understanding of the impact of different aspects on temporal performance by employing widely-used methods. Our objective is to perform a hitherto lacking investigation into the extent to which a model's performance drops due to temporal language evolution [Alkhalifa et al., 2021] (Chapter 5), as well as when and why it occurs [Alkhalifa et al., 2023] (Chapter 4), with the aim of devising best practices for the development of models with their temporal persistence as the objective.

Related Papers	Feature type
Basic Features	
Sun et al. [2016]	Theme word with target (TWT)
Wojatzki et al. [2018], Sun et al. [2016]	Length
Sobhani et al. [2019]	Stylistic features
Linguistic Features	
Somasundaran and Wiebe [2009], Walker et al. [2012b], Mandel et al. [2012], Rekik et al. [2019]	lexicon cue and Polarity
Somasundaran and Wiebe [2009]	Syntactic rules
Somasundaran and Wiebe [2009]	Modal verb
Somasundaran and Wiebe [2009]	Unsupervised methods
Wojatzki et al. [2018]	Text similarly
Mandel et al. [2012], Wojatzki et al. [2018]	Semantics of words
Hasan and Ng [2014]	Frame-semantic features
Anand et al. [2011]	Repeated Punctuation.
Anand et al. [2011]	Syntactic Dependency
Sun et al. [2016]	Syntax tree (STPH)
Sun et al. [2016]	Dependency tree (DPGD)
Sun et al. [2016]	POS
User Profiling Features	
Nguyen et al. [2012], Graells-Garrido et al. [2020]	Profiling attributes
Addawood and Bashir [2016], Anand et al. [2011], Somasundaran and Wiebe [2009]	Personal Word Choices
Mandel et al. [2012], Volkova et al. [2016], Graells-Garrido et al. [2020]	Demographic Analysis

Table 2.2: Classical feature engineering techniques

2.2.2 Semantic Evolution in Textual Data

The dynamic nature of language usage is an ideal setting for exploring the impact of temporal changes in NLP. As linguistic expressions shift and transform over time, understanding the influence of semantic evolution becomes crucial. In this subsection, we delve into the field of diachronic lexical semantics, which focuses on exposing changes in word-level language across temporal dimensions. Our investigation of semantic shifts goes beyond mere detection; we navigate the intersection of this semantic evolution and the dynamic world of textual data. Text classifiers, dealing with the dynamic and evolving nature of language, face unique challenges and opportunities as they interact with the constantly changing, evolving vocabulary dynamics and adaptations in language use. This investigation provides context for discussing the significant implications of temporally-evolving data for text classifiers, providing useful insights into the evolving nature of NLP text classifier data that extend beyond its dynamic semantic dimension.

Understanding Semantic Shifts

The quantitative analysis of language evolution over time is an emerging research area within Natural Language Processing (NLP) [Turney and Pantel, 2010, Hamilton et al., 2016c, Dubossarsky et al., 2017]. The study of Diachronic Lexical Semantics [Tahmasebi et al., 2021, Kutuzov et al., 2018] contributes to detecting word-level language evolution, bringing together researchers from computational linguistics, cognitive science, statistics,

mathematics, and historical linguistics. Identifying words whose lexical semantics have changed over time has applications in historical linguistics and NLP.

Unsupervised diachronic lexical semantics detection approaches can be categorized based on the type of word representations used in a diachronic model, such as graph or probability distributions [Frermann and Lapata, 2016, Azarbonyad et al., 2017], temporal dimensions [Basile and McGillivray, 2018], frequencies or co-occurrence matrices [Sagi et al., 2009, Cook and Stevenson, 2010], and neural- or Transformer-based methods [Hamilton et al., 2016c, Del Tredici et al., 2019, Shoemark et al., 2019, Schlechtweg et al., 2019, Giulianelli et al., 2020].

Systems operating on representations like Skip-gram or Continuous Bag-of-Words often use deterministic approaches, employing mathematical matrix transformations [Hamilton et al., 2016c, Azarbonyad et al., 2017, Tsakalidis et al., 2019] or machine learning models [Tsakalidis and Liakata, 2020]. The goal is to learn a mapping between independently trained word vectors from different time periods, with cosine distance being a common measure for diachronic semantic change [Turney and Pantel, 2010].

However, using cosine distance can introduce bias due to word frequency variations [Dubossarsky et al., 2017]. Some approaches, like that of Tan et al. [2015], mitigate this by considering only vectors of the top frequent terms in the transformation matrix calculation. Incremental update strategies [Kim et al., 2014, Del Tredici et al., 2019] use intersections of words between datasets in each time frame to compare word shifts across different years. Temporal Word Embeddings with a Compass (TWEC) [Di Carlo et al., 2019] leverages freezing selected vectors based on the model's architecture, learning a parallel embedding for all time periods from a base embedding with frozen vectors.

Breaking Down Semantic Variations in Temporal Text Classification

The dynamic nature of textual data over time introduces profound implications for text classifiers. As language evolves, so do the challenges and opportunities faced by these classifiers in effectively capturing and generalizing patterns. Understanding the implications of temporally-evolving data is crucial for developing robust and adaptable models. In broader view, when analysing textual data, features usually transformed into a numerical matrix using variations of two main approaches: classical approaches using frequency-based methods and vector representations using context-based methods. Frequency-based method are based on counting (e.g., co-occurrence matrix), while prediction-based methods are based on predicting contextual meaning form surrounding words (e.g., continuous bag of words). In the following, we summarize some related work under these two categories and discuss

how data can impact text classifiers performance based on our understanding of semantic-shift literature.

Classical feature-based learning In social media application using NLP and traditional machine learning model, features can be extracted by using **stylistic signals** from text such as bag of n-grams, char-grams, part-of-speech labels, and lemmas, **structural signals** such as hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet, and **pragmatic signals** related to author's profile. Table 2.2 summarize related work in these regards.

Studies which target users' profile and human opinion modeling are also taking two directions. In first branch, some studies that target individuals focused on evolution of singular user profiles including user posts, full name, biography, URLs, hashtags and number of following/followers/posts, posting time. Although, state-of-art user profiling tools can participate in aggregating more insights about both author and respondent in social media, in our work, we focus on collective analysis which represent the second branch with more focus on aggregated features which is go beyond user profiling by include linguistics and contextualized feature representation which can guarantee long term performance for longitudinal-based applications without violating general term of use for social media platforms.

In [Graells-Garrido et al., 2020], with an objective to identify users who have changed their stance before and after an event, they find out that over-time people adopt new technologies to express their stance such as using emoji and pictures. This is different than the classical assumption of using textual tokens within an utterance and removing all other segments. They monitor semantic meaning of emojis with an attempt to identify how demographic and user profile explain variations in stance. Though, they choose two culturally similar for the longitudinal analysis, Argentina and Chile, they have different legalization low for selected topic, Abortion. In their work, they use classifier to predict gender and location while keeping only profile with high confidence. With that, thy only kept the largest connected component (LCC) of the discussion network containing text of retweets, mentions, replies, and quotes. Similar to number of researches that uses network homophily (see Section 2.3.2), during legislative session discussion peaks while return to normality afterwards.

In [Volkova et al., 2016], their visualisation tool revealed a fascinating pattern of collective sentiment dynamics using a state-of-the-art collective prediction model [Nguyen et al., 2012, O'Connor et al., 2010] designed to understand public opinion by shedding light on

the contextual variations: their context, different languages and geographical locations. Their work utilised machine learning models to predict opinion dynamics towards the USA in Russia and Ukraine over time. Their storyline visualisation is a clear illustration of how people change their opinion from neutral to positive or negative over time. They further correlated these changes with incidents that occur in the same time period, making their work very beneficial for longitudinal stance analysis. However, they used sentiment and emotion features to understand opinion dynamics to overcome shortage of general use stance detection models.

Neural context-based learning With modern deep learning model more shift toward **contextualised representations** using word vector representation algorithms, either by having personalized language model trained on task specific language or as a pre-trained language model offered after training using complex architecture and billions of documents. This is mainly because context is important to define the semantic meaning of the word.

Distributional semantics of words aims to collectively categorize semantic similarities between linguistic items using distributional properties, occurrences, and contexts from a given corpus. This have so far boost performance in many NLP downstream tasks including stance detection. Though, many semantic shift literature discussed referential effect as a factor which add noise into word meaning and generate temporal meaning shift [Zhang et al., 2016, Tan et al., 2015, Kim et al., 2014, Del Tredici et al., 2019]. They argue that corpora with smaller time spans are useful for analysing socio-cultural semantic shifts [Kahmann et al., 2017].

Moreover, they detected this change by compare the meaning of a word in one space to its meaning in another space and measure the size of the semantic shifts. It is worth noting that this task is different than capturing multiple senses a word might have which is more related to polysemy. In such that they believe that same word can have different meanings in different contexts [Tian et al., 2018, Azarbonyad et al., 2017]. However, sometimes contextual information changes due to undefined factors, and associated contexts changes or shifts cross time creating new contextual neighbours without changing the word meaning.

Considering these insights, the exploration of semantic variations in temporal text classification becomes imperative. The detection of contextual variability and model stability measures becomes essential to reasoning about text classifiers' performance in dynamic linguistic environments. This exploration aims to bridge the gap between the evolving semantic landscape and the challenges faced by classifiers over time.

2.3 Capturing Dynamics in Stance Detection Classifiers

With the proliferation of social media and blogs that enable anyone to post and share content, professional accounts from news organisations and governments aren't any longer the sole reporters of events of public interest [Kapoor et al., 2018]. Posting a tweet or a video, or writing an article that goes viral and reaches millions of individuals is now more accessible to ordinary citizens [Mills, 2012]. Where anyone can post their views on social media, the use of social media gains ground as a data source for public opinion mining. This data source provides a goldmine for nowcasting public opinion by aggregating the stances expressed by individual social media posts on a particular issue.

Stance detection is indeed a crucial task where time has a substantial impact, not least on topics where public opinion evolves quickly due to societal evolution. Therefore, stance is expected to constitute a particularly challenging task in the realm of achieving temporally persistence text classifiers. Because of this, stance detection constitutes a core part of this thesis which is inherently present in the datasets we use for our experiments and analysis.

Within this subsection, we introduce a comprehensive survey delving into the convergence of computational linguistics and the temporal dimension of human communication in digital media. Through an in-depth review of emerging research, we explore the impact of dynamics, semantic nuances, and pragmatic factors on linguistic data in general and stance classification in particular. Furthermore, we delve into current strategies for capturing stance dynamics in the realm of social media and address challenges associated with such evolving stances. Additionally, we pinpoint ongoing challenges and outline future avenues in three essential dimensions: utterance, context, and influence. Our goal is to draw a framework that links the current trends in stance detection from an interdisciplinary perspective, covering computational challenges bridging broader linguistics and social science angles.

2.3.1 Computational Perspective on Stance Detection

Research in stance detection has recently attracted an increasing interest [Küçük and Can, 2020], with two main directions. One of the directions includes determining the stance of posts as supporting, denying, querying or commenting on a rumour, which is used as a proxy to predict the likely veracity of the rumour in question [Zubiaga et al., 2016, Zubiaga et al., 2018, Hardalov et al., 2022]. The other direction, which is the focus of this chapter, defines stance detection as a three-way classification task where the stance of each post is one of supporting, opposing or neutral [Augenstein et al., 2016], indicating the viewpoint of a post towards a particular issue. This enables mining public opinion as the

aggregate of stances of a large collection of posts.

In using stance detection to mine public opinion, most research has been operationalised by evaluating on temporally constrained datasets. This presents important limitations when one wants to apply the models on temporally distant test datasets, as recent studies demonstrate. Due to the rapidly evolving nature of social media content, as well as the rapid evolution of people's opinions, a model trained on an old dataset may not perform at the same level on new data [Alkhalifa et al., 2021].

Linguists are interested in understanding human language, which is often dependent on its context [Englebretson, 2007]. The ethnographic definition of stance in everyday language may vary from the academic definition of stance given in the literature [Englebretson, 2007]. Consequently, the definition of stance can be analysed from different perspectives, while most NLP work tends to focus on one of them. The prevalent definition of stance in NLP research stems from a usage-based perspective defined in the field of linguistics and is described by Englebretson [2007] in which stance is dependent on personal belief, evaluation or attitude. Additionally, stance can be seen as the expression of a viewpoint, and it relates to the analysis and interpretation of written or spoken language using lexical, grammatical and phonetic characteristics [Cossette, 1998]. For example, everyday words or phrases used by people during working hours or in performing specific tasks can express subjective features [Cossette, 1998] which can be used by NLP researchers in different applications. However, the stance term may appear and be used differently by researchers as it is strongly relevant to one's own interpretation of the concept.

Stance, as a message conveying the point of view of the communicator, is the opinion from whom one thing is discovered or believed. As a computational task, stance detection is generally defined as that in which a classifier needs to determine if an input text expresses a supporting, neutral or opposing view [Aldayel and Magdy, 2019]. It is framed as a supervised classification task, where labelled instances are used to train a classification model, which is then applied on unseen test data.

While humans can easily infer whether an author is in favour or against a specific event, the task becomes more challenging when performed at scale, due to the need for automated NLP methods. Consequently, the stance detection task has attracted an increasing interest in the scientific community, including scholars from linguistics and communication as well as computational linguistics. However, the need to automate the task by means of NLP methods is still in its infancy with a growing body of ongoing research.

Understanding stance expressed in text is a critical, yet challenging task and it is the main

focus of this chapter. Stance is often implicit and needs to be inferred rather than directly determined from explicit expressions given in the text; indeed the target may not be directly mentioned [Somasundaran and Wiebe, 2009, Mohammad et al., 2016]. However, given the scale of social media data, understanding attitudes and responses of people to different events becomes unmanageable if done manually. Current stance detection approaches leverage machine learning and NLP models to study political and other opinionated issues [Volkova et al., 2016, Al-Ayyoub et al., 2018, D'Andrea et al., 2019, Johnson and Goldwasser, 2016, Lai et al., 2017a]. However, using persuasive writing techniques and word choices [Burgoon et al., 1975] to convey a stance brings important challenges for current state-of-the-art models as there is a need to capture these features in a large-scale dataset. Recent research is increasingly considering pragmatic factors in texts, adopting stance dynamics and the impact of language evolution. Research in this direction can shed light into other dimensions when defining and analysing stance. However, building representations for complex, shifting or problematic meanings is still an open problem that needs exploration.

2.3.2 Capturing Dynamics in Stance Detection

The stance detection task overlaps with, and is closely related to, different classification tasks such as sentiment analysis [Chakraborty et al., 2020], troll detection [Tomaiuolo et al., 2020], rumour and fake news detection [Zubiaga et al., 2018, Rani et al., 2020, Collins et al., 2020], and argument mining [Lawrence and Reed, 2020]. In addition, stance can be impacted by the discursive and dynamic nature of the task [Mohammad et al., 2017, Somasundaran and Wiebe, 2009, Simaki et al., 2017].

In reviewing the literature on stance dynamics, we break down our review into three different dimensions (see Figure 2.1), which cover the different aspects impacting how stance is formed and how it evolves:

- **stance utterance**, referring to a single message conveying a particular stance towards a target.
- **stance context**, referring to the pragmatic, spatiotemporal and diachronic factors that make stance an evolving phenomenon.
- **stance influence**, referring to social factors including the author of a post, as well as reactions towards, and activity around, messages expressing a particular stance.

In what follows, we delve into each of these dimensions and associated literature.

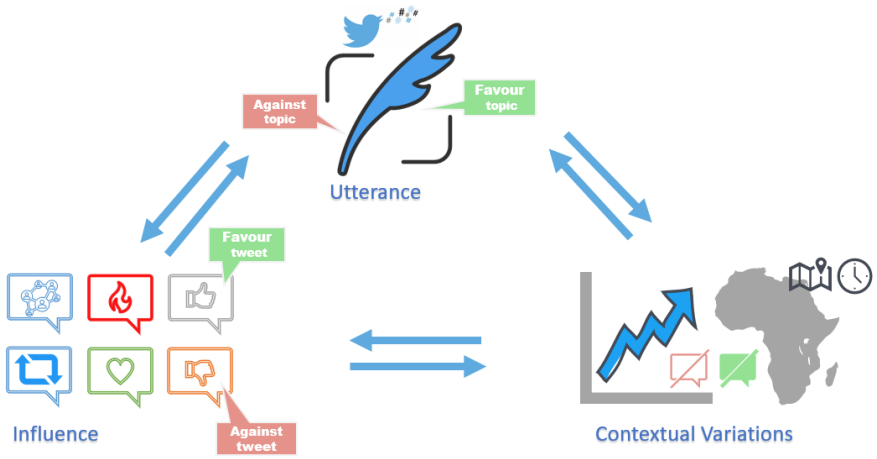


Figure 2.1: Three dimensions around stance detection.

Stance Utterance

Stance utterance refers to the stance expressed in a single message [Mohammad et al., 2016], and reflects human interpretation of an event. It represents the features that form the textual viewpoint and they are essential for human inference and interpretation. Textual data can be analysed based on different features, which previous work have tackled by looking at a range of different challenges, which we discuss next.

One of the challenges in detecting the stance of a single utterance is **target identification**, i.e. determining who or what the stance is referring to. For example, in the utterance "I am supportive of A, but I'm totally against B", the author expresses a supporting stance towards target A and an opposing stance towards target B. The **target** may be implicit and not always directly mentioned in the text [Schaefer and Stede, 2019]. The target may be implicitly referred to [Sobhani et al., 2019], or only aspects of it may be mentioned [Bar-Haim et al., 2017]. These cases present the additional challenge of having to detect the target being referred to in a text prior to detecting the stance. Retrieval of messages likely referring to a target, as a first step to then do the target identification, is a challenge. Achieving high recall in relevant message retrieval can be difficult in the case of implicit messages [Mohammad et al., 2017]. In addition, there is a risk of detecting false positives where a message may not be about the target at all, may not contain data expressing a stance, or may hold multiple stances in the same utterance [Lai et al., 2019, Simaki et al., 2017].

Nuances in the **wording** of an utterance can present another challenge in detecting stance.

Opinions are not always explicitly expressed, and can also be implicit, explicit, ironic, metaphoric, uncertain, etc. [Sun et al., 2016, Al-Ayyoub et al., 2018, Simaki et al., 2018], which make stance detection more challenging. Moreover, surrounding words and symbols can alter the stance of an utterance [Sun et al., 2016], e.g. negating words or ironic emojis inverting the meaning of a text, which are especially challenging to detect. Recent models increasingly make use of more sophisticated linguistic and contextual features to infer stance from text. For example, looking at the **degree of involvement** by using special lexical terms, e.g. slang, jargon, specialist terms, and the informal lexicons associated with social intimacy [Tausczik and Pennebaker, 2010, Hamilton et al., 2016a, Rumshisky et al., 2017, Liu et al., 2015]. Also, use of **embeddings** where concept meanings can be biased and highly impacted by the cultural background and beliefs may lead to varying interpretations [Shoemark et al., 2019, Kutuzov et al., 2018, Hamilton et al., 2016b, Dubossarsky et al., 2017, Xu et al., 2019b].

The **framing** of an utterance can also play an important role in the detection of stance. Framing refers to the adaptation of the wording to convey a specific interpretation of a story to a targeted audience [Walker et al., 2012a] as in language and word choices, this can be seen in the following forms:

- **Reasoning/supporting evidence** about the target or aspects of it [Hasan and Ng, 2014, Addawood and Bashir, 2016, Bar-Haim et al., 2017, Simaki et al., 2017]. For example, Bar-Haim et al. [2017] define the claim stance classification task as consisting of a target, a set of claims and the stance of these claims as either supporting or opposing the target. Further, they simplify the task by looking for sentiment and contrast meaning between a given pair of target phrase and the topic candidate anchor phrase.
- **Attitude**: using single lexical terms holding polarity features related to the author sentiment and reaction to an event. Such word choices can be positive, negative, offensive, harmful, suspicious, aggressive, extremist [Blšták and Rozinajová, 2017]. For instance, emotion and sentiment expressed in the text [Deitrick and Hu, 2013, Xu et al., 2011], which may express the author's view on the importance (or lack thereof) of the target.
- **Persuasion and quality of communication** [Hamilton, 2015, Aune and Kikuchi, 1993] through using grammatically correct sentences. For example, Lai et al. [2019] concluded that people connected to users taking cross-stance attitudes become less polarised and use neutral style when expressing their stance.

Another challenge in stance detection is determining if **stance is present** in an utterance, as the text may be neutral and not opinionated towards the target. This involves determining whether or not an utterance expresses a stance, and subsequently determining the type of stance the author is taking [Mohammad et al., 2016, Zubiaga et al., 2016, Simaki et al., 2017]. For example, the most basic way of stance-taking could be the more extreme positions such as in favour or against to less extreme positions such as asserting, questioning, responding, commanding, advising, and offering which may lead to conversational and threaded stance context [Zubiaga et al., 2016]. In such a context stance moves from its singular utterance structure to its augmenting component (see Section 2.3.2).

Stance Context

Stance context refers to the impact of external factors in an utterance, including the collective viewpoint of a society in relation to the interpretation of a particular target. Context aims to capture the stance occurring in longitudinally evolving contexts, and can be impacted by shifts in opinions over time, locations, or cultures, among others. Any stance inference requires consideration of other points of view, potential stereotypes, as well as how public opinion evolved over time. This represents our understanding of the world dynamics and how stance may change over time [Lai et al., 2019, Volkova et al., 2016]. Stance towards a topic may be considered stable only when it has the same polarity over time. Context involves factors causing changes in public opinion over time, such as real world events. Context constructs complex, shifting or problematic meanings which change the entire view of an event [Azarbonyad et al., 2017, Stewart et al., 2017]. We discuss the two main aspects that are considered when modelling stance context, which include **spatiotemporal changes** and **social changes**.

Collective and individual stance towards a target can be impacted by **spatiotemporal factors** [Volkova et al., 2016, Jackson et al., 2019]. Events occurring in different locations/times get different attention depending on how likely they are to happen again and how unusual they are [Baly et al., 2018a, Hamborg et al., 2019]. Consequently, the audience judging the events have their own biases depending on the cultural and ideological background, which leads to variations in stance across regions.

Even when we restrict geographical locations, there are other factors leading to **social changes** that have an impact on public opinion and stance. Social changes around a topic can lead to shifts in opinions [Volkova et al., 2016, D'Andrea et al., 2019, Lai et al., 2018, Lai et al., 2019]. This can pose a significant challenge, particularly with the tendency in NLP to using distributed representations of words driven by co-occurrence frequency of words using sliding windows, and considering polysemy in more advanced language mod-

els. Words are treated based on their contextual similarity rather than solely based on their isolated frequencies. In order to build these models, one needs large collections of documents with a diverse vocabulary to produce high quality vector representations for different words. Consequently, these models rely on the amount of training data available, and the dimensionality of the word vectors [Mikolov et al., 2013a]. The emergence of out of vocabulary (OOV) words, not seen by these models, can be one of its main limitations. Different methods have been proposed to mitigate these limitations, for example through character-level representations in ELMo or FastText, and sub-word representations in BERT allowing models to incorporate segmented representations for unseen words [Ha et al., 2020]. In prediction models, character-level and subword representations can lead to performance improvements with a trade-off on reduced model explainability; ongoing research is however investigating how to improve model explainability, exploiting for example attention scores produced by BERT [Bodria et al., 2020]. Moreover, newly emerging words or words that shift their meaning over time would lead to outdated models. Challenges relating to social changes can be further broken down into the following:

- **Linguistic shift**, which is defined as slow and regular changes in the core meaning of a word. For example, "the word GAY shifting from meaning CAREFREE to HOMOSEXUAL during the 20th century" [Kutuzov et al., 2018]. This is also reflected in the semantic meaning of emoticons across different contexts, languages and cultures [Robertson et al., 2018]. In multilingual settings, code-mixing of two languages in the same utterance [Khanuja et al., 2020], or borrowing a word from a different language due to influence from other languages, rather than internal changes in the same language.

- **Usage change**, which is the local change of a word's nearest semantic neighbours from one meaning to another, as in the shift of word the "prison CELL to CELL phone" which is more of a cultural change than a semantic change [Hamilton et al., 2016b]. Thus, different viewpoints allow collective stance to change especially when a story is viewed through different eyes and interpreted differently. For example, focusing on UK politics, Azarbonyad et al. [2017] revealed that "The meaning given by Labours to MORAL is shifted from a PHILOSOPHICAL concept to a LIBERAL concept over time. In the same time, the meaning of this word is shifted from a SPIRITUAL concept to a RELIGIOUS concept from the Conservatives' viewpoint. Moreover, two parties gave very different meanings to this word. Also, the meaning of DEMOCRACY is stable over time for both parties. However, Conservatives refer to democracy mostly as a UNITY concept, while Labours associate it with FREEDOM

and SOCIAL JUSTICE."

- **Changes in cultural associations**, which is measured as the distance between two words in the semantic space, as in "IRAQ or SYRIA being associated with the concept of WAR after armed conflicts had started in these countries" [Kutuzov et al., 2018]. Also, the type of sentiment bore by a word can change over time. For example, "the word SLEEP acquiring more negative connotations related to sleep disorders, when comparing its 1960s contexts to its 1990s contexts" [Gulordava and Baroni, 2011]. Moreover, studies have also looked at the relatedness of words over time, by looking at how the strength of the association between words changes. For example, Rosin et al. [2017] introduced a relationship model that supports the task of identifying, given two words (e.g. Obama and president), when they relate most to each other, having longitudinal data collections as input.

Stance Influence

Stance influence refers to the aggregated importance surrounding an individual message expressing a particular stance, and can be measured by using different qualitative and quantitative metrics. These include the author's profile and others' reactions to a message.

Influence defines the quality of an utterance to make an impact, and can vary depending on the popularity and reputation of the author, as well as the virality of a post, among others. Next we discuss three aspects which are relevant to stance influence, i.e. **threading comments**, **network homophily** and **author profile**.

Social media platforms provide a place for conversations to develop, which lead to **threaded conversations** or **tree-structured conversations**. The formation of these conversations enables exchanging viewpoints on top of the initial author's stance [Zubiaga et al., 2016, Guerra et al., 2017, Lai et al., 2019]. For example, Lai et al. [2019] observed that users make use of replies for expressing diverging opinions. Research looking at whether retweeting a post indicates endorsement is so far inconsistent. Lai et al. [2018] observe that people tend to retweet what they agree on. Conversely, Guerra et al. [2017] argued that a retweet does not indicate supporting its underlying opinion.

There is evidence showing that social media users tend to connect and interact with other like-minded users [Lai et al., 2017b, Conover et al., 2012], which is also known as the phenomenon of **network homophily**. Lai et al. [2019] looked at the impact of different characteristics of social media in sharing stance, showing for example that opposing opinions generally occur through replies as rather than through retweets or quotes, polarisation varies over time, e.g. increasing in the proximity of elections.

The identity of the person posting a piece of text expressing a stance, or the **author's profile**, can also play an important role in the development of stance, for example if an influential user expresses an opinion. Two key factors of an author's profile include:

- **Author's ideology and background**, often inferred by observing the user's profile [Elfardy and Diab, 2016, Conover et al., 2012, Yan et al., 2018, Lai et al., 2019], can be used as additional features to determine the stance expressed by a user, rather than solely using the textual content of a post [Mohammad et al., 2016].
- **Author's stance in the temporal space** [Garcia et al., 2015]. For instance, media organisations may express viewpoints through different frames [Hasan and Ng, 2014], which takes time to be assessed [Zubiaga et al., 2016] and may also impact stance evolution and people's stand points. This can also have an impact on how threaded conversations are developed.

2.3.3 Stance Detection Datasets

To study the stance detection task, different datasets covering various topics and pragmatic aspects have been created by researchers.

Table 2.3 shows the list of stance datasets available, along with their key characteristics; these include the time frame they cover, a key aspect in our focus on stance dynamics, as we are interested in identifying the extent to which existing datasets enable this analysis. For ten of the datasets we found, the time frame covered by the data is not indicated (marked in the table as N/A), which suggests that temporal coverage was not the main focus of these works. The rest of the datasets generally cover from a month to a maximum of 1 or 2 years; while the latter provides some more longitudinal coverage, we argue that it is not enough to capture major societal changes. The exception providing a dataset that covers a longer period of time is that by Conforti et al. [2020] and Addawood et al. [2018], with five years' worth of data.

Despite the availability of multiple stance datasets and their ability to solve different generalisability problems (e.g. across targets, languages and domains), this analysis highlights the need for more longitudinal datasets that would enable persistence for temporal stance detection and temporal adaptation, ideally across cultures and languages. For the few datasets that contain some degree of longitudinal content, such as [Conforti et al., 2020] covering 57 months and [Addawood et al., 2018] covering 61 months, the available data is sparsely distributed throughout the entire time period. This again urges the need for more longitudinal datasets, which in turn provide more density for each time period. While data

labelling is expensive and hard to afford at scale, possible solutions may include use of distant supervision [Purver and Battersby, 2012] for data collection and labelling or labelling denser datasets for specific time periods which are temporally distant from each other, despite leaving gaps between the time periods under consideration. Distant supervision has been widely used for other tasks such as sentiment analysis [Go et al., 2009], leading to datasets covering in some cases over 7 years [Yin et al., 2021], however its applicability to stance detection has not been studied as much.

In summary, we observe that existing datasets provide limited resources to capture language dynamics and leverage longitudinal analysis, which would then give rise to more research aiming to capture stance dynamics.

ref.	time frame	months	topics (#)	source	language
Conforti et al. [2020]	Apr 2014 - Dec 2018	57	Finance (2)	Twitter	English
Addawood et al. [2018]	Jan 2012 - Jan 2017	61	Women to Drive Movement (1)	Twitter	Arabic
Rajadesingan and Liu [2014]	Apr 2013	1	US Issues (1)	Twitter	English
Volkova et al. [2016]	Sep 2014 - Mar 2015	7	Politics (1)	VKontakte	Russian Ukrainian
Zubiaga et al. [2016]	Aug 2014 - Oct 2015	15	News events (9)	Twitter	English German
Mohammad et al. [2016]	Jul 2015	1	US Issues (6)	Twitter	English
Schuff et al. [2017]	Jul 2015	1	US Issues (6)	Twitter	English
Simaki et al. [2017]	Jun - Aug 2015	3	Political blogs(1)	the BBC	English
Küçük and Can [2020]	Aug - Sep 2015	2	Sport (2)	Twitter	Turkish
Sobhani et al. [2019]	Oct 2015 - Feb 2016	5	US Issues (4)	Twitter	English
Addawood and Bashir [2016]	Jan - Mar 2016	3	Products (1)	Web	English
Lai et al. [2019], Lai et al. [2018]	Nov - Dec 2016	2	Italian Referend. (1)	Twitter	Italian
Yan et al. [2018]	Jan - Dec 2016	12	US Issues (2)	Twitter	English
D'Andrea et al. [2019]	Sep 2016 - Jun 2017	10	Health (1)	Twitter	Italian
Lozhnikov et al. [2018]	Nov 2017	1	Politics(1)	Twitter Meduza Russia Today	Russian
Baly et al. [2018b]	Jan 2016 - Dec 2017	12	Middle East(1)	News articles	Arabic
Somasundaran and Wiebe [2009]	N/A	N/A	Products (1)	Convinceme	English
Anand et al. [2011]	N/A	N/A	Politics (12)	Convinceme	English
Walker et al. [2012b]	N/A	N/A	US Issues (12)	4forums Createdebate	English
Skanda et al. [2017]	N/A	N/A	Indian Issues (4)	Facebook	Indian
Hercig et al. [2018]	N/A	N/A	Czech Issues (2)	News	Czech
Xu et al. [2016]	N/A	N/A	Different topics (7)	Sina Weibo	Chinese
Ferreira and Vlachos [2016]	N/A	N/A	News Articles (1)	News articles	English
Hasan and Ng [2014]	N/A	N/A	US Issues (4)	Createdebate	English
Bar-Haim et al. [2017]	N/A	N/A	Open domain	IBM dataset	English
Taulé et al. [2018]	N/A	N/A	Catalan Independence (1)	Twitter	Spanish Catalan

Table 2.3: Stance detection datasets, including the time frame covered.

2.3.4 Exploration of Other Longitudinal Datasets

In the previous sections, we examined the extent of dataset coverage for the task of stance detection and sentiment analysis and in chapter 3) we will provide a comprehensive introduction to the datasets utilized in the context of our thesis. Within this section, we aim to draw attention to additional large-scale datasets that have been introduced in the broader body of academic literature for the purpose of temporal analysis and understand longitudinal perception of different domains during our thesis work. These datasets are comprehensively outlined in Table 2.4.

ref.	time Frame	months	topics (#)	source
Lykousas et al. [2019]	Jan 2012 - Jan 2017	61	Emotion and network analysis (705)	Vent
Robertson et al. [2021]	2012-2018	-	Second-order similarity of emoji	Twitter
Suhavi et al. [2022]	Apr 2014 - Dec 2018	57	Mental health disorders (8)	Twitter
Rozado et al. [2022]	2000–2019	-	Sentiment and emotion in news headlines	Twitter, Reddit, student reports, TV
Effrosynidis et al. [2022]	2016-2019	-	Geolocation, user gender, climate, sentiment, aggressiveness, temperature, topic modeling	Twitter
Hofmann et al. [2022]	Jan 2008 - Dec 2019	-	Non/political comments from various perspectives	Reddit

Table 2.4: Summary of different large scale data introduced in the literature

It is important to note that, due to current limitations in data retrieval on Twitter (currently referred to as the ”X” platform), the utilization of unshared datasets by researchers has become increasingly challenging. This difficulty arises from the fact that Twitter APIs are no longer operational without restrictions for researchers. As a result, researchers may be compelled to explore alternative data sources to study social behavior beyond Twitter. This shift highlights the evolving landscape of data accessibility and the need for researchers to adapt their data collection strategies accordingly.

2.4 Identifying Research Gaps and Motivation

In the previous sections, we have discussed the three key factors relevant to stance and impacting its formation and temporal evolution, as well as existing datasets. In what follows, we discuss the main research challenges and set forth our key research directions. We first discuss general challenges, which are broader and impact various NLP tasks, followed by specific challenges related to social media and core challenges specific to stance detection.

2.4.1 General Challenges in Maintaining Temporal Persistence

We also identified gaps in the literature that are not exclusive to stance but have significant impact in stance prediction models such as the impact of **predefined lexicon** word resolution on the model's accuracy [Somasundaran and Wiebe, 2009]. This is especially true when models dependent on a lexicon fail to capture the polarity of evolving words. Research in this direction has used pre-trained word embeddings such as GloVe [Pennington et al., 2014], FastText [Bojanowski et al., 2017], Elmo [Peters et al., 2018], and BERT [Devlin et al., 2019] among others which proved to mitigate the problem of polysemy though word vector representations. This is due to the fact that these models are fed news articles and web data from different sources may be inherently biased [Ruder, 2017]. Moreover, it has been shown that variations of architecture in state-of-the-art language models can significantly impact the performance of the model in downstream tasks. Other work focuses on flipped polarity and negation [Polanyi and Zaenen, 2006]. Even though embedding models consider preceding and following words of a centre word for a given sentence (context), the temporal property of the word itself and its diachronic shift from one meaning to another has not been studied in the context of stance. The identification of diachronic shift of words has however been tackled as a standalone task [Fukuhara et al., 2007, Azarbonyad et al., 2017, Tahmasebi et al., 2021, Shoemark et al., 2019, Dubossarsky et al., 2017, Stewart et al., 2017, Hamilton et al., 2016a, Kutuzov et al., 2018, Hamilton et al., 2016b, Rumshisky et al., 2017]. This is however yet to be explored in specific applications such as stance detection. This may also impact the models' performance across different domains and time frames.

The use of models developed in the field of **NLP** has been barely explored in the context of stance detection, which have been more widely studied for other tasks such as co-reference resolution [Somasundaran and Wiebe, 2009] and named entity recognition [Küçük and Can, 2020, Liu et al., 2013]. Previous research has however highlighted problems in this direction [Lozhnikov et al., 2018, Küçük and Can, 2020, Borges et al., 2019, Sobhani et al., 2019, Lai et al., 2019, Lai et al., 2018, Simaki et al., 2017], which suggests that further exploration and adaptation of NLP models may be of help.

Current deep learning models and the existence of large pre-trained embeddings can offer highly accurate results using training datasets. However, it can lead to biased results when applied to new, **unseen data**, e.g. data pertaining to a different point in time to the one seen during training. This highlights the difficulty of the task and the need to advance research in developing models that are independent of a specific use case and dataset, which can keep evolving as the data changes. Also, there is a need to develop data from different languages to mitigate the cultural biases in existing datasets. This can help detect and explain different perspectives while using specific topics to reason, compare and contrast a model's performance. This would also help further research in stance detection models that are more stable in performance. Moreover, in the case of certain languages, such as Arabic, the use of dialectical language instead of the modern standard language presents an additional challenge. More methods need to be investigated to improve a model's performance considering contextual variation (see Section 2.3.2).

Understanding and detecting **semantic shift**(Section 2.2.2) [Stewart et al., 2017, Rumshisky et al., 2017, Tahmasebi et al., 2021, Shoemark et al., 2019] in the meaning of words has been of much interest in linguistics and related areas of research, including political science, history. However, the majority of this literature focuses their efforts on uncovering language evolution over time, with a dearth of computational research assessing its impact in context-based prediction models such as those using embedding models. Moreover, combining a contextual knowledge using word embeddings in prediction models can help improve performance of stance detection models by leveraging their vector representations. However, current state-of-the-art research ignores the impact of contextual changes due to pragmatic factors such as social and time dimensions when building their models. This may impact a model's performance over time and can result in **outdated datasets and models**. This is due to the dependence of these models to use static data and pre-trained word embeddings to train models. While still training on data pertaining to a particular time period, models need to leverage the evolving nature of language in an unsupervised manner to keep stance detection performance stable. Temporal deterioration of models is however not exclusive to stance detection, and has been demonstrated to have an impact in other NLP tasks such as hate speech detection [Florio et al., 2020]. While some social and linguistic changes may take time [Hamilton et al., 2016c] before they occur, recent literature proved that they may also occur in short periods of time [Shoemark et al., 2019, Azarbonyad et al., 2017]. Most importantly, unlike semantic changes which capture word fluctuations over time, temporal contextual variability may occur in corpus-based predictive models.

2.4.2 Challenges in Social Media and Specific to Temporal Stance Detection

There are numerous open challenges that are specific to the stance detection task. To the best of our knowledge, few studies have specifically focused on the **evolving nature of topics** and its impact on stance detection models. Moreover, fluctuation of word frequencies and distributions over time highlights both the challenge and the importance of the task. Commonalities between the source and target tasks tend to be crucial for successful transfer [Vu et al., 2020]. However, recent NLP models have shifted to transfer learning and domain adaptation where target tasks contain limited training data [Xu et al., 2019a], source data pertains to a different domain [Zhang et al., 2020] or to a different language [Lai et al., 2020]. We anticipate two main directions that would help extend this research: (1) furthering research in transfer learning that looks more into transferring knowledge over time, as opposed to the more widely studied subareas looking into domain adaptation [Ramponi and Plank, 2020] or cross-lingual learning [Lin et al., 2019], and (2) increasing the availability of longitudinal datasets that would enable further exploration of temporal transfer learning.

The majority of existing datasets are from the **domain of politics and to a lesser extent business**, and are hence constrained in terms of topics. Broadening the topics covered in stance datasets should be one of the key directions of research. In this thesis we contribute to this gap by creating a new dataset in the domain of gender equality.

In general, existing datasets cover **short time spans** in languages including English [Ferreira and Vlachos, 2016, Mohammad et al., 2016, Simaki et al., 2017, Somasundaran and Wiebe, 2009, Hercig et al., 2018, Walker et al., 2012b, Anand et al., 2011, Conforti et al., 2020], Arabic [Baly et al., 2018b, Addawood et al., 2018], Italian [Lai et al., 2018], Chinese [Xu et al., 2016], Turkish [Küçük and Can, 2020], Spanish and Catalan [Taulé et al., 2018], Kannada [Skanda et al., 2017], German [Zubiaga et al., 2016], Russian [Lozhnikov et al., 2018]. Recent efforts in multilingual stance classification have also published datasets including German, French and Italian [Mohtarami et al., 2019, Vamvas and Sennrich, 2020], and English, French, Italian, Spanish and Catalan [Lai et al., 2020], but are still limited in terms of the time frame covered. Longitudinal datasets annotated for stance would enable furthering research in this direction by looking into the temporal dynamics of stance. This thesis is also limited to studying the English language in the case of longitudinal datasets, and the study of other languages is left for future work.

The **quality and persistence of the data** are also important challenges that need attention. Annotation of stance is particularly challenging where a single post may contain

multiple targets, or where users change their own stances towards a particular target, i.e. cross-stance attitude. These are challenges that lead to lower inter-annotator agreement and produce confusion even for humans [Lai et al., 2019, Sobhani et al., 2019]. Moreover, relying on social media data under the terms of service of the platforms, reproducibility of some datasets is not always possible [Zubiaga, 2018]. There is also a need for stance detection models that also consider context, for which suitable datasets are lacking. There are also cases where concepts including sentiment, stance and emotion are conflated, with few efforts to define stance [Mohammad et al., 2016, Simaki et al., 2017] or to experimentally prove the difference between these concepts [Mohammad et al., 2017, Aldayel and Magdy, 2019].

The existence of social media accounts run by **bots** leads to fabricated viewpoints of events. These accounts may have been created to manipulate the true view and harm specific targets (for example, businesses or people). This manipulated information can in turn have an impact on specific points in time where the bots operate, and can jeopardise the applicability of stance detection models for certain points in time (e.g. during elections where bot participation may increase) if bots are not detected and removed from the dataset.

Similar to most approaches for social media data, **pragmatic opinions** [Somasundaran and Wiebe, 2009] including short opinions with few lexicon cues can negatively impact prediction performance, including hedging [Somasundaran and Wiebe, 2009], **rhetorical questions** [Hasan and Ng, 2014, Mohammad et al., 2017], **inverse polarity** [Mohammad et al., 2017, Skanda et al., 2017], **sarcasm** [Hasan and Ng, 2014, Skanda et al., 2017], all of which can have a significant impact in the classifier, especially in the case of two-way classification models.

In stance particularly, we can define these problems in four levels: (1) **utterance level** as changing stance from being in favour to being against, (2) **time level** as collective stance [Nguyen et al., 2012] of public pool change from highly in favour to highly against over time, (3) **domain level** where some words change its polarity from one domain to another (such as high prices indicating a favourable stance in the context of a seller but an opposing stance for customers), and (4) **cultural level** which represents stance shift between languages or various geographical locations. Indeed, use of a machine learning model training from old data may not be directly applicable to future datasets, e.g. due to suffering from domain bias, co-variate shift and concept drift. This can be cause by the nature of controversial topics and the impact of pragmatics such as time, location and ideology.

2.5 Conclusion

This chapter providing a background on temporal persistence of text classification models discusses the impact of temporal dynamics in the development of models for classification tasks, with a particular focus on stance detection, by reviewing relevant literature in both stance detection and temporal dynamics of social media. Today's computational models are able to process big data beyond human scale, building on digital humanities and computational linguistics. This however poses a number of challenges when dealing with longitudinally-evolving data. The changes produced by societal and linguistic evolution, among others, both of which are prominent in social media platforms, have significant impact on the shift of social beliefs by means of spreading ideas. With the proliferation of historical social media data and advanced tools, we argue for the need to build models that better capture this contextual change of stance. This thesis makes one of the first such efforts, and arguably the most comprehensive effort, in establishing a benchmark methodology, investigating and quantifying the problem, and studying the mitigation of temporal performance drop in text classification models.

Chapter 3

Temporal Analysis Methodology, Evaluation Metrics, and Longitudinal Datasets

---

Researchers have predominantly adopted a longitudinal perspective to gain deeper insights into the evolution of various social behaviors over time, including sentiment, emotional expression, the use of emojis, and concerns about global warming. Social media posts have emerged as valuable resources for temporally analyzing public perceptions within these diverse domains. However, there is a conspicuous lack of quantitative assessments regarding the persistence of performance scores over time for the machine learning models used and the potential deterioration of model performance as language use evolves over time. This highlights the significance of our research, which not only addresses these issues but also lays the groundwork for future expansions into other tasks.

This chapter aims to investigate the dataset characteristics that may influence model performance over time, particularly when incorporating new data for evaluation (refer to Chapter 4 for more details on dataset characteristics impacting model performance). In this chapter, we introduce the datasets used for our experiments and perform a preliminary analysis of their characteristics. Finally, we provide other longitudinal datasets that can be used by researchers in the future to expand this research work to other tasks and datasets.

3.1 Problem formulation

3.1.1 Problem Statement

In this thesis, our primary focus lies in evaluating the temporal persistence of text classifier performance. Specifically, we aim to assess how effectively a text classifier retains its efficacy when applied to data collected at different points in time. Our investigation encompasses scenarios where the classifier may need to perform on older data than its training corpus or on newer data. To address this challenge, we introduce a novel task known as temporal classification persistence. In this task, we systematically analyze the classifier's performance over time, utilizing a longitudinal, labeled dataset comprising annotated textual data that represents temporal utterances within a specific domain (e.g., gender equality, healthcare). Each text in this dataset carries one of two binary labels, denoted as $S \in \{x1, x2\}$, spanning T years, represented as $Y = \{y_1, ..., y_T\}$. Each y_t corresponds to textual data from year t.

For the purposes of this thesis, we focus on years instead of months or other time periods. This decision is rooted in the practical challenges of collecting and analyzing large-scale, temporally annotated datasets, which are more feasible to manage on a yearly basis. Furthermore, focusing on annual intervals allows us to capture more significant shifts in linguistic usage and societal discourse, which might not be as apparent in shorter time spans.

Our goal is to evaluate the classifier's ability to maintain its performance over time. We adopt a systematic approach in which we train the classifier on data from a specific year, y_i, where $i \in \{1, ..., T\}$, and subsequently assess its performance in each of the subsequent years, y_j, where $j \in \{i+1, ..., T\}$, as well as in years preceding its training data. By aggregating performance scores based on the temporal gap between training and testing sets, we aim to uncover temporal performance trends over time.

We can formally define our temporal classification problem as follows:

Given a temporal annotated corpus from a non-stationary environment (e.g., News, Social Media), denoted as $\mathbf{U}$*, representing temporal utterances within the same domain (e.g., gender equality, health issues), along with corresponding binary labels* $\mathbf{S}$ *and contextualized embeddings* $\mathbf{X}$*, our objective is to develop an optimal approach that enhances a classifier* $\mathbf{F}$*'s performance over time. This entails enabling the classifier to confidently assign labels to unseen texts in temporal test datasets, such as those from year* $t+1$ *or from years older than its training corpus.*

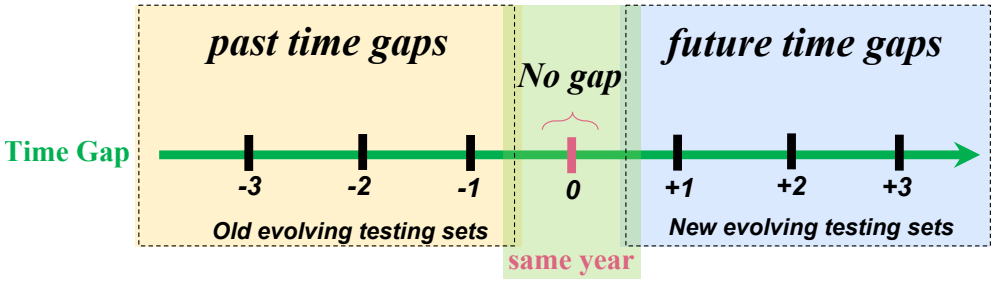


Figure 3.1: An overview of our evolving benchmark settings used for monitoring temporal generalisability of models' performance.

3.1.2 Experimental Settings

To systematically explore temporal classifier persistence, we introduce the concept of the "Temporal Gap". The Temporal Gap represents the temporal difference, measured in years, between training and test data. It accounts for the direction of time, taking on negative values when the test data predates the training data (past) and positive values when the test data is more recent than the training data (future). Figure 3.1 provides an overview of our experimental setup, illustrating different temporal gap scenarios.

Our investigation unfolds across three primary settings:

1. **Learning from Older Training Sets:** This setting involves training on data from previous years.
2. **Learning from Same-Period Training Sets:** Here, the classifier is trained on data from the same temporal period. This is primarily used as a baseline experiment showing the upper-bound performance when there is no temporal gap between the training and test data.
3. **Learning from Newer Training Sets:** This setting explores training on data from more recent years.

The Temporal Gap (G) is formally defined as the difference between the target year for testing (j) and the source year for training (i). We calculate this gap for all possible pairs of years within the dataset, spanning N years, resulting in $G \in [0,N]$. The temporal gap (G) is defined as follows.

$$G = j - i, \forall i, j \in N \tag{3.1}$$

where the resulting $G \in [0,N]$.

In our experiments, we select a source year (i) and a target year (j) and proceed to train the model using the training data from year i (D^i_{tr}). We subsequently evaluate its performance on the target year j (D^j_{ts}), considering both future and past years.

We systematically iterate through all possible source i and target j year combinations within the dataset, which includes N years. Finally, we aggregate the results based on the temporal gap G between years i and j, presenting the average performance metrics for each gap.

i / j	2015	2016	2017	2018	2019
2015	G=0	1	2	3	4
2016	-1	0	1	2	3
2017	-2	-1	0	1	2
2018	-3	-2	-1	0	1
2019	-4	-3	-2	-1	0

Figure 3.2: Illustration of time gaps between training and testing years. Positive values indicate future test years, while negative values indicate past test years.

For instance, if a dataset encompasses data from 2015 to 2019, the temporal gaps range from -4 to 4, representing various combinations of training and testing years. Figure 3.2 illustrates how we calculated the time gaps for every training and test dataset combination, calculated as $G = j - i$ (See 3.1), where i and j represent the training and testing years, respectively, and measure the temporal distance. It accounts for the direction of time, with negative values when testing predates training and positive values for the opposite, as outlined below:

- **G = -4:** 2019-2015
- **G = -3:** 2018-2015, 2019-2016
- **G = -2:** 2017-2015, 2018-2016, 2019-2017
- **G = -1:** 2016-2015, 2017-2016, 2018-2017, 2019-2018
- **G = 0:** 2015-2015, 2016-2016, 2017-2017, 2018-2018, 2019-2019
- **G = 1:** 2015-2016, 2016-2017, 2017-2018, 2018-2019

- **G = 2:** 2015-2017, 2016-2018, 2017, 2019
- **G = 3:** 2015-2018, 2016-2019
- **G = 4:** 2015-2019

When evaluating models, we aggregate performances for different combinations of years pertaining to each gap by averaging them. It is worth noting that this setup leads to varying experiment counts for each temporal gap, with a smaller number of held-out testing sets for larger gaps.

3.1.3 Evaluation Metrics

In our evaluation, we use the following metrics: (1) **Model Evaluation Metric** which measures the model performance (macro-averaged F1-score) for each training and test pair; (2) **Average Performance Metric** which quantifies the average performance of the model (macro-averaged F1-score) by aggregating all testing sets from the same age using the difference in years (Time Gap) between test and training sets. (3) **Relative Performance Drop** which quantifies the actual drop in performance using Average Performance Metric scores using the difference between model performance on the testing set from the Time Gap 0 as an anchor score and on another Time Gap > 0 distant in time.

Our experimental setup is summarised in Algorithm 1, showing the procedure for aggregation of year pairs by their temporal gap. The key metrics defined are:

- **Model Evaluation Metric (MEM):** Uses the macro F1-score to assess the performance between training and test sets for any given model. For each training (D^i_{tr}) and test sets (D^j_{ts}), we use the macro F1-score (F-macro) to assess the performance between a single pair of training and test sets for any given model design choice. F1 score is a weighted average of the recall and precision.

$$MEM(F^i, D^i_{tr}, D^j_{ts}) = F - macro(D^i_{tr}, D^j_{ts}) = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3.2}$$

- **Average Performance Metric (APM):** Averages the F-Macro scores for each time gap, reflecting performance fluctuations over time. As such, we quantify the performance fluctuation for each time gap by averaging F-Macro over all training and test f-score pairs with the same temporal distance. We use an evolving longitudinal benchmark with multiple experiments for each time gap to accurately measure the temporal performance score for each task.

Algorithm 1 Temporal Performance Evaluation

Input: Longitudinal training sets, development sets, and test sets D_{tr}, D_{dv}, and D_{ts} for each dataset D covering N temporal years, language representations R, and classification algorithms C.

Output: Average performance score for all possible time gaps between training source years $i \in I$ and test target years $j \in J$ where ($I \subset N$ and $J \subset N$). ▷ We use N for both training source years I and test target years J as they are equal sets of all N years in our settings.

1: **for** all $F \in Models^{R\&C}$ **do**
2: **for** all training sets D_{tr} from source year $i \in N$ **do**
3: Train F using D^i_{tr} to get a classifier F^i
4: Apply early stopping strategy using held-out D^i_{dv}
5: **for** all testing sets D_{ts} in target year $j \in N$ **do**
6: Predict classes for D^j_{ts} using F^i
7: Calculate the **Temporal Gap** (See Equation 3.1) to the temporal distance between training and test set $(j-i)$ to determine the temporal distance of classifier F^i predictions for given held-out test set.
8: Calculate **Model Evaluation Metric** (See Equation 3.2) to evaluate the performance D^j_{ts} of the given classifier using the classifier F^i.
9: **end for**
10: **end for**
11: Compute **Average Performance Metric** for all years per time gap (See Equation 3.3).
12: Compute **Relative Performance Drop** between each time gaps and time gap 0 (See Equation 3.4).
13: **end for**

$$APM(G) = \overline{\sum_{i \in N} \sum_{j \in N} F - macro(D^i_{tr}, D^j_{ts})} \text{ if} j - i = G \tag{3.3}$$

- **Relative Performance Drop (RPD):** Quantifies performance drop comparing APM at time gap 0 to APMs at subsequent time gaps. We quantify the performance drop using the difference of APM(0) when time gap is 0 ($G = 0$) compared to APM(G) score of another time gap that is temporal distance than time gap 0 ($G > 0$). Each RPD measures the temporal performance drop in model's performance using all APMs of $G > 0$ compared to initial year's APM model score at $G = 0$.

$$RPD(APM(0), APM(G)) = \frac{APM(G) - APM(0)}{APM(0)} \text{ when} G > 0 \tag{3.4}$$

This framework enables us to systematically evaluate the temporal persistence of NLP classification models, crucial for applications where data evolves over time. In summary, our study introduces the concept of the temporal gap to systematically investigate the persistence of NLP classification models across different temporal scenarios. We analyze the performance of these models under varying temporal gaps, allowing us to gain insights into their adaptability to changing data distributions over time. We employ three key metrics: the **Model Evaluation Metric**, which measures performance for each training and test pair; the **Average Performance Metric**, aggregating performance across all testing sets of the same temporal distance; and the **Relative Performance Drop**, quantifying the performance decline over different time gaps. This comprehensive approach enables us to assess the persistence and temporal dynamics of NLP models in a wide range of real-world scenarios. The detailed experimental procedure is outlined in Algorithm 1, providing a robust framework for our investigation.

3.2 Collection of Longitudinal Datasets

In this thesis, we use five datasets pertaining to three different tasks. We collected four new datasets and borrowed one from previous work. Our selected datasets come from two sources: social media and e-commerce platforms. These sources contain user-generated content that reflects individual opinions, preferences, and stances. Analyzing datasets from both domains offers a more comprehensive understanding of how stance and sentiment are expressed in different online contexts and how they evolve over time. Additionally, using a diverse range of datasets across multiple tasks enhances the generalizability of our experiments.

We collect and label the following datasets for the following tasks:

- **Stance Detection (SD),** including the Gender Equality Stance Detection (GESD) and HealthCare Stance Detection (HCSD) datasets.
- **Temporal Sentiment Analysis (TSA),** including the Temporal English Sentiment Analysis (TESA) and LongEval-Temporal English Sentiment Analysis (LE-TESA) datasets.
- **Review Rating Prediction (RRP),** which includes the Amazon Books Rating Reviews (**ABRR**) dataset originally collected by Ni et al. [2019].

Given the need for longitudinal, labeled datasets covering several years, we opted for relying on distant supervision for the collection and labeling of datasets. Distant supervision

has become a widely recognized and effective method for collecting labeled datasets from social media sources, with a primary focus on sentiment analysis [Go et al., 2009]. This approach has more recently been extended to encompass a variety of other tasks, including stance classification [Kumar, 2018, Mohammad et al., 2017]. Essentially, distant supervision consists in automatically or semi-automatically labeling datasets by using certain signals inherent in the data which give a hint on the labels, while then removing those signals from the data to avoid data contamination in the dataset. For example, one can rely on the fact that a tweet contains a happy or sad emoji face to determine whether the tweet conveys positive or negative sentiment; while the emoji can help determine the sentiment label of the tweet, these emojis can then be stripped from the final dataset, which would contain the text of the tweet without the emoji and its associated label. This method can help obtain large-scale labeled datasets at low or no cost, with the caveat that labels can have some levels of noise as there can be, for example, cases where happy or sad emojis can be used ironically. In this thesis, while relying on distant supervision, we also perform manual checks on small samples of the distantly labeled datasets to estimate the amount of noise in the datasets.

Moreover, our datasets encompass both labeled and unlabeled data, offering distinct advantages for our research. While labeled data allows us to directly evaluate the performance of our models through testing, unlabeled data serves a crucial role in model adaptation. By leveraging unlabeled data, we can employ techniques such as incremental learning, which utilizes unlabeled data to update model lexicons and adapt to evolving language patterns. This approach is particularly valuable in scenarios where acquiring labeled data is expensive or impractical, especially when the language model is fine-tuned with outdated training data, where however obtaining large amounts of unlabeled data is generally cheap and affordable.

In what follows we describe the data collection and labelling methodology we follow for the five datasets pertaining to the three tasks.

3.2.1 Stance Detection (SD) datasets: GESD and HCSD.

Due to the lack of large-scale temporally annotated datasets for stance classification, we collected new datasets using distance supervision. In this case for a six-year time period from 2014 to 2019, the data collection is based on predominantly supporting or opposing hashtags trends over years though manual search for relatedness to gender equality and healthcare posts. Labeling involved distant supervision, employing opinionated hashtags to categorize tweets as “support” or “oppose,” enhancing its utility for research [Alkhalifa

et al., 2021].

GESD Dataset: Gender Equality	
Support	#MeToo, #feminist, #Feminism, #WomenEqualityDay, #WomensRights, #GenderEquality, #WomenEmpowerment, #truefeminism, #Imatter, #GirlsInCrisis, #ProtectPregnantWorkers, #equalityinsports, #GenderPayGap, #StopEnslavingWomen, #alienatedmother, #WomenPeaceSecurity
Oppose	#MenToo, #MensRights, #FakeFeminism, #GenderBiasedLaws, #fakefeminist, #CrimeHasNoGender, #CrimeByWomen, #CrimeAgainstMen, #ConspiracyHasNoGender, #MenToo_UntoldTruth, #ToxicFemininity, #MenWillBeMen, #alienatedfather, #submissivewife, #feminismiscancer, #equalparenting
HCSD Dataset: Healthcare	
Support	#fitness, #eatright, #weightloss, #gamusa_day, #dite, #wightloss, #patientadvocacy, #4Patients, #Caregivers, #ThinkGP, #Doctors, #patientvoiceheard, #healthandwellness, #VaccinesSaveLives, #Gprecruitment, #GPForwardView, #TeamGP
Oppose	#Carb, #Fat, #bellylover, #dontdiet, #Gainer, #endweightstigma, #ContaminatedBlood, #medicalnegligence, #Insurancemisconduct, #clinicalnegligence, #shadowdoctors, #staffAbovePatients, #PatientEmpowerment, #VaccinesareNOTsafe, #MedicalIndemnity, #MedicalMalpractice, #patientsafety

Table 3.1: List of hashtags used to build our datasets.

This resulted in two Twitter stance datasets by using hashtags spanning the same time period (2014-2019) constructed using hashtags shown in Table 3.1 [1]: (1) with hashtags supporting and opposing gender equality (GESD), involving issues such as feminism and gender pay gap, and (2) with hashtags supporting and opposing the use of healthcare (HCSD), involving issues such as dieting and medical care. For both healthcare and gender equality, our stance detection approach discerns between support and opposition. In healthcare, hashtags represent advocacy for medical practices or critique healthcare systems, reflecting varied public opinions. In gender equality, we distinguish between hashtags that support movements like #MeToo and #Feminism, promoting rights and equality, and opposing hashtags such as #FakeFeminism and #MensRights, stemming from countermovements or critiques of feminism. These opposing hashtags are crucial for a balanced

[1] `https://github.com/OpinionChange2021/opinion_are_made_to_be_changed.git`

	Source		Target	Labels	
	Train.	**Dev.**	**Test.**	**% Class 1**	**% Class 2**
GESD	35,100	3,900	9,000	76.9% (Support)	23.1% (Oppose)
HCSD	22,500	2,500	5,040	53.6% (Support)	46.4% (Oppose)
TESA	74,850	9,980	14,970	50% (Pos)	50% (Neg)
ABRR	15,840	2,620	3,930	50% (5)	50% (1)

Table 3.2: Dataset Statistics for GESD, HCSD, TESA, and ABRR. Source and target tweet counts per year with consistent temporal label distribution.

analysis, capturing the dynamic nature of language use within gender equality and healthcare posts on social media overtime. These collected datasets acquired via the Twitter API using relevant hashtags. Historical tweet retrieval was facilitated by the user-friendly GetOldTweets3 API, allowing researchers to specify search parameters, including hashtags and date ranges.

Our hashtag compilation was manually incremental, starting with key hashtags and expanding through manual "snowballing". This method uncovered a wide spectrum of hashtags, ensuring our datasets accurately reflect diverse public discussions on these issues, including essential opposing viewpoints in gender equality and healthcare for comprehensive social media discourse analysis.

Unlabelled Datasets. We collected more extensive domain-specific Twitter datasets pertaining to the topics of gender equality and healthcare. These datasets were constructed using the same hashtags as the labeled datasets which will be described in the following section. However, in this case, we intentionally omitted labels and removed hashtags from the text to avoid introducing any form of supervision during the training of word embedding models. As a result, we compiled a total of 578K tweets related to gender equality and 343K tweets related to healthcare. It is worth noting that these datasets played a crucial role in the development of our temporal adaptive model, which will be explored in detail in Chapter 5.

Labelled Datasets. Distant supervision consists in defining a set of keywords (e.g. hashtags) which serve as a proxy to data labels, subsequently removing these keywords from the resulting dataset and leaving the rest of the text of the posts. To assess the quality of the distantly supervised labels, we manually inspected a subset of 225 random tweets from the resulting datasets. We observed that only 11% of the instances are noisy, i.e. opposite stance. This is in line with previous work on distant supervision (cf. [Purver and Battersby,

2012]), and hence continue with this dataset given the inevitably trade off between dataset size and label quality. We then randomly selected a stratified sample from each year based on minimum count of each label, which is split then into train, evaluation and test data. Table 3.2 shows the per-year statistics of the resulting datasets.

3.2.2 Temporal Sentiment Analysis (TSA).

For the temporal sentiment analysis task, we leveraged distant supervision to collect large collections of tweets spanning a seven-year time period from 2013 to 2020 [Yin et al., 2021]. Collected tweets, spanning from January 2013 to June 2020, are accessible through the Twitter Stream Grab (TSG) project on the Internet Archive[2], curated from the archived collection of Twitter's 1% public stream of tweets. This compilation amounts to 3.8TB of tweets in compressed (bz2) format.

These tweets were initially sampled for seven different languages: Arabic (ar), German (de), English (en), Spanish (es), French (fr), Italian (it) and Chinese (zh). While the full dataset for all seven languages was published, for the purposes of this thesis we focused on the English language subset of the dataset.

We enhanced the **emoticon** list from Go et al. [2009] with those used by Byrkjeland et al. [2018] for sentiment embedding training. This list was refined against manually labeled datasets and SemEval Twitter sentiment datasets (2013-2017) [Rosenthal et al., 2017], excluding ambiguous emoticons for better suitability in distant supervision.

For **emojis**, we initiated our list from all available emojis as of September 2020, selectively including only full faces or gestures to preclude cultural or gender biases. Through rigorous assessment based on Emojipedia[3] descriptions, we ensured the inclusion of emojis that explicitly expressed positive or negative emotions, excluding those related to external factors for clarity.

We evaluated the quality and the extent of noise in the resulting distantly labelled datasets. We did this by matching the aggregation of *SemEval* sentiment analysis tweets from 2013-2017 shared tasks Rosenthal et al. [2017] based on the predicted emotion by the distantly supervised approach in comparison with the manually annotated labels as shown in Table 3.4. While the manual labels are three-way (positive, neutral, negative), the distantly supervised approach only captures two different categories (positive and negative). This shows the difference in the inclusion thresholds used – the manually labelled "neutral"

[2] `https://archive.org/details/twitterstream`
[3] `https://emojipedia.org/`

Emoticons				Emojis
Positive				
:)	:-)	;D	:D	😀 😃 😄 😁
=D	:-]	:]	:-3	😆 ☺️ 😊 😇
:3	:->	:>	8-)	😌 😍 🤩 🥰
:-}	:}	:o)	:c)	😘 😗 😚 😙
:ʔ)	=]	=)	:-D	😋 😝 🤗 🙄
8--D	x--D	xD	X--D	🤤 🥳 😎 😺
XD	=D	=3	B^D	😸 😻 😽 ✌️
:-))	:'--)	:')	;--)	👍
;)	*-)	*)	;--]	
;]	;ʔ)	:--,	<3	
Negative				
:(	:-(	: (	:'(	🙃 😐 😑 😶
:L	=L	:-c	:c	😒 🙄 😬 😔
:--<	:<	:--[	:[	🤢 🤮 😵 😕
:-||	>:[	:{	:@	😟 ☹️ ☹️ 😦
>:(	D-':	D:<	D:	😧 😨 😰 😢
D;	D=	:--/	:/	😭 👎 😖 😣
:--.	>:\	>:/	:\	😞 😓 😩 😫
=/	=\	>.<	v.v	😡 😠 🤬 😿
:S	</3	<\3		🖕

Table 3.3: Positive and negative sentiment keywords: (1) Emoticons including 40 positive and 35 negative expressions, and (2) Emojis including 29 positive and 33 negative expressions.

instances that our distant supervision captured with polarity mainly consists of mildly positive or mildly negative ones.

Considering the rest of the instances that were captured, only few were noise (coloured red), showing the effectiveness of the distantly supervised method to automatically label tweets with little noise. These noise instances have similar distributions of emoticons as the valid ones, showing ironic use. Users were more likely to ironically use positive emoticons in a negative setting than vice versa.

This dataset serves as a foundational resource for conducting comprehensive longitudinal analyses, making use of the aggregated sentiment labels derived from emojis and emoticons. Allowing our distantly supervised approach to cover a total of 137 balanced sentiment keywords: 69 positive and 68 negative expressions. This curated list comprises 137 sentiment keywords, evenly distributed between positive and negative expressions, enriching

	positive	neutral	negative
positive	510	95	19
negative	9	27	71

Table 3.4: Manual labels (columns) vs distantly supervised labels (rows) on the subset of SemEval tweets matching emoticons or emojis.

our dataset for accurate sentiment analysis as detailed in Table 3.3.

In the subsequent paragraphs, we will delve into further details regarding the Temporal English Sentiment Analysis (TESA) component employed within this thesis and used in Chapter 4 where we used distantly annotated texts sets and Chapter 6 where we introduced LongEval TESA (LE-TESA) human-annotated evaluation sets.

Temporal English Sentiment Analysis (TESA).

TESA Labelled Datasets. For TESA training and testing sets, we sample the same amount of data and preserve the same distribution of labels for each year, which helps us avoid other confounding factors to solely focus on the impact of temporal change. The dataset used after removing Emojis and emoticons provided by Yin et al. [2021] and keeping tweets with a 3-word minimum length. In all the settings, we applied a stratified split with 75%, 10%, 15% for train, development and test per year respectively. The resulting distribution of data is shown in Table 3.2.

LongEval Temporal English Sentiment Analysis (LE-TESA).

LE-TESA was introduced as part of the evaluation shared task discussed in Chapter 4. The LE-TESA dataset is distinctive from original TESA in that it comprises both human-annotated testing sets and distantly annotated training sets. This dataset spans eight years from 2014 to 2021 and consists of tweets with binary sentiments, categorized as either “positive” or “negative”.

LE-TESA Unlabelled Datasets. The LE-TESA dataset underwent rigorous preprocessing to ensure quality. Duplicates and near duplicates were removed. Also, to enforced a diversity of users and removed tweets from most frequent users with bot-like behaviour. Finally, user mentions were replaced by ’@user’ for anonymization, except for verified users that remained unchanged. For all these preprocessing steps, we relied on the same pipeline and script used by Loureiro et al. [2022]. On average, each year in the dataset comprises approximately 119,999 lines of data after prepossessing, with a slight variation observed in some years. For instance, the year 2019 contains slightly fewer lines, with an average count

of 119,997. The dataset is publicly available.[4]

Dataset	Time Period	Size
Training	Feb 2014 - Dec 2016	49,608
Practice-2016 [within]	Jan 2016 - Dec 2016	1,344
Practice-2018 [distant]	Jan 2018 - Dec 2018	1,344
Test-within	Jan 2016 - Dec 2016	908
Test-short	Jan 2018 - Dec 2018	908
Test-long	Jan 2021 - Aug 2021	908

Table 3.5: LE-TESA Dataset statistics summary of training, practice and testing sets.

LE-TESA Labelled Datasets. We began with a sample excluded from the **LE-TESA unlabelled dataset**, and further preprocessing steps were introduced. These steps included the exclusion of all retweets and replies, as well as prioritizing English posts with a 5-word minimum and a 140-character maximum. Posts with at least one stop word were also selected. Emojis and emoticons provided by Yin et al. [2021] were removed prior to sampling. Subsequently, the dataset was systematically sampled from the preprocessed data to create a balanced annotated set, considering sentiment distribution, the month of tweet creation, and normality in tweet length distribution.

The resulting distribution of data is shown in Table 3.5. The LE-TESA **training** set covered a two-year period, from 2014 to 2016. For the LE-TESA practice sets, we introduced both "within" and "distant" time sets. The **Practice-2016** set had a temporal gap of 0 years from the training data, given that it overlapped with the training period. Additionally, the **Practice-2018** set was provided as a distant test set for practice, featuring a temporal gap of two years from the training data. In terms of evaluation sets, the **within** set shared a 0-year time gap, covering the same period as the within **Practice-2016** set. The **Test-short** set had a 2-year time gap, aligning with the distant **Practice-2018** set. Lastly, the **Test-long** set had a 5-year time gap, representing a long-term evaluation scenario.

The **Practice and evaluation sets annotation** testing sets were annotated using Amazon Mechanical Turk (AMT) [5]. Workers on AMT underwent a selection process involving two qualification tasks. The first task aimed to identify experienced workers located in English-speaking countries, ensuring language proficiency and familiarity with AMT. The second task presented candidates with 5 tweets, allowing only those correctly annotating

[4] `https://colab.research.google.com/drive/1gOwCqtRRNcYsNozRYeF_2pAl8VAoGulv?usp=sharing`

[5] `https://www.mturk.com/`

3 or more to proceed to the actual annotation task. In total, 4,032 tweets were annotated, comprising 1,874 positive, 741 neutral, and 1,417 negative examples. Each tweet received annotations from 5 different workers, and the final label was determined by calculating the mode of the annotations.

3.2.3 Amazon Books Rating Reviews (ABRR).

Total Reviews	38,480,365
Average Reviews per Year	2,137,798
Total '5' Ratings	30,581,057
Total '1' Ratings	1,146,047
Average '5' Ratings per Year	1,698,947
Average '1' Ratings per Year	80,336

Table 3.6: Basic Statistics of Amazon Book Reviews Dataset (2000-2018)

This is based on the Amazon Product Reviews dataset [Ni et al., 2019][6]. The original dataset contains reviews and their corresponding rating scores across various product categories on Amazon. Our study, however, narrowed its scope to book reviews, specifically from 2000 to 2018. Even though the Amazon book reviews dataset spans from 1996 to 2018, the decision to exclude data preceding the year 2000 was guided by the need for a sufficient number of labels to facilitate cross-temporal testing across consecutive years. Table 3.6 presents general statistics of the dataset, including the total number of reviews, the average number of reviews per year, and the total counts of '5' and '1' ratings, along with their respective averages per year.

Labelled Datasets. Data selection includes review rating scores assigned by the users themselves along with the review texts. We used two columns from the dataset, **text** of the review and **overall rating** of the product. The original ratings in the dataset range from 1 to 5; in order to frame the task as a binary problem, we sample the reviews rated as either 1 or 5, removing the remainder of the reviews with scores 2, 3 or 4, as well as duplicates and empty lines with simple reprocessing. The data resulting from this is shown in Table 3.7. By using this approach, we ensured that our research remained focused within the context of binary classification, allowing for a more in-depth exploration of the temporal persistence of binary classifiers.

To ensure balanced representation, both sentiment categories were downsampled to target

[6] `https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews`

Year	Total Reviews	'5' Count	'1' Count
2000	348,471	213,344	18,287
2001	310,029	185,754	17,220
2002	297,913	177,404	17,658
2003	299,921	177,511	19,689
2004	345,949	201,635	26,842
2005	494,120	286,634	41,262
2006	551,745	328,487	40,854
2007	733,532	451,180	43,147
2008	787,699	477,736	48,953
2009	972,338	590,288	61,494
2010	1,114,410	633,082	66,567
2011	1,436,114	872,015	90,887
2012	2,530,939	1,574,994	127,637
2013	6,037,968	3,864,774	210,269
2014	8,225,749	5,401,730	284,022
2015	8,398,852	5,659,919	297,847
2016	8,209,061	5,627,440	291,376
2017	7,281,384	5,121,207	266,588
2018	2,743,746	1,968,514	103,755

Table 3.7: Analysis of '5' and '1' Ratings in Amazon Book Reviews (2000-2018)

sizes of 7,920, 1,965, and 1,310 instances for each label within the training, testing, and evaluation sets, respectively. This strategy resulted in a total of 22,390 instances per year across all splits. Further details will be discussed in the following chapters. This approach ensured that the dataset maintained a balanced dataset with suitable size for testing, facilitating fair and unbiased analysis throughout the entire temporal analysis over classification expressions.

3.3 Analysis of the temporal dynamics of language use

We look at the word frequency in the datasets over time, i.e. how does word usage persist over time and to what extent is word usage ephemeral. Figure 3.3 shows the statistics for different types of words according to their lifetime, which we define as follows:

- **Dying words (red):** Words that, having been used for 2+ years, are no longer used in future years.
- **Unique words (gray):** Words only used in that year and not in any other year.

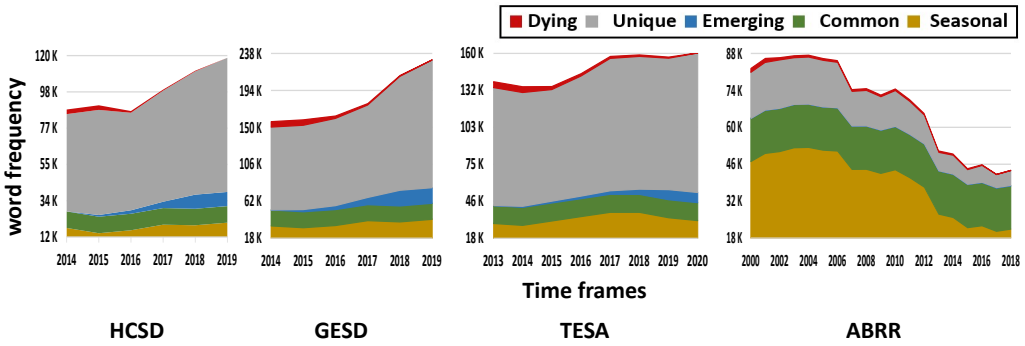


Figure 3.3: Temporal usage of different word types. Dying words (red), unique words (gray), emerging words (blue), common words (green) and seasonal words (brown). See main text for descriptions.

- **Emerging words (blue):** New words in that year which had not been used in the past.
- **Common words (green):** Words consistently used in all the years. Note that the set of common words is the same across all years, and hence absolute values of common words don't change.
- **Seasonal words (brown):** Words used in 2+ years but not in all the years.

The patterns observed in this analysis lead to the following hypotheses for our experiments:

- **Total number of words.** We observe an increase of vocabulary size for the social media datasets (GESD, TESA, and HCSD), and a decrease for the ABRR dataset of reviews. This may be due to the nature of the data, i.e., ABRR coming from a more constrained domain (book reviews), whereas the social media domain attracts more diverse participants and is more dynamic and informal. The vocabulary increase is particularly expected for TESA and HCSD, which are not restricted by a particular domain and cover a broad range of topics. We hypothesize that these patterns in word counts may have an impact on temporal model performance.
- **Emerging and dying words.** We observe that the proportions of emerging and dying words in the vocabulary of the ABRR dataset are very small. This suggests that the constrained nature of book reviews leads to a more established vocabulary, which varies to a lesser extent between years. However, emerging and dying words are more prominent in social media datasets (GESD, TESA, and HCSD). This shows a higher vocabulary variation in these datasets, which could lead to challenges for the

models to classify instances from these datasets over time. While one may initially expect a higher variation in the TESA dataset due to its domain independence, the percentages of emerging and dying words are especially high for the GESD dataset. We believe that the larger variation in GESD can be explained by the evolving nature of the gender equality domain, where people's opinions are likely to evolve over time, leading to a larger vocabulary variation. Hence, we hypothesize that GESD might be a dataset (and a task setup) where models might more prominently drop in performance over time, and more regular model updates may be required to achieve persistent performance. The HCSD dataset exhibits similar patterns.

- **Unique and seasonal words.** We observe that the proportion of words of ephemeral nature, such as unique and seasonal words, does not vary much in the social media datasets TESA, GESD, and HCSD. Unique words are frequent, covering approximately 60% of words in the vocabulary, whereas seasonal words are rare. Interestingly, ABRR shows a very different trend. Seasonal words are more frequent than unique words, with both showing a decreasing tendency over time. The decreasing vocabulary size over time in ABRR is likely a contributing factor, i.e., a smaller vocabulary over time leads to more common words and fewer ephemeral words.

  Seasonal words We observe that some words are used in multiple years but not consistently across all years. These words can be considered "seasonal" in the sense that they are not constant but have periods of activity followed by inactivity. The presence of seasonal words in the vocabulary of social media datasets suggests that certain terms or phrases may become more relevant during specific periods and then lose relevance in subsequent years. Models trained on these datasets will need to account for the temporality of these words and potentially adapt to their changing usage patterns.

- **Common words.** The set of common words used consistently across all years is larger for ABRR than for GESD, TESA, and HCSD. This is remarkable given that ABRR is the dataset covering the longest temporal period (19 years), which reduces the likelihood of words to consistently occur annually over such a long period of time. However, we believe that the more restricted nature of book reviews leads to such vocabulary consistency.

In summary, the analysis of word usage patterns over time reveals that social media datasets exhibit dynamic vocabulary changes, including the emergence and disappearance of words and the presence of seasonal terms. These observations suggest potential challenges for NLP

models when classifying text over time, as they may need to adapt to shifting language use patterns. Our experiments in the following chapter will explore how well models handle these challenges and whether their performance is influenced by changes in word usage over time.

3.4 Conclusion

In this chapter, we have laid the foundation for understanding the temporal dynamics in NLP classification tasks by presenting our methodology, evaluation metrics, and the longitudinal datasets utilized in our research. We introduced five datasets across three tasks, highlighting the significance of using both social media and e-commerce data to capture a diverse range of user-generated content. This approach not only enhances the generalizability of our experiments but also provides a comprehensive understanding of how stance and sentiment are expressed and evolve in different online contexts.

We have also discussed the concept of temporal classification persistence and introduced the notion of the temporal gap, which allows us to systematically analyze the performance of text classifiers over time. Our evaluation framework, including metrics such as Model Evaluation Metric (MEM), Average Performance Metric (APM), and Relative Performance Drop (RPD), provides a robust mechanism to assess the temporal stability of NLP models. Furthermore, we explored the characteristics of our datasets, including the frequency and nature of word usage over time. This preliminary analysis revealed dynamic vocabulary changes in social media datasets and more stable vocabulary patterns in e-commerce reviews, underscoring the different challenges posed by each domain.

The subsequent chapters will build on this foundation, using the introduced datasets and methodologies to empirically investigate the temporal persistence of text classifiers. We will analyze how well models adapt to shifting language patterns and explore strategies to enhance their temporal persistence, ultimately contributing to the development of more robust and reliable NLP text classifiers.

Chapter 4

Assessing the temporal persistence of text classifiers.

Performance of text classification models tends to drop over time due to changes in data, which limits the lifetime of a pretrained model. Therefore an ability to predict a model's ability to persist over time can help design models that can be effectively used over a longer period of time. In this chapter, we look at this problem from a practical perspective by assessing the ability of a wide range of language models and classification algorithms to persist over time, as well as how dataset characteristics can help predict the temporal stability of different models. We perform longitudinal classification experiments on three datasets spanning between 6 and 19 years, and involving diverse tasks and types of data. We find that one can estimate how a model will retain its performance over time based on (i) how well the model performs over a restricted time period and its extrapolation to a longer time period, and (ii) the linguistic characteristics of the dataset, such as the familiarity score between subsets from different years. Findings from these experiments have important implications for the design of text classification models with the aim of preserving performance over time.

- We shed light into the temporal persistence of existing language models.
- We analyse when and why model performance drops over time, which informs when a model needs adapting.
- We investigate the impact of classification model choice in cross-temporal perfor-

mance.

- We analyse the impact of the dataset properties on performance drop over time.
- We assess the potential and limitations of contextual language models to improve temporal persistence.

4.1 Introduction

A supervised text classification model relies on labelled datasets to train the model [Sebastiani, 2002, Cunha et al., 2021]. From an experimental perspective, the design and evaluation of classification models typically rely on data pertaining to fixed periods of time. Recent research demonstrates that such models, while showing competitive performance in their experimental environment, underperform when they need to classify new data that is distant in time from that observed during training [Alkhalifa and Zubiaga, 2022]. This deterioration of performance has been demonstrated for different classification tasks, including topic classification [Rocha et al., 2008], sentiment classification [Lukes and Søgaard, 2018], hate speech detection [Florio et al., 2020], stance detection [Alkhalifa et al., 2021] and political ideology detection [Röttger and Pierrehumbert, 2021]. This performance drop can happen for multiple reasons, including among others the evolution in language use [Smith, 2004] or the evolution of public opinion [Claassen and Highton, 2006, Bonilla and Mo, 2019] and its extent may vary [Alkhalifa et al., 2021]. This poses an important challenge and limitation on such models when one plans to continue using the model over a long period of time to classify new, incoming data, as can be the case with a stream of user-generated contents [Cheng et al., 2021].

Despite this evidence of performance deterioration over time, previous research hasn't explored the nature of this deterioration, i.e. when, how and why it occurs, such that it could inform design and maintenance of text classification models which can continue to be used as effectively as possible over time. Our study fills this gap by performing a comprehensive study into model performance over time, i.e. by keeping all variables in the experiments fixed, where the only factor that changes is time, with the associated evolution of data over this time. This helps us shed light into the causes of performance deterioration, as well as to devise possible solutions to mitigate this deterioration.

In this chapter, by using three large-scale, longitudinal text classification datasets involving user-generated content, we perform a set of experiments to learn more about the temporal persistence of text classifiers. In these experiments, we look at the problem from different angles including embedding models used for language representation, algorithms used for

the classification and underlying characteristics of the datasets. We discuss how these factors impact model performance over time. Our study focuses on quantifying the impact of these different factors (representation models, classification algorithms, datasets) as well as helping understand its implications for the design of temporally-persistent text classification models. Our work is the first to delve into the problem of temporal persistence of text classification models, shedding light into the development of temporally persistent models.

Through this study, we tackle the research aims defined in Section 4.1.1 by focusing on the research questions in Section 4.1.2, and make the novel contributions discussed in Section 4.1.3.

4.1.1 Research aims

The focus of this study is on delving into the understanding of the impact that the temporal evolution of datasets has in text classification tasks, in a way that can inform future design of classifiers with the stability of temporal performance in mind. To do so, our study has two overarching aims: (1) assessing how different factors of the datasets and models affect performance over time, and (2) where one only has access to annotated data covering a small timeframe, determining whether one can predict the temporal performance that different classifiers will exhibit over time. An ability to design classifiers after only seeing a short timeframe of annotated data can be very important to enable design of temporally robust classifiers where annotation of longitudinal data is costly and unaffordable.

To address these two larger aims, we break down our research into six smaller research objectives:

1. assessing the temporal persistence of existing **language models**, quantifying their performance drop in different situations,
2. investigating the impact of **classification model** choice on temporal performance,
3. understanding **when and why model performance drops** over time, which informs when a model needs adapting, based on factors including content representation, classification model and dataset characteristics,
4. understanding the **potential and limitations of contextual language models** to develop temporal persistence, which in turn informs annotation practices,
5. looking at **dataset patterns from a longitudinal angle**, to assess the impact of different factors, and

6. assessing how different metrics extracted from small timeframes of the dataset can determine performance over time, such that we can improve **predictability of temporal performance** when longitudinally annotated datasets are not available and/or affordable.

We further discuss the implications of temporal performance drop on classification experiment design, highlighting the importance of considering the temporal persistence as an additional dimension in the evaluation.

4.1.2 Research questions

Our overarching research question is "how do different factors in the experiment design determine the temporal decay (or lack thereof) of classification results?". We break down this research question into five smaller ones that we address in a set of experiments:

- **4.RQ1.** How does the choice of a **language model** impact classification performance over time across different datasets?
- **4.RQ2.** How does the choice of an **algorithmic architecture** impact classification performance over time across different datasets?
- **4.RQ3.** How does prediction **performance vary across longitudinal datasets of different types**?
- **4.RQ4.** What are the **linguistic features that, in the absence of sufficient labelled data for direct testing, can help us estimate the temporal persistence on a particular dataset**?
- **4.RQ5.** How **stable are contextualised language model representations** with respect to temporally **evolving context changes**?

4.1.3 Contributions

In this chapter, we make the following novel contributions:

- We perform comprehensive experiments on three longitudinal text classifier datasets, with the aim of assessing the factors that impact model performance. We focus on key factors including language representation approaches, classification algorithms and dataset characteristics; concluding how, when and the extent to which they impact classification performance.
- We perform a comprehensive analytical study of dataset characteristics to investigate

how factors such as word frequencies in datasets can help predict model performance where one lacks sufficient labelled datasets over time.

- We propose a novel scalable methodology to quantify contextual meaning drift for dynamic (diachronic) aspects over time, which can help assessing how well a contextual model recognises texts from different periods.
- Our study provides insights into the factors that impact model performance drop over time, looking at five key dimensions: language representations, classification algorithms, time, lexical features and context.

We devise a set of best practices to consider when designing text classification models with the aim of keeping their performance as stable as possible over time. Among others, our study highlights that (i) the classifier that shows top performance in year 0 is likely to consistently perform best over time, whose performance drop is comparable to other models and does not impact the ranking among models, (ii) linguistic variability of the dataset at hand can help make an informed decision on the optimal classifier design, which depending on the amount of data available can be based on quantitative metrics or qualitative estimates (e.g. whether it is a social media dataset and whether it is a quickly evolving domain), and (iii) contextual language models provide a solid methodology to maximise temporal persistence thanks to their capacity to model sub-words, but also show that there is still room for improvement.

4.1.4 Chapter structure

While Chapter 2, Section 2.2 introduces background on the problem of temporal performance on NLP classification tasks. This chapter is organised as follows. In Section 4.2, we describe our research methodology, including our evaluation settings. We then describe and analyse the longitudinal datasets we use in Section 4.3.1, followed in Section 4.3 by the language models and classification algorithms we used, and lexical analysis methods used. In Section 4.4, we present our experimental objectives, evaluation and analysis of results. We provide a critical analysis of our findings in Section 4.5, concluding the chapter in Section 4.6.

4.2 Methodology

In this section, we delve into our experimental setup and evaluation metrics for temporal analysis of classifiers. The methodology for data collection and preparation, as extensively detailed in Chapter 3, with 3.1 focuses on the experimental setup and evaluation process.

Our experiments are dedicated to exploring the temporal persistence of classification models, considering scenarios where these models must perform in different temporal settings. This concept refers to a model's ability to maintain performance when applied to a testing set from a different, possibly more temporally distant, year than the one it was trained on. We include testing in past, future, and the same period, enabling a comprehensive assessment of temporal adaptability.

We use datasets collected and labeled uniformly over extended periods, as discussed in Chapter 3, Section 3.2, to minimize the influence of confounding factors and isolate the impact of temporal evolution on model performance.

We segment our datasets into 1-year intervals to align with their longitudinal nature, enabling us to observe significant changes over time. Although testing on larger time intervals might be preferable, we are constrained by the availability of datasets.

Our evaluation methodology revolves around the "Temporal Gap", a concept introduced in Chapter 3 Section 3.1.3, which allows us to systematically investigate how NLP models adapt to different temporal scenarios. Utilizing the Temporal Gap, Model Evaluation Metric, and Performance Change Metric, we assess the robustness and temporal dynamics of NLP models across various time gaps. This analysis provides insights into how these models generalize and evolve in response to changing data distributions over time.

In summary, Chapter 3 thoroughly covers data collection methodology, while this section focuses on our experimental setup and evaluation metrics, enabling us to study how NLP classification models persist over time, adapt to diverse temporal scenarios, and respond to evolving data distributions.

4.3 Datasets, Language models, classification algorithms and methods for lexical analysis

We perform experiments with a wide range of state-of-the-art static and contextual language representations, and neural classification models to evaluate by-design adaptability to temporal changes. The feature space is always generated using word-level vectors extracted from the selected set of pretrained language models, including static and contextual representations, and the model parameters are learned from the training data, and hence can also be impacted by the quality and temporal persistence of the Pretrained Language Model (PLM) in question.

4.3.1 Datasets used

We chose three datasets for our study based on their longitudinal nature, i.e. they cover the span of several years, where each year is dense containing sufficient data for training and testing, and they are labelled for text classification tasks: Two of the datasets come from social media (GESD and TESA), whereas the other one (ABRR) is made of user-generated contents in a more constrained setting (i.e. book reviews). Table 4.1 shows a summary of the datasets, including the time covered, size, annotation methodology as well as the task. Detailed analysis of the datasets are discussed in Chapter 3, Section 3.2.

Datasets	**GESD**	**TESA**	**ABRR**
Classification Labels	Binary: favour/against	Binary: positive/negative	Binary: 5/1
Labels balance	unbalanced	balanced	balanced
Annotation	Aggregated hashtags	Aggregated emojis	Human annotation/tags
Time range	2014-2019	2013-2020	2000-2018
Size per year	48,000	99,800	22,390
Total size	288,000	798,400	497,800
Sampling method	GetOldTweets API	Twitter stream archive	-
Source	Twitter	Twitter	Amazon
Data specificity	Gender Equality	Generic	Book Reviews

Table 4.1: Summary description of the selected datasets.

4.3.2 Pretrained language models

We experiment with two types of word representations, with three different methods of each type. The models are selected with the aim of having a diverse, competitive and widely used set of word representation methods.

Static Word Representations (SWRs):

- **Google Word2Vec (G-W2V)** [Mikolov et al., 2013b]: G-W2V widely used word embedding model, trained on roughly 100 billion words from Google News data.[1]
- **FastText (FT)** [Mikolov et al., 2018]: FT has 300 dimensional vectors trained from English Wikipedia by using the skip-gram method [Bojanowski et al., 2017].[2]
- **Twitter Glove (Glove)** [Pennington et al., 2014]: Glove is trained on the non-zero entries of a global word co-occurrence matrix. The model we use is trained from 2 billion tweets.[3]

[1] `https://code.google.com/archive/p/word2vec`
[2] `https://fasttext.cc/docs/en/pretrained-vectors.html`
[3] `https://nlp.stanford.edu/projects/glove/`

Contextual Word Representations (CWRs):

- **Bidirectional Encoder Representations from Transformers (BERT)**[a] [Devlin et al., 2019]. Google trained BERT in 2018 out of 11K unpublished books and English Wikipedia articles, using 110M parameters. A static masked language modelling (Static-MLM) loss objective was used to fine-tune BERT, with this masking is done only once as part of the preprocessing step prior training. BERT is trained as a non-auto-regressive model that generates contextual representations from all previous and next tokens in a text, then generates representation output all at once. WordPiece [Wu et al., 2016] tokeniser, a greedy approach for dividing words into smaller tokens during training, was used to produce these tokens from text. When used in downstream task, each out-of-vocabulary (OOV) word is split into known subwords. BERT's vocabulary contains 30,522 words.This reduces ambiguity in sentence meaning and chances of losing important signals due to OOVs.
- **Robustly Optimized BERT Pretraining (RoBERTa)**[b] [Liu et al., 2019]. RoBERTa inherits many of BERT's capabilities, with improved pretraining efficiency. It is trained from the same data as BERT, in addition to the CC-News dataset. RoBERTa uses Byte-Pair Encoding (BPE) [Shibata et al., 1999] for data compression, in a similar fashion to BERT's WordPiece tokeniser. During the pre-tokenisation step, the WordPiece algorithm selects symbol pairings that increase the probability of the training corpora, whereas the BPE algorithm selects the most common combinations. The number of words in the vocabulary is 50,265. The model is trained with a loss objective based on dynamic masked language modelling (Dynamic-MLM). This distinguishes it from BERT in that it changes the masked word for the same sentence at each training epoch. Similar to BERT, the model accepts tokenised inputs ranging from 3 to 512 and produces 12 neural layers with 768 hidden units.
- **Generative Pretrained Transformer 2 (GPT)**[c] [Radford et al., 2019]. GPT-2 was introduced in February 2019 by OpenAI, as a model trained on a corpus of 8 million web documents. Like RoBERTa, it employs BPE with space tokenisation. It differs from BERT and RoBERTa in that it is an auto-regressive model that generates outputs iteratively and predicts tokens unidirectionally based on their context through reading tokens from left to right. Unlike prior models, GPT-2 is trained with a causal language modelling (CLM) loss objective since it is intended primarily to create human-like writing in text generation tasks. Furthermore, it employs decoder attention blocks from the transformer design, as opposed to BERT, which uses encoder blocks. The vocabulary size is 50,257. GPT-2 accepts tokenised inputs ranging from 3 to 1024 and produces 12 neural layers with 768 hidden units.

[a] `https://huggingface.co/bert-base-uncased`
[b] `https://huggingface.co/roberta-base`
[c] `https://huggingface.co/gpt2`

4.3.3 Machine learning models

We use a set of **classification models,** including two types of classifiers, and three algorithms of each type of classifier. All architectures were chosen based on their robustness in many state-of-art text classification tasks.

Traditional classifiers:

- **Linear Support Vector Machine (LinearSVC)** [Joachims, 1998]. LinearSVC is an algorithm that searches for a hyperplane (decision boundary) and linearly separates classes, and is used in numerous NLP classification problems. The introduction of kernel functions allows LinearSVC to generalise to high dimensional and sparse feature spaces. Therefore, feature engineering is not required even when the number of dimensions exceeds the number of data samples. Moreover, it is able to handle large numbers of features.
- **Logistic Regression (LogisticRegression)**. A supervised classification approach, it is used in binary classification because it fits a single line to split the space exactly into two using a sigmoid function. Using a probability curve to predict classes makes LogisticRegression different from LinearSVC, and more similar to deep learning models. The LogisticRegression classifier can interpret model coefficients [Bruin, 2011], an effective indicator of feature importance.
- **Multinomial Naive Bayes (MultinomialNB)**. MultinomialNB is a probabilistic learning algorithm that assumes features to be conditionally independent. Its probabilistic nature makes it vulnerable to data sample distributions but robust to deal with infrequent features, unlike LogisticRegression and LinearSVC.

Deep learning models:

- **Convolutional Neural Network (CNN)** [Kim, 2014]. Despite its capacity to retain the order of feature elements, CNN is a shift-invariant classifier. CNN may process the entire text or a portion of it by taking reduced signals from the sequence and sliding over it with a single conventional head. A deep CNN may include additional layers, allowing it to decrease signal overruns for each phrase. CNN can also learn several levels of structured n-gram co-occurrences. This is mostly owing to its mapping during the reduction phase via a filtering mechanism with a certain size and sliding window. This enables the CNN to compress the feature space of words into smaller latent feature representations.
- **Long-short Term Memory network (LSTM)** [Hochreiter and Schmidhuber,

1997]. With LSTM the information persistence is high [Salton and Kelleher, 2019] due to the use of recurrent units connected as chains and temporal backpropagation. This allows a classifier to learn based on long term dependencies between sequences regardless of textual length or dataset size. Moreover, the ability of LSTM to model temporal historical dependencies makes it useful to capture subtle semantic changes [Elman, 1990]. The is done though temporal memory sequencing for modeling each sentence word by word while training the model. This allows LSTM to reveal the latent syntactic and semantic long term signals for any word even from noisy data. Thus, it generalises better than other non-memory dependent models. This is different than classical models which are usually trained to reduce the mean square error by considering current inputs only and without using memory units to keep historical states of inputs.

- **Hierarchical Attention Network (HAN)** [Augenstein et al., 2016, Li et al., 2019]. The HAN architecture we use includes GRU units and an attention mechanism. The first layer is a bidirectional GRU [Cho et al., 2014] layer, which is a kind of RNN and an LSTM variant. GRU, on the other hand, is less complex and hence quicker than LSTM. The next layer is an attention mechanism between word vectors and sentence representations. Text classifiers can generalise prediction by identifying the latent properties that exist between individual word vectors and the average representations of all words in a sentence.

4.3.4 Preprocessing and classifier hyperparameters

We preprocess the inputs to maximise the coverage of words in the training set, following two steps to clean the dataset and to tokenise the words to match the SWR. Out-of-Vocabulary words in the training set added with zero vectors to SWR. The sentence inputs are padded and truncated to 128.

We use a softmax layer to obtain class probabilities as the final layer in all our classifiers' architectures. We also use early stopping strategy with restoring best weights by measuring the validation loss on same-period training data. The number of training epochs is set to 25 with the same hyperparameters.[4]

Each pretrained language model's text classifier is a feed-forward neural network architecture in which word-level representations are used as initial weighted layers for input

[4]As noticed that models performed similarly to one another, we assume that our findings apply to other hyperparameters; more study is required with consideration to computational costs.

tokenised sentences. Then fed into a flattened layer followed by a final sigmoid layer that outputs class probabilities. In the case of 4.RQ2, addressed in Section 4.4.2, all traditional classifiers used the default parameters and TF-IDF features. All deep leaning models use the Adam optimiser with the learning rate fixed at $2e^{-5}$.

4.3.5 Methods for lexical analysis

Along with our experiments, we perform lexical analyses of the datasets to support our findings, particularly for 4.RQ4 and 4.RQ5. To do this, we propose several measures looking into aspects that can explain the causes of performance patterns we observe in our experiments.

We categorise these measures in two groups: (1) lexical variations across datasets where we propose word-level statistical measures across training and test sets, and (2) contextual variations over time where we propose a model to track context-level meaning drift over time using cosine similarities and variance.

Lexical variations across datasets

To address **4.RQ4** in Section 4.4.4, we calculate various unsupervised metrics estimating the similarity between training and testing sets. We then analyse whether these metrics are able to predict how well a model will perform on the new test set.

We calculate the Pearson correlation coefficient between the following four statistical measures and model performance to quantify their potential impact on model performance. Table 4.2 shows a summary of all quantitative measures, which we discuss next. We discuss the results of this analysis in section 4.4.4.

Measure	***Definition***
Familiarity score	$familiarity(U,V) = \frac{\lvert U \cap V \rvert}{\lvert V - U \rvert}$
Jaccard index	$jaccard(U,V) = \frac{\lvert U \cap V \rvert}{\lvert U \cup V \rvert}$
TFIDF similarity	$similarity(U,V) = tfidf(t,d,U) * tfidf(t,d,V)$
Information rate	$H(V \mid U) = -\sum_{v,u} p_{V,U}(v,u) \log_b \frac{p_{V,U}(v,u)}{p_U(u)} = H(V,U) - H(U)$

Table 4.2: Summary of different evaluation measures used to quantify Lexical variations cross-datasets using unlabelled data. Note: U: vocabulary of training and V: vocabulary of test set.

- **Familiarity score:** This is measured by relying on two metrics extracted from two sets of words: (i) overlap ratio, as the proportion of words, which occur in both sets of words, and (ii) uniqueness ratio, as the ratio of words that occur only in one of

the sets of words. The familiarity score is calculated by dividing the overlap ratio by the uniqueness ratio.

- **Jaccard Index (JI):** The JI is based on the intersection and union of words in two vocabularies. The JI is calculated as the size of the intersection divided by the size of the union.

  The resulting coefficient is asymmetric.

- **TF-IDF:** By calculating the TF-IDF weights of words in two different vocabularies, we can then calculate the similarity between the two vocabularies. The similarity is calculated using the cross product similarity.

  This is also an asymmetric measure and avoids computation of sentence-by-sentence similarities.

- **Information rate:** By training a Markov model on a source vocabulary, and testing it on a target vocabulary, we rely on conditional entropy to estimate the information rate of the target vocabulary.

  The benefit of using information rate is to test the model's ability to predict the next word given the conditional probabilities from the training set. The information rate quantifies the amount of information required to describe the testing set V given training set statistics U. With p_U the probability distribution of U, and $p_{V,U}$ the distribution for the joint distribution (V,U), the base- b conditional entropy.

Temporal language variations

To answer **4.RQ5** we aim to analyse language variations over time by looking at the context that is surrounding the aspects extracted from the corpora. In this case, we choose to work with aspects that capture opinionated multi-word expressions, owing to two reasons: (i) given the nature of our datasets (sentiment, stance, reviews), to analyse how context surrounding these core elements changes, and (ii) because aspects are less likely to be polysemous than isolated tokens, hence reducing the potential of noise from multiple co-existing meanings in the analysis.

Year-specific discrete representations of aspects can help us analyse how their context changes over time.

To achieve this, we develop a novel approach to generate time-specific representations of aspects, allowing to quantify their change over time based on how the surrounding context changes.

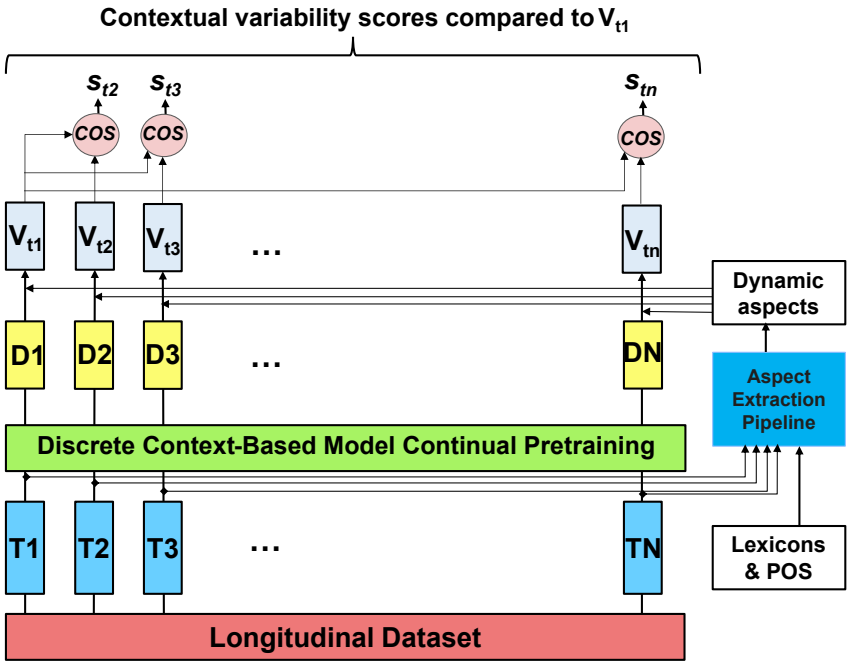


Figure 4.1: Our approach for assessing contextual coverage for dynamic aspects over time. The dataset is first split into years. For each year, we then pretrain a contextualised model (e.g. BERT, RoBERTa) using a masked language modeling strategy. Having this year-specific language models, we can then calculate similarity scores across years. When these similarity scores are low for a particular aspect, they indicate a prominent change in its surrounding context over time.

Our approach to measure the temporal change of context is illustrated in Figure 4.1. It consists of two main steps: (i) Aspect Extraction Pipeline, where we identify aspects present in texts, and (ii) Contextual Change Measurement Model, where we measure how much the context of these aspects has changed.

Step 1: Aspect Extraction Pipeline: We first extract aspects, or multi-word expressions, from texts. To extract aspects, we rely on two resources:

- A lexicon of 7K words we build from two source dictionaries: (1) opinion lexicon [Biber, 2006] including adverbs and adjectives of different opinion modalities which covers different semantic categories: possibility, necessity, certainty, likelihood, attitude, communication, evaluation (210 lexicon words); (2) polarised sentiment lexicons [Liu et al., 2005] (2K positive and 5K negative lexicon words), which also covers misspelled words.

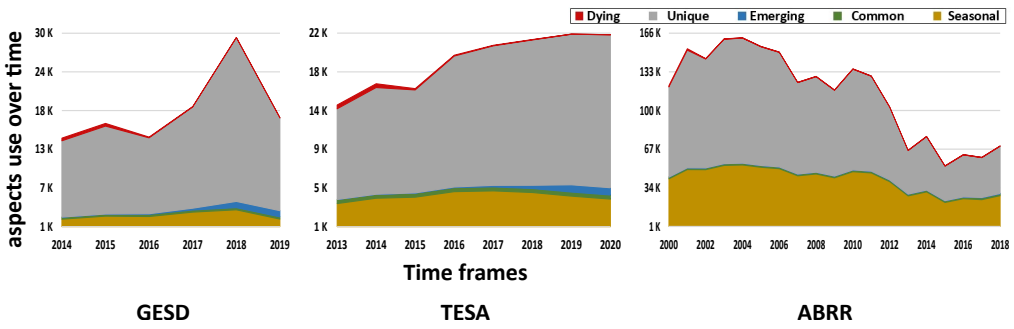


Figure 4.2: Temporal usage of different aspect types. Dying aspects (red), unique aspects (gray), emerging aspects (blue), common aspects (green) and seasonal aspects (brown).

- The Spacy part-of-speech (POS) tagger [Hu and Liu, 2004, Honnibal and Montani, 2017], which we use to tag the texts.

Next, we look at sequences of adjacent words that both (i) contain a word present in the lexicon, and (ii) based on the output of the POS tagger, match one of the 45 POS regular expressions defined by Hu and Liu [2004]. The set of sequences matching both criteria constitute our final set of aspects.

By analysing the frequencies of the resulting aspects over time, we can categorise them based on their use frequency over time into one of dying, unique, emerging, seasonal or common, following the same approach as in Section 3.3. Figure 4.2 shows the result of the analysis, demonstrating that dynamic aspects are frequent and their use varies over time.

Step 2: Contextual Change Measurement Model: The change measure model consists in turn of the following three steps: (1) generate time-specific embeddings of aspects, (2) measure pairwise similarities between time-specific embeddings of an aspect over time and (3) rank aspects by their contextual change.

1. **Step 2.1: Generating time-specific embeddings:** Given a longitudinal dataset D that spans several years $\{T_1, T_2, \ldots, T_N\}$ where each year has the same number of sentences per year (SpY), and a context-based PLM. A discrete set of time-specific embedding can be trained using masked language strategy (MLM) for each year.

   We obtain 6 discrete time-specific embeddings for GESD dataset, 8 for TESA and 19 for ABRR.

2. **Step 2.2: Measuring similarities between two time-specific representations of an aspect:** The contextual change of an aspect from time T_1 to each time T_i can be calculated by measuring the cosine similarity between the relevant time-specific

representations [Hamilton et al., 2016c, Shoemark et al., 2019, Tsakalidis et al., 2019]. The inverse of the cosine simlarity then indicates contextual change, i.e. lower cosine similarity indicating bigger change. In the interest of simplicity and focus, similarity scores are computed for each year with respect to year T_0, which is kept constant as a pivot.

$$s_{ti}(a, D1, Di) = \frac{\sum_{j=1} v_j^{T1} v_j^{Ti}}{\sqrt{\sum_{j=1} v_j^{T1^2}} \sqrt{\sum_{j=1} v_j^{T1^2}}}, \tag{4.1}$$

where V_{T1} and V_{Ti} are the embeddings of aspect a for years T_0 and T_i, trained using datasets $D1$ and Di, respectively. We repeat this process several times for all $i \in \{t+1, ..., T\}$ for all N temporal data frames in a given longitudinal dataset D. For each aspect, we then get a final list of cosine similarities s between T_1 and all T_i.

3. **Step 2.3: Ranking aspects by contextual change:** To measure the contextual variability of target aspects over time, we use the variance (σ^2) to measure the fluctuation of cosine similarity scores from the mean similarity score (μ) over time of the same aspect, defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (\cos(V_{T1}, V_{Ti}) - \mu)^2}{N} \tag{4.2}$$

4.4 Experiments, results and analysis

In this section, we present the results of the five experiments we conduct to address the five research questions we set forth. We first assess the extent to which different language models (4.RQ1, Section 4.4.1) and algorithmic architectures (4.RQ2, Section 4.4.2) persist their classification performance. Then, by looking at temporal gaps (4.RQ3, Section 4.4.3), linguistic features (4.RQ4, Section 4.4.4), and the similarity of a given set of dynamic aspects contextual variations (4.RQ5, Section 4.4.5), we concentrate our studies on the linguistic complexity through temporal prospective to determine when and how diachronic changes impact text classifiers' generalisability.

4.4.1 Experiment 1: Impact of language representations on performance (4.RQ1)

In this experiment, we aim to answer 4.RQ1 to assess the impact of language representations on the classification performance. We use the six different language models presented in Section 4.3.2. We also present grouped performance averages for each group consisting of static (SWRs) or contextual (CWRs) word representations.

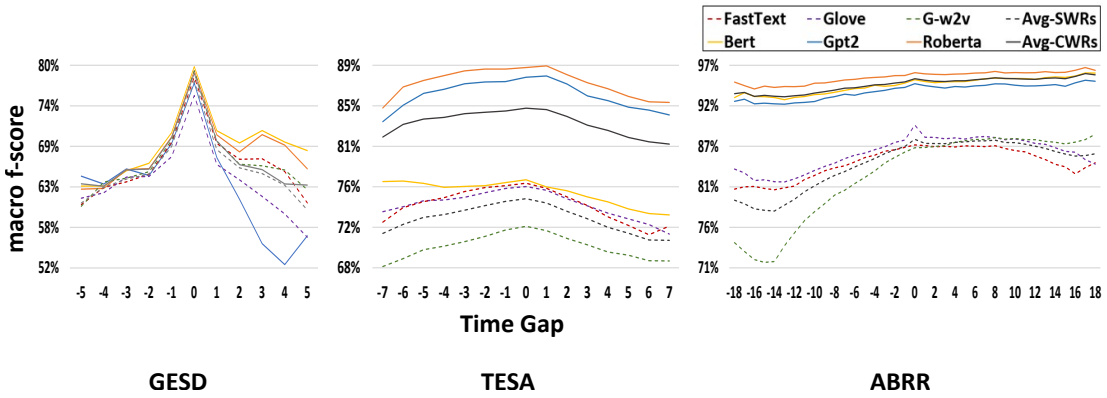


Figure 4.3: Temporal performance of different language representations across the three longitudinal datasets. *Dashed-line*: average f-score results for average static-based representations, *Bold-line*: average f-score results for all representations for average context-based representations.

Classification results for different language representations across the three datasets are shown in Figure 4.3. Most importantly, we observe performance decay for all datasets and language models at least in one direction (future or past) comparing to the initial temporal gap of 0 (present). Moreover, the performance decay for each PLM and dataset remains consistent, where all PLMs show similar decaying trends. Despite these commonalities, we make some interesting observations:

- The performance decay is not the same for all datasets. We observe that the decay is particularly steep for the GESD dataset, where even the top-performing representation (BERT) drops performance by more than 15% for data 5 years in the past and by more than 10% for data 5 years in the future. This drop is comparatively modest for the other two datasets, TESA and ABRR. In the case of TESA, we observe that all representations drop slightly in performance for past and future data, with the exception of BERT that achieves some degree of stability for the past data. The trend line is different for the ABRR dataset, which shows performance stability on future data, likely due to the shrinking nature of its vocabulary over time, as shown

in Section 3.3.

- SWRs exhibit a substantial drop in performance for the past data, and lesser drop for the future data; the more pronounced drop of SWRs compared to CWRs is likely due to the ability of the latter to model context, which enables some readjustment of words that change meaning over time thanks to information extracted from the surrounding context. To better understand and explain these different patterns we observe across datasets, we will delve into dataset characteristics in 4.RQ3 (section 4.4.3).
- Looking at the absolute performance scores of different representations, we observe that CWRs consistently outperform SWRs, which is in line with the expectations. However, we do observe some variation across the different CWRs, as for example BERT performs best for GESD, but its performance leaves much to be desired for TESA. We observe an overall best generalisation across datasets for the RoBERTa model, which performs best for TESA and ABRR, and close to the best for GESD. The improved generalisability of RoBERTa over BERT may be due to its larger pretrained vocabulary, as RoBERTa is trained from the same corpora as BERT, with the addition of CC-News.

4.4.2 Experiment 2: Algorithmic architecture impact on performance (4.RQ2)

In this experiment, we address 4.RQ2, assessing the extent to which different algorithmic architectures enable temporal persistence (please refer to Section 4.3.3 for the full description of the algorithms used). In the interest of focusing on algorithm impact, we average performances for each type of language representation, i.e. average of SWRs and average of CWRs.

Figure 4.4 shows the temporal performance scores for the different classification algorithms under study. We observe the following:

- The overall tendency is for all models to exhibit similar performance trends over time. Despite some models achieving better absolute performance than others, the performance drop across models is largely consistent. The performance drop we observe for future and past periods is mostly symmetric for GESD and TESA, where the performance drop increases for larger temporal gaps. ABRR shows a similar trend for past periods, however with a different pattern for future periods, where performance remains stable or even increases slightly on occasions.

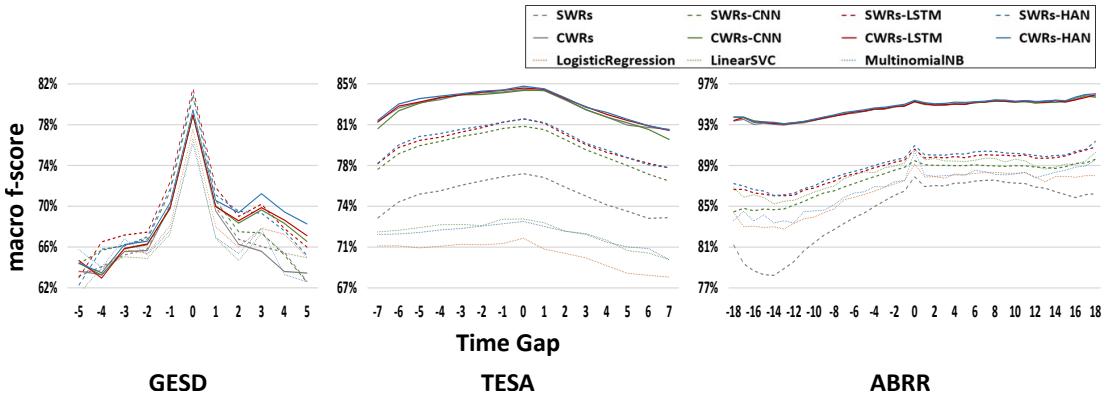


Figure 4.4: Temporal performance of different algorithmic architectures across the three datasets. *Dotted-line*: f-score results for each traditional machine learning model with frequency-based representations, *Dashed-line*: average f-score results for average static-based representations, *Bold-line*: average f-score results for all representations for average context-based representations.

- In terms of absolute performance scores, we observe that HAN-based models achieve an overall better performance than other models, which however also suffers from the same performance drop pattern as the other models. This is consistent with the findings in 4.RQ1, and hence shows limited impact of the classification algorithm when we look at temporal persistence as the key factor.

Similar to what we observed in the first experiment dealing with language representations, and leaving absolute performance scores aside, again we see predominantly consistent performance trends regardless of the algorithm choice. The shape of these performance trends is in fact very similar for different language representations and algorithms selected, which highlights the impact that the data itself has on this performance drop. While we have seen that some pretrained language models can achieve improved generalisability in temporal persistence (as is the case of RoBERTa thanks to its broader vocabulary than BERT), we can't conclude the same for classification algorithms, as we don't see a similarly clear difference in this case. We can therefore that, when it comes to temporal persistence, language representations obtained through pretrained language models have a bigger impact than classification algorithms, which are not as crucial.

Aside from pretrained language models and classification algorithms, looking at the different datasets we observe that the performance drop is more prominent and consistent for the TESA and GESD social media datasets. The trends are different for ABRR, which, as a dataset collected from reviews associated with a particular domain, i.e. books, one

could hypothesise that leads to a lesser impact over time due to the predictably more constrained vocabulary. The clear differences in performance across datasets leads to our third research question, where we conduct more quantitative experiments to investigate our selected datasets structure to gain deeper understanding of models temporal performance fluctuation.

4.4.3 Experiment 3: How the temporal gap impacts performance (4.RQ3)

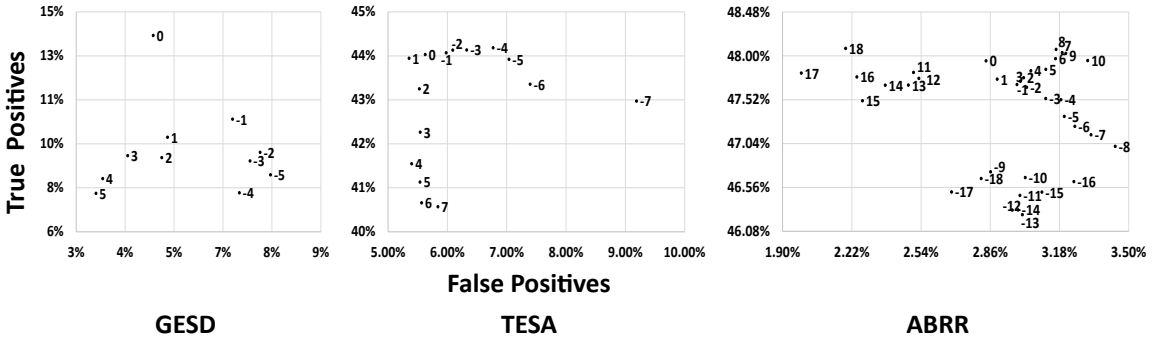


Figure 4.5: Model performance across temporal gaps. The best performing sets are those positioned in top left corner of the grids, i.e. those minimising false positive rates and maximising true positive rates.

In this section we address 4.RQ3 by looking at how different factors in the inherent characteristics of the datasets impact model performance over time.

Figure 4.5 shows TP (true positive) and FP (false positive) rates as a scatter plot for different temporal gaps, where data points are labelled with the corresponding temporal gap.[5]. This allows us to rank temporal gaps based on TP and FP rates. Note that the optimal results are those minimising FP and maximising TP, hence values positioned in the top-left corner of the figures are ideal. In the interest of clarity, we focus here on results produced by CWRs-HAN model as the best-performing model in the previous experiments.

For both GESD and TESA, the best performance is achieved for the temporal gap 0, i.e. same-year train and test. In the case of GESD, we observe that performance is better for future tests, whereas performance is similar for future and past test sets for TESA. The trend is quite different for ABRR, where performance for future test sets is even better than for temporal gap 0, which confirms our finding in 4.RQ2 showing that the constrained vocabulary in this dataset leads to improved performance. This makes future test sets

[5]Similar to ROC Curve [Flach, 2012], as we have data-centric approach for using same model we refrain from using same name

Vocabulary Test/Dataset	*GESD*	*TESA*	*ABRR*
Familiarity score	+54%	+48.5%	+57.7%
Jaccard index	+73.3%	+29%	+65.5%
TF-IDF similarity	+49.1%	+46.8%	+14.1%
Information rate	+64.2%	+34%	+52.5%

Table 4.3: Lexical variation correlation scores of the evaluation matrics (Familiarity Score, Jaccard Index, TF-IDF Similarity, and Information Rate) (and two-tailed p-values < 0.05 significant results) of metrics against model's performance at best overall prediction model (CWRs-HAN) for each dataset (GESD, TESA, and ABRR).

less linguistically diverse, with better coverage of common words over time, facilitating classification of future data in ABRR. This can be observed in the cluster of larger temporal gaps (11+ years) in the top-left corner of the ABRR figure.

4.4.4 Experiment 4: How lexical variations across datasets help determine temporal persistence (4.RQ4)

With 4.RQ4 we aim to investigate if there is a metric that we can use to estimate the temporal persistence (or lack thereof) that a dataset will exhibit, given the realistic scenario where we lack longitudinally labelled data. Can we estimate this temporal persistence for the linguistic characteristics of an unlabelled dataset and, if so, with what metrics?

To address 4.RQ4, we analyse the Pearson correlation scores between four different factors in dataset vocabularies and model performances with CWRs-HAN model predictions (see Section 4.3.5 for implementation details).

Table 4.3 shows the resulting correlation scores for the four factors under study (See Section 4.3.5 for more details about the measures). The scores represent the Pearson correlation coefficients between various lexical variation metrics (Familiarity Score, Jaccard Index, TF-IDF Similarity, and Information Rate) and the performance of the best overall prediction model (CWRs-HAN) for each dataset (GESD, TESA, and ABRR). Significant results are indicated by two-tailed p-values < 0.05. A first look at the correlation scores shows consistent positive correlation scores across the board. However, the values of these positive correlation scores vary substantially, showing different degrees of strength in these correlations. The familiarity score is similarly high for all three datasets. In the case of GESD and ABRR, both the Jaccard index and the information rate are particularly high, whereas the TF-IDF similarity score is higher for both GESD and TESA than for ABRR. All in all, however, we could determine that, among the metrics under study, the familiarity score,

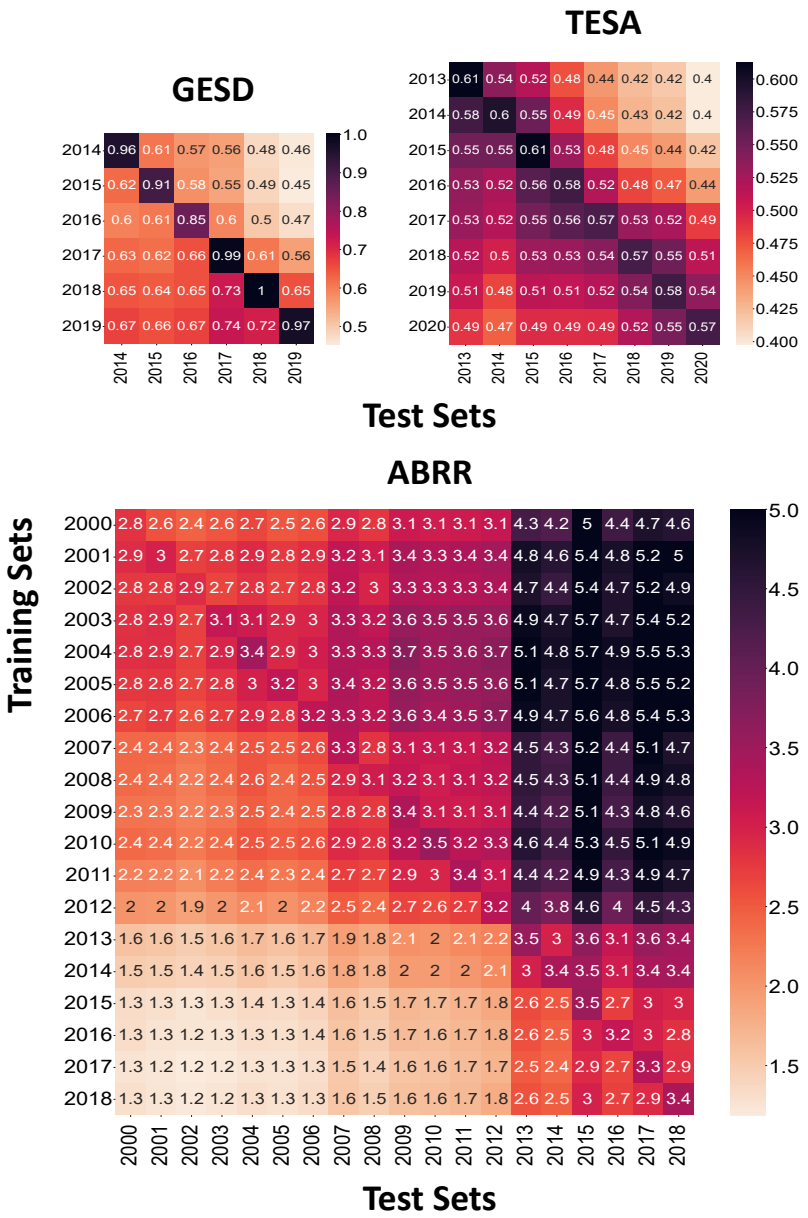


Figure 4.6: Temporal effect of familiarity score in each train-test pair. Darker cells indicate higher rates.

given its consistency across the dataset, is the best metric to predict model performance deterioration. Indeed, where one lacks labelled data for years beyond the current year 0, one could estimate model performance drop by using the familiarity score as a metric.

To further understand the insights derived from the familiarity score, we delve into its results with finer granularity. Figure 4.6 illustrates pairwise familiarity score values for each pair of years. The first clear pattern difference we observe is between ABRR and the other two datasets. With ABRR, we see that the highest familiarity scores are between initial years for training (e.g. 2000) and final years for testing (e.g. 2018), a similarity that decreases even for same-year comparisons for early years (e.g. 2000) and for past year comparisons (e.g. 2018 training and 2000 testing). This observations through the familiarity score is in line with the performance scores observed in earlier sections for ABRR. On the contrary, both TESA and GESD show show higher familiarity scores for years that are closer to each other, which decrease as years are further apart from each other. This is again in line with the performance scores we have observed.

This again reaffirms the validity of the familiarity score as a metric to estimate the model performance deterioration for a particular dataset where we lack labelled data for more than a single year, i.e. we can calculate the familiarity score based on the overlap and uniqueness of the vocabulary sets in different years of unlabelled data, and we can expect some degree of correlation with the actual performance.

4.4.5 Experiment 5: Generalisability of contextual characteristics of context-based models (4.RQ5)

To address 4.RQ5, following the methodology described in Section 4.3.5, we quantify how the changing context of unique aspects impact the resulting representations of contextualised language models. The aspects we use to measure these changes are short, multi-word expressions, such as:

``@realDonaldTrump great'' from TESA

This is an example of an aspect occurring in only one of the years in the TESA dataset.

For this specific example, we observe that its contextual similarity has dropped gradually as we move further away from the first year, with the following similarity scores: 2014 (0.8542), 2015 (0.7838), 2016 (0.6222), 2017 (0.6222), 2018 (0.6119), 2019 (0.5368), 2020 (0.5171). Using multiple examples like this, we can determine whether transformer-based models produce stable representations over time despite the contextual changes. To do

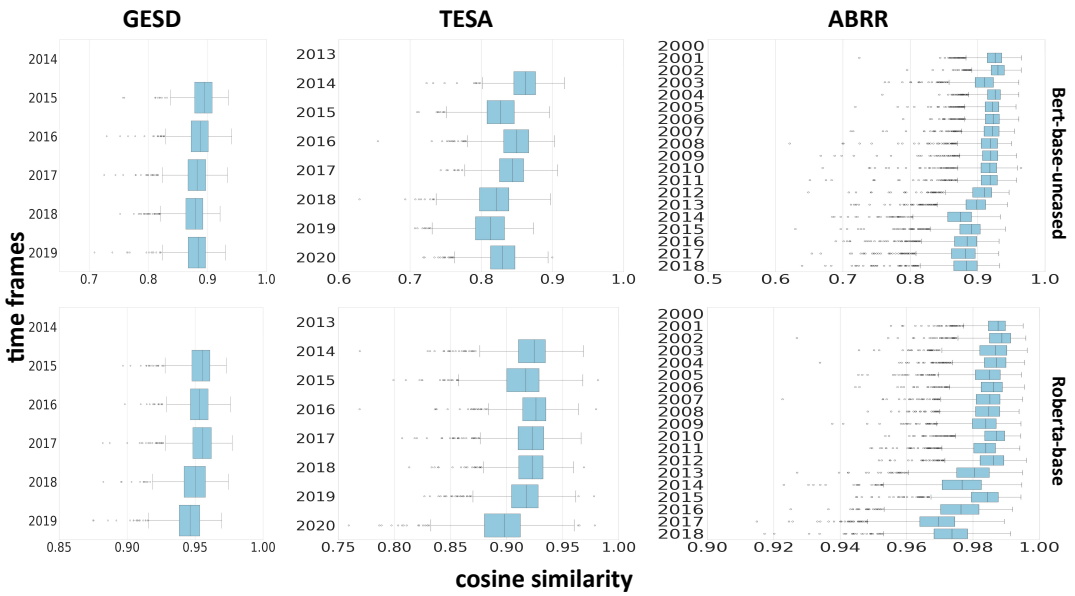


Figure 4.7: Assessing context-based temporal semantic similarity decay using discrete temporal context-based representations for BERT and Roberta by measuring the cosine similarity of dynamic seasonal aspects.

so, we next assess whether the trend observed in the example above generalises to other aspects in the dataset.

Figure 4.7 shows the similarities of all aspects across years and across datasets. Each box plot represents the similarity scores for the set of aspects under consideration. The box plot for each year represents the similarities in that year with respect to the first year in the dataset, which is taken as a reference. For figures looking at common aspects, unique aspects and seasonal aspects, we observe a consistent tendency for the similarity scores to decrease over time, as the year in question is further apart from the first year.

This indicates that meanings and contextual representations keep varying over time, where this change does not happen all of a sudden, but it gradually happens showing a decreasing tendency of similarity scores. This in turn indicates that context-based language models do not fully capture the evolution of these aspects. Even if context-based language models provide the means to process sub-word, contextualised representations of texts, this has still room for improvement in furthering temporally stable representations to enable temporal persistence of models.

Indeed, learning a meaningful contextual representation for a unique aspect A at time T is

much harder for models distant in time compared to models closer in time although A has never been seen in any other time-frame. While both BERT and RoBERTa showed slight temporal changes in the context when using MLM in which temporal splits are used to generate the discrete temporal embeddings. This change is smaller in RoBERTa compared to BERT which is potentially a result of the dynamic masking strategy for sentences utilised in RoBERTa. This again indicates, as discussed earlier in 4.RQ1, that RoBERTa proves to be more stable than BERT when it comes to temporal persistence.

DATASETS					
GESD		**TESA**		**ABRR**	
BERT-based aspects temporal cosine similarities variance					
common aspects	σ^2	**common aspects**	σ^2	**common aspects**	σ^2
dowry death	0.0065	most welcome	0.0092	the bad piece	0.0019
interesting report	0.0064	I miss you	0.0068	super book	0.0018
amazing group of woman	0.0014	the good	0.0026	and entertaining book	0.0005
male hatred	0.0014	head hurt so bad	0.0026	very disappointing book	0.0004
seasonal aspects	σ^2	**seasonal aspects**	σ^2	**seasonal aspects**	σ^2
husband ask why	0.0110	love you	0.0143	in timely manner	0.0122
misandry strike	0.0072	thank mummy	0.0076	good seller list	0.0093
I see they	0.0011	I need food	0.0018	protect a family	0.0003
the legal protection	0.0010	I love grill	0.0018	they claim he be	0.0003
unique aspects	σ^2	**unique aspects**	σ^2	**unique aspects**	σ^2
gap well	0.0074	@realDonaldTrump great	0.0341	a prompt fashion	0.013
be willing to make mistake	0.0058	just stunned	0.0097	in true geoff johns	0.0106
fantastic group of college	0.0011	just love that baseline	0.0015	he use apparently british slang	0.0003
the quiet recess	0.0010	I wish they tape	0.0015	fun and offer practical	0.0003
Roberta-based aspects temporal cosine similarity variance					
common aspects	σ^2	**common aspects**	σ^2	**common aspects**	σ^2
political empowerment	0.0032	lovely lady	0.0031	a complete rip	0.0001
so inspiring	0.0019	so handsome	0.0026	m sorry to say that	0.0001
amazing group of woman	0.0002	miss they so	0.0006	big waste of time	0.0000
to strong woman	0.0001	so so sad	0.0005	mess of a book	0.0000
seasonal aspects	σ^2	**seasonal aspects**	σ^2	**seasonal aspects**	σ^2
never be more relevant	0.0018	good beauty	0.0042	significant character	0.0007
racist people	0.0013	be as bright	0.0031	strong magnifying	0.0004
sick of	0.0002	really wanna get	0.0003	I really appreciate the way	0.0000
work on the basis	0.0002	plan for this weekend	0.0002	compliment the story	0.0000
unique aspects	σ^2	**unique aspects**	σ^2	**unique aspects**	σ^2
an epic fail	0.0029	good nazi	0.0036	due to damage	0.0006
breach of the FA	0.0023	the correct response	0.0035	use editorial help	0.0005
sensitised student on issue	0.0001	I honestly wish	0.0003	all about the bullying	0.0000
to false rape victim	0.0001	quiet about kmm	0.0002	quite right with this book	0.0000

Table 4.4: Aspects represented using BERT and RoBERTA using high and low contextual variability score examples for each dynamic language use type.

Complementing the analysis showing specific examples of aspects and their scores in Table

4.4 shows specific examples of aspects for each dataset, grouped by type. The table shows the two aspects with the highest and the two with the lowest similarity variations.

4.5 Discussion

Our work provides a first-of-its-kind study looking into both quantifying the challenges of developing temporally persistent text classifiers using dynamically evolving datasets, as well as further understanding the impact of dataset factors into this performance drop. Beyond simply confirming or quantifying that model performances drop over time, we further dig into getting a broader understanding into the problem by getting insights applicable in future experiments. We envisage the situation where one may need to choose a persistent classification strategy, but where longitudinally annotated datasets are lacking at the time of making this decision. Hence our overarching research question revolves around: what are the properties of the dataset, model or task at hand that can determine its temporal persistence?

A first look at the results from our experiments confirms that state-of-the-art language models and algorithms experience a performance drop over time as data evolves. This performance drop shows a similar trend compared to deep learning models, despite achieving overall better performance scores in terms of absolute values. This is especially true with user-generated content and social media datasets, where less formal and less established vocabularies are prone to evolve quickly. Both our experiments and our analyses demonstrate this in terms of performance scores and lexical changes, respectively. Where previous research has also posited the importance of capturing social changes, e.g. Syria and Iraq associated with war context [Kutuzov et al., 2018], where new words emerge and word meanings evolve, there is a growing need to develop temporally persistent models.

Along with these general findings, a deeper look into the results, by investigating the five research questions we defined, leads to interesting and important findings that helps inform the design of text classifiers with their temporal persistence as the objective.

4.5.1 Summary of key findings and best practices for classifier design

We next summarise our key findings from this research, along with our suggestions for best practices for classifier design, which we highlight in bold.

Our initial vocabulary analysis of three longitudinal datasets shows varying patterns of different types of words. We observe that social media datasets (GESD and TESA) experienced an increase of vocabulary size over time, whereas the book review dataset (ABRR)

experiences the opposite effect by exhibiting a decrease of vocabulary size. A possible explanation of this is that a more constrained domain such as book reviews may tend to settle into a fixed vocabulary, as opposed to more open domains. Having a more fixed, less varying vocabulary, also means having fewer emerging words, as well as fewer ephemeral words, both of which have shown to have a strong impact on model performance drop over time. Indeed, the reduced vocabulary size and variation leads to improved persistence of model performance over time in the case of ABRR.

Language models (4.RQ1) and algorithms (4.RQ2) both exhibit similar trends in their performance drops. This performance drop is particularly prominent for GESD, a dataset that deals with stance classification on the timely topic of gender equality, which is expected to have fluctuated over recent years. This means that the GESD dataset may be also impacted by how people expressed their opinions on this evolving topic, which in turn translated into a more prominent vocabulary change and a more noticeable performance drop in a shorter period of time (5 years) than in the other two datasets (7 and 18 years). Among the other two datasets, TESA experiences a larger performance drop than ABRR, again owing to the vocabulary stability and lack of vocabulary growth observed in the latter, as explained above.

Where all language models and algorithms show a similar trend in their performance decay, there is a relatively high consistency with the best-performing combination of model and algorithms in temporal gap 0 also performing best over time. Based on this observation, where one only has accessed to labelled data from year 0, **a wise approach to design a robust classifier can be based on the combination of language model and algorithm leading to top performance in year 0**.

Despite a consistent trend in performance drops across language models and classification algorithms, however, we observe that language models do exhibit different abilities to generalise better in terms of temporal persistence across datasets. This is particularly true when we compare RoBERTa and BERT, where we see that the former achieves better generalisability across datasets. Hence, despite having room for improvement in terms of temporal persistence, **when designing a persistent text classifier, it is safer to choose RoBERTa over BERT provided its improved generalisability across datasets**. One possible explanation is that the extended vocabulary of RoBERTa enables this improved generalisability, and therefore it opens an avenue for future research in further increasing its vocabulary. While we see this interesting difference between the two pretrained language models RoBERTa and BERT, we don't see a similar, informative difference be-

tween different classification algorithms, all showing similar trends in terms of persistence and generalisability.

When we look at time (4.RQ3), we observe that –predictably– the best model performance is generally achieved when the training and test datasets pertain to the same year (gap 0). However, surprisingly this is only true for two of the datasets (GESD and TESA), and we observe that performance scores can be higher when a model trained on older data is tested on future ABRR data, even when they are 18 years apart. Despite this unexpected outcome, it again highlight the nature of the ABRR dataset which has a very different vocabulary pattern than the other two datasets. A closer look at the characteristics of ABRR indeed shows a temporally shrinking vocabulary growth, indicating that the vocabulary in a more constrained domain like Amazon book reviews (compared to more informal, evolving social media data) makes it easier to achieve close to persistent performance. Hence, **an analysis of the vocabulary growth exhibited by a dataset is an important factor to consider when designing a persistent text classifier, with datasets showing low growth requiring less effort to achieve persistence**.

Further looking at correlations between linguistic features and model performance drop (4.RQ4), we find that the four linguistic factors we studied (familiarity score, jaccard index, similarity and information rate) play an important role in determining model performance over time. Indeed, one can effectively use these factors to estimate the temporal persistence of models. Where **one has labelled data from year 0 and unlabelled data from other years, a classifier design informed by these metrics can lead to a better decision**. More specifically, we have found that, among the metrics under study, the familiarity score provides the best estimate of model deterioration without using labelled data beyond year 0. However, where one only has labelled data from year 0, but **no further unlabelled data yet, one can instead attempt to estimate these metrics by relying on other more qualitative factors**, such as whether the data comes from social media and whether the domain contained in the dataset is expected to vary substantially.

Looking at the temporal persistence of contextual language models (4.RQ5), we found that while the emergence of new words can be solved by sub-word tokenisation in contextual-based embedding models, words dynamic between training and test pairs still play an intrinsic role in determining text classifier performance. Indeed, a large number of unique words and low overlap of vocabularies leads to higher uncertainty in text classifier predictions, where sub-word tokenisation proves insufficient. Despite their ability to model out-of-vocabulary words, we observe that models still get outdated. We quantify this by

measuring the change of meanings of word and aspect representations over time in discrete language models trained over time. The good performance of contextual language models shows their strong potential, but in turn demonstrates that there is still room for improvement in furthering their temporal persistence. However, **contextual language models provide a solid alternative among the state-of-the-art solutions**.

4.5.2 Suggestions for the temporal evaluation of text classifiers

In our work we propose a novel evaluation framework which can effectively quantify the persistence of different text classifiers. Driven by the analysis of our results, we next provide a set of key suggestions to take into account for the design of experiments evaluating text classifiers with their temporal persistence as the key objective, as well as to enable furthering research in this direction:

- **Evaluating text classifiers across different time periods.** Where possible, evaluating models across test data pertaining to different time periods can help perform a more comprehensive assessment towards making models more generalisable and persistent over time. Where longitudinal datasets are not available, insights from our study, summarised in Section 4.5.1, can help design experiments with the temporal persistence as the objective.
- **Using complex longitudinal benchmarks** for evaluating the temporal persistence of models. As we observe in our research, there is substantial variation in the complexity of achieving temporal persistence. Indeed, persistence is relatively straightforward in a dataset exhibiting a more stable vocabulary, such as ABRR with Amazon book reviews, and much more complicated for datasets exhibiting more linguistic variations, as is the case with the social media datasets GESD and TESA. Hence, in order to make a fair assessment of models over time, it is ideal to include a range of datasets of different levels of complexity when it comes to language variation.
- **Studying temporally adaptive architectures** that can gradually learn with continually evolving datasets, including supervised models, but ideally focusing on unsupervised models, provided that labels are rarely available for large-scale, longitudinal datasets.
- **Making the most of existing pretrained language models.** Given the cost, complexity and often prohibitive use of computational resources to train new models [Treviso et al., 2023], it is also important to consider the efficiency of existing models in making them persistent over time. Improvement over temporal consistent would

be ideal if it is achieved through computationally efficient resources that do not need expensive resources to train and that can be accessible to a wider range of users.

4.5.3 Limitations

Our study presents a comprehensive analysis of a wide range of combinations of language models and algorithms across three longitudinal datasets. However, in using this wide range of combinations, our objective has not been to achieve the best possible performance in each dataset, but instead to extensively assess temporal persistence through comparative experiments. In doing so, we have not conducted extensive hyperparameter tuning and one could expect that higher performance scores could be achieved by further tuning the models, but not necessarily improving temporal persistence. Likewise, in focusing on models tested across years, we assess model generalisability over time by exclusively leveraging knowledge acquired from the training data. To further leverage unlabelled data from subsequent years, one could consider further investigation into alignment methods [Alkhalifa et al., 2021].

While splitting our datasets into blocks of one year each and experimenting across these years, we present the aggregated results for sets of years which are the same number of years apart from each other, e.g. a gap of 1 year between training and test data aggregates results for a set of years that are one year apart from each other. Consequently, it is worth noting that experiments in larger temporal gaps have fewer experiments to aggregate and can be less stabilised (e.g. the 18-year gap in ABRR includes only experiments between 2000 and 2018, whereas the 1-year gap combines 2000-2001, 2001-2002, 2002-2003, etc.).

Last but not least, and largely constrained by dataset availability, our experiments focus all on binary classification experiments. We can expect that the findings from this study would largely extend to multiclass classification experiments too. However, additional experiments would be needed to further confirm this.

4.6 Conclusion

Despite the evidence of model performance deterioration over time due to changes in data, previous research did not delve into the factors impacting this deterioration. Through extensive exploration of and experimentation on three longitudinal temporal datasets, our comprehensive analysis provides insights into the role of five dimensions on temporal performance: language representations, classification algorithms, time, lexical features and context. Our study identifies the main challenges to focus on towards achieving temporally persistent classification models.

While our study has focused on understanding the capacity and limitations of widely-used classification approaches, and drawing a set of best practices from this analysis, future research could further look into tackling the problem through domain adaptation and transfer learning. The study of the computation complexity and effective design for hyperparameter tuning to support classifier persistence over time was also not considered as part of this study and was left for future work.

Chapter 5

Implementation of Temporally Adaptive Classification Methods

In the rapidly evolving nature of social media and the dynamic nature of people's views, the language and terminology they use also change over time. This temporal evolution poses a significant challenge for natural language processing (NLP) classification models, especially when they are trained on older textual data and then tested on newer information. Such scenarios can result in a dramatic drop in classifier performance. While the field of stance classification has seen considerable advancements in recent years, there has been a notable gap in research when it comes to ensuring the persistence of classifier performance over time. This chapter addresses this critical issue by introducing methods for increasing the temporal persistence of stance classifiers using two large-scale datasets. These datasets provide the foundation for our in-depth exploration into the temporal dynamics of stance classifiers, revealing how their performance degrades as the temporal gap between training and testing data widens.

To address this performance decay over time, we propose an innovative approach based on the temporal adaptation of word embeddings, which are fundamental to training effective stance classifiers. This adaptation enables us to harness readily available unlabelled data from the current time period, eliminating the need for costly and time-consuming annotation efforts. Throughout this chapter, we introduce and compare various approaches to embedding adaptation, seeking the most effective strategy for mitigating performance drop

over time. Our findings highlight the Incremental Temporal Alignment (ITA) model as the top-performing solution for reducing performance decay in temporally adaptive classification methods.

In the subsequent sections, we delve into the details of our research, datasets used, experimental methodology, and the results and insights gained from our investigation into temporal adaptation in stance classification models. This chapter is a critical contribution to the evolving field of NLP, addressing the pressing need for classifiers to maintain their effectiveness in the face of linguistic and temporal changes.

5.1 Introduction

Word meanings drift over time, with new words emerging, words adopting new senses and the frequency of word usage varying. Vocabulary and usage patterns in social media evolve rapidly [Hamoodat et al., 2020], and people's views change over time [Kelman, 1961, Alkhalifa and Zubiaga, 2022]. This can have an impact on stance classification in social media as the data used for training may not generalise well to future data with different patterns. Previous research has either assumed that a classifier trained on static, temporally-restricted data would suffice to track public opinion over time [Deng et al., 2013], or focused on short time periods, analysing stance on trending topics such as Brexit, death penalty or climate change [Simaki et al., 2017, Mohammad et al., 2016]. Our work contributes to research in stance classification by focusing on the impact of a hitherto overlooked aspect: time.

A recent study by Florio et al. [2020] demonstrated that social media hate speech detection models do not perform well on newer data when simply trained on older data. Despite highlighting the existence of this problem, their work did not propose any solutions to the problem. Here we show that this problem is not exclusive to hate speech detection and that it also impacts the performance of social media stance classifiers [Alkhalifa and Zubiaga, 2020]. We collect two longitudinal stance detection datasets that we use for the classifier performance evaluation over time (Section 4). In our experiments we reproduce a real world scenario in which training data remains unchanged while new testing data is generated over years. Our findings indicate that a regular stance classifier can drop up to 18% in relative performance in only five years (Sections 5.6). We then propose novel methodology that makes a social media stance classifier more robust when applied to data that is temporally distant from the training data, which would in turn enable improved tracking of public opinion.

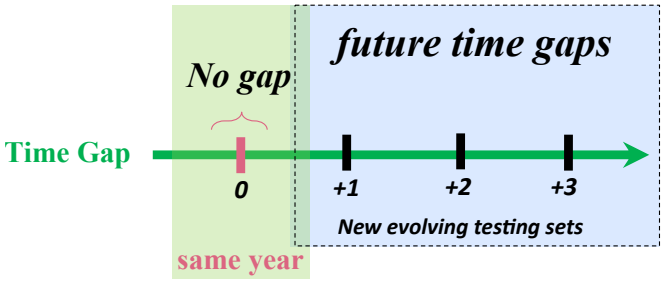


Figure 5.1: An overview of our evolving experimental settings used for monitoring temporal generalisability of models' performance.

While one can choose the costly option of annotating new stance data regularly to re-train a classifier, here we investigate the scenario where one needs to make the most of the originally labelled data, e.g. due to limited resources. Hence we propose to use temporally adapted word embeddings, to re-train the classifier on the unchanged training data. This approach adapts the model to the vocabulary changes that happened over time while making use of readily available unlabelled data. We compare two types of approaches to update the word embeddigns: (1) incrementally updating the same embedding model with new unlabelled data over time and (2) creating a temporally contextualised embedding for the testing year by incrementally aligning a new embedding with preceding embedding models over time. We find that the second approach is more successful at mitigating the performance drop over time. We can obtain improved performance with a substantially reduced performance drop of up to 5%.

5.2 Experimental Settings

In our experimental setup, we focus on the stance classification task, which involves determining whether the author of a post supports or opposes a particular topic. Our goal is to maximize the performance of a stance classifier when tested on new data that are several years apart from training data in the future, as shown in Figure 5.1. This optimization is essential for ensuring the classifier's persistence over time.

Our proposed approach, known as adaptive stance classification, is centered around adapting the word embeddings used in training the classifier. To investigate this approach, we utilize two main types of datasets: (1) **a longitudinal, unlabelled dataset** D, divided into T equally sized temporal slices where $\mathrm{D} = \{D_1, D_2, \ldots, D_T\}$, and (2) **a longitudinal, labelled dataset** of annotated stance tweets representing temporal utterances from a particular domain (e.g., gender equality, healthcare) with a corresponding set of binary stance

labels $s \in \{support, oppose\}$ spanning T years, $Y = \{y_1, ..., y_T\}$, where y_t is a set of tweets from year t.

We use the unlabelled data to generate a sequence of temporal embeddings $\mathbf{X} = \{X_1, X_2, \ldots, X_T\}$, where each $X_t, t \in [1, T]$ contains vector representations of words generated using the temporal slice D_t representing the ground truth of temporal representation at time t.

Our experimental approach follows the methodology outlined in Chapter 3, specifically detailed in Section 3.1.1. We evaluate classifier persistence by training it on data from earlier years (y_i, where $i \in 1, ..., t-1$) and subsequently testing it on data from later years (y_j, where $j \in t+1, ..., T$). However, our primary objective is to continuously update the representation, adapting it to evolving vocabulary changes and maximizing persistence in stance classification for any given pair of years, y_i and y_j in the future.

5.3 Datasets used

We employ two social media datasets in our work: (1) labeled datasets, comprising distantly annotated GESD for gender equality and HCSD for healthcare, which serve as the foundation for evaluating stance detection models, and (2) larger unlabeled datasets, encompassing in-domain GESD and HCSD textual posts, essential for constructing temporal word embedding models. These datasets share a consistent time frame, enabling us to conduct in-depth experiments on stance evolution over time through the labeled datasets, while simultaneously facilitating the incremental adaptation of word embeddings through the unlabeled datasets.

Further, to measure the temporal evolution of the datasets, we compute the Jaccard similarities between the vocabulary observed for each year. Figure 5.2 shows the pairwise Jaccard similarity scores for the two datasets. We can observe that these similarity scores consistently decrease as the distance between the years increases, indicating an increasing variation of vocabulary over time. To gain a comprehensive understanding of these datasets, including their characteristics and label distribution, please refer to the dedicated section 3.2.1 in Chapter 3.

5.4 Methods for Incorporating temporal knowledge into word embeddings

We assess the potential of word embeddings [Mikolov et al., 2013b] to aid classifiers to have a temporally persistent performance, and propose novel methods to further their tempo-

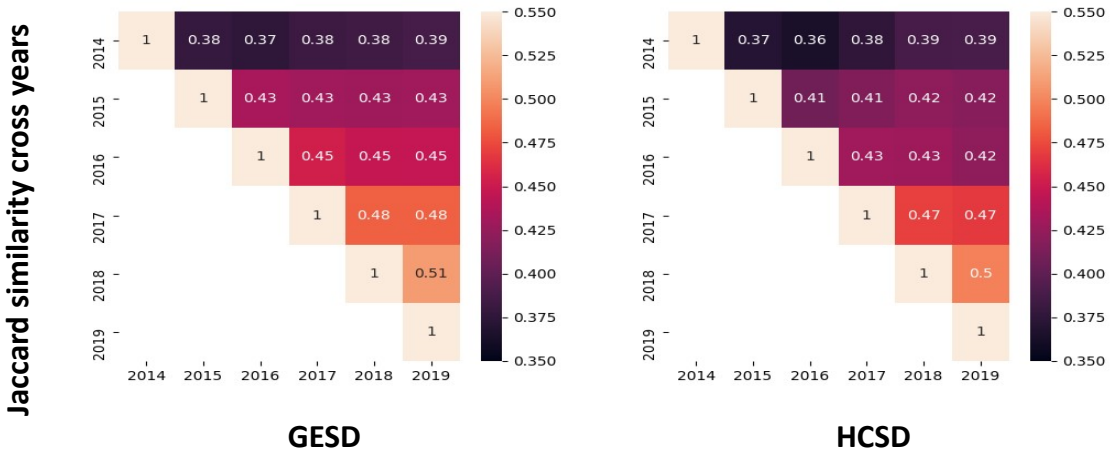


Figure 5.2: Jaccard similarity between vocabularies for test sets of our annotated data collections.

	Embedding	**Learning Strategy**	**Data**
DTE	Discrete	None	S
ITE	Incremental	Model Update	SPT
2TE	Incremental	Model Update	ST
ITA	Incremental	Diachronic Alignment	SPT
2TA	Incremental	Diachronic Alignment	ST

Table 5.1: Temporal embedding methods. Data sources: source year (S), target year (T), and preceding years (P).

ral persistence (see Figure 5.3). We use the CBOW model [Mikolov et al., 2013a], which outperformed skip-gram [Mikolov et al., 2013a] for linguistic change detection [Kulkarni et al., 2015]. We control for other variables (e.g. prediction models, label distributions) by keeping them stable across experiments.

Method 1. Discrete Temporal Embedding (DTE), a baseline method that lacks awareness of the temporal evolution. DTE learns CBOW word vector representations given a collection of tweets pertaining to a particular time frame as input. For example, where our classifier needs to train from data pertaining to year y_1 and test it on y_2, a DTE embedding is generated from the unlabelled data pertaining to y_1. We can formally represent it as follows: Discrete Temporal Embedding (DTE) are the embeddings X generated using temporal slice D_s where s represents the time frame of the source set.

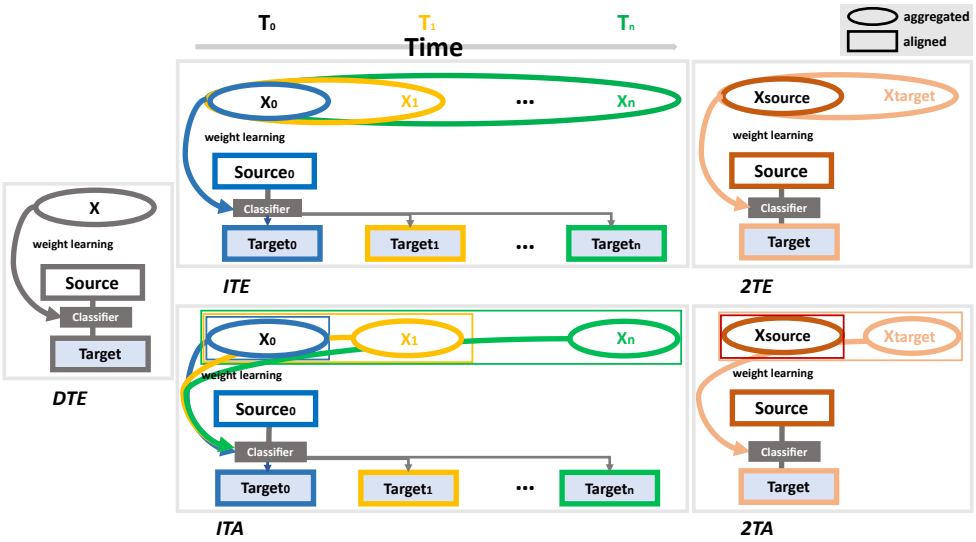


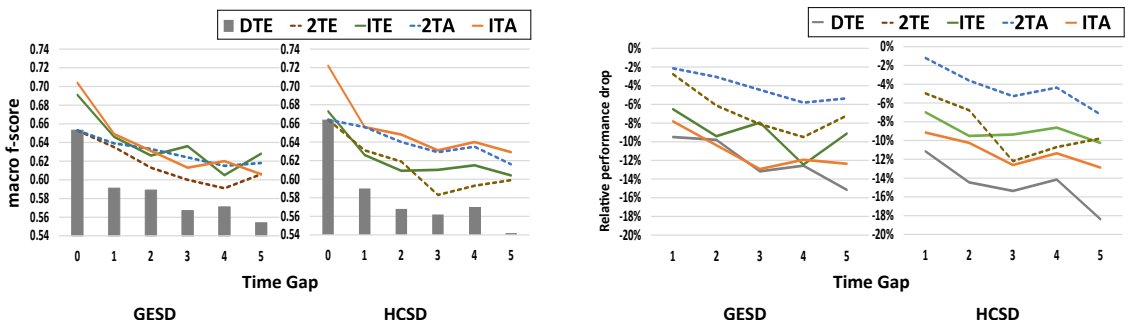
Figure 5.3: Our adaptive stance classification models. DTE represents the baseline, static classifier. The other four are temporally evolving classifiers: 2TA and 2TE only consider source and target periods to model the evolution, whereas ITA and ITE consider all the periods between the source and the target.

In this work we propose four models to incorporate knowledge over time by leveraging unlabelled data.

Method 2. Incremental Temporal Embedding (ITE). New embeddings X are trained using the unlabelled data incrementally aggregated from all years preceding and including the target year, i.e. D_p, where $p \in [2014, t]$ and t is the target year. Then the stance classifier is retrained using the labelled data from the source year represented using the new up-to-date embeddings X.

Method 3. Source-Target Temporal Embedding (2TE). New embeddings X are generated using the unlabelled data aggregated from the source D_s and the target D_t years only, while ignoring all years in between. These embeddings are then used to represent source year training data for the stance classifier.

While ITE and 2TE incorporate temporal knowledge, they do not explicitly handle other phenomena such as semantic shift of vocabulary [Kim et al., 2014], which we anticipate may lead to performance limitations. To address this, we propose alternative methods that perform temporal word alignment. Our proposed solution comes from using a *compass* [Di Carlo et al., 2019] method for temporal alignment. With *compass* each temporal embedding becomes temporally contextualised to the testing year semantic-meaning. With



(a) (b)

Figure 5.4: Performance (a) of temporal embeddings by temporal gap between training and test data (0-5 years) and (b) Relative performance drop. Flat line indicates consistent temporal performance.

Time gap		0	1	2	3	4	5
Gen. Equality	**DTE**	0.653	0.591 (-9.5%)	0.589 (-9.8%)	0.567 (-13.2%)	0.571 (-12.6%)	0.554 (-15.2%)
	ITE	0.691	0.646 (-6.5%)	0.626 (-9.4%)	**0.636** (-8.0%)	0.605 (-12.4%)	**0.628** (-9.1%)
	2TE	0.653	0.635 (-2.8%)	0.613 (-6.1%)	0.600 (-8.1%)	0.591 (-9.5%)	0.606 (-7.2%)
	ITA	**0.704**	**0.649** (-7.8%)	0.631 (-10.4%)	0.613 (-12.9%)	**0.620** (-11.9%)	0.617 (-12.4%)
	2TA	0.653	0.639 (-2.1%)	**0.633** (-3.1%)	0.624 (-4.4%)	0.615 (-5.8%)	0.618 (-5.4%)
Healthcare	**DTE**	0.664	0.590 (-11.1%)	0.568 (-14.5%)	0.562 (-15.4%)	0.570 (-14.2%)	0.542 (-18.4%)
	ITE	0.673	0.626 (-7.0%)	0.609 (-9.5%)	0.610 (-9.4%)	0.615 (-8.6%)	0.604 (-10.2%)
	2TE	0.664	0.631 (-5.0%)	0.619 (-6.8%)	0.583 (-12.2%)	0.593 (-10.7%)	0.599 (-9.8%)
	ITA	**0.722**	**0.656** (-9.1%)	**0.648** (-10.2%)	**0.631** (-12.6%)	**0.640** (-11.4%)	**0.629** (-12.9%)
	2TA	0.664	**0.656** (-1.2%)	0.640 (-3.6%)	0.629 (-5.3%)	0.635 (-4.4%)	0.616 (-7.2%)

Table 5.2: Experiment results by temporal gap between training and test data (0-5 years). Reported values in brackets indicate relative performance drops with respect to same-year (temporal gap of 0) experiments for the same method.

this method, we assume words contextual usage fluctuates over time as in social associations creating subtle meaning drift. For example, the word *'Clinton'* shifted from being related to administration to presidential context over time [Di Carlo et al., 2019]. The *compass* aligns the embeddings of different temporal years using pivot non-shifting vocabularies. It constructs a dynamic temporal context embedding matrix that changes over time, allowing the context embedding to be more time relevant. These settings allow natural selection of vocabularies in terms of temporal contextual words of the target year, and a time-aware representation of the target year in general. We show that this approach is more useful in some cases than model update as the model trained considering the semantic meaning of the target year without additional contexts.

Method 4. Incremental Temporal Alignment (ITA). *X* is incrementally aligned using *compass* from all preceding D_p, where $p \in [2014, t]$.

Method 5. Source-Target Temporal Alignment (2TA). *X* performs temporal alignment using *compass* of D_s and D_t.

We summarise all five models in Table 5.1 and Figure 5.3, which enable us to test the impact of three different parameters: (i) the use of discrete vs incremental embedding models, (ii) the use of different learning strategies (none, model update, diachronic alignment), and (iii) the use of different data sources for building embedding models (source, target and preceding years).

5.5 Experimental Setup

To control for the impact of model choice [Lukes and Søgaard, 2018, Rocha et al., 2008], we consistently use a Convolutional Neural Network (CNN) model with 32 filter and 5-gram region sizes. Our 5-gram kernels encompass a Rectified Linear Unit (ReLU) activation function, and a max-pooling operation. We use a softmax activation function, the Adam optimiser with the learning rate fixed at $2e^{-5}$ and 10 epochs. Sentence length for vector representations is also fixed in all experiments to 32. By keeping the CNN model intact, our aim is to assess the effectiveness of the proposed temporal embedding-based representations for the task, highlighting the practical benefits of temporal adaptation for text classifiers in the context of evolving data.

We experiment with all 21 possible combinations from 2014 to 2019 of y_{source} and y_{target} for training and testing. In each case, we are interested in the temporal gap between train and test data, measured in number of years. For brevity and clarity, we report the mean average performance of models with the same gap (e.g. 4-year gap performance averages 2014-2018 and 2015-2019). In each case, we report the macro-averaged F1 score, as well as the Relative Performance Drop (RPD) to measure the sharpness of the drop in our model, defined as:

$$\text{RPD} = \frac{f_{score_{t_j}} - f_{score_{t_0}}}{f_{score_{t_0}}}$$

Where t_0 represents performance when temporal gap is 0; t_j represents performance when temporal gap is one of 1-5. More details about evaluation matrics are in Chapter 3, Section 3.1.3.

5.6 Results and Discussion

Table 5.2 and Figure 5.4 show the results of our experiments. Results are aggregated by temporal gap. We observe that the best results are obtained for same-period experiments (i.e. temporal gap of 0) and a decrease in performance as the temporal gap increases, i.e. confirming our hypothesis that model persistence drops as training data gets older. Furthermore, the performance drops (percentage, shown in brackets) indicate that the drop has an upwards tendency for larger temporal gaps, demonstrating that the older the training model, the less accurate the model becomes when dealing with new data. Temporal dynamics in the stance datasets can indeed lead to deterioration in model persistence.

When we look at the methods separately, we observe that ITA achieves an overall best performance. This is especially true for the healthcare dataset, where ITA is the best method for all temporal gaps under consideration; for the gender equality dataset, ITA is the best method for small temporal gaps (0-1), and while it achieves competitive performance for larger gaps, 2TA and ITE occasionally achieve better performance.

We observe interesting trends when we look at performance scores and performance drops in conjunction. The use of the baseline DTE, solely relying on source-year embeddings, leads to the lowest performance and also the largest performance drops. This reinforces that embeddings from a particular time period gradually become less useful for subsequent periods, more so when the target period is more distant in time. Among the four proposed methods, ITA yields the best same-period performance, however it is also the method experiencing the highest performance drop for larger temporal gaps; this demonstrates ITA's competitive performance for shorter temporal gaps, but its performance on longer temporal gaps is more uncertain. For methods relying on source and target years, 2TE and 2TA, we observe a modest performance for same-period experiments (equivalent to DTE), which however experience a substantially smaller performance drop for larger temporal gaps. While their performance is not as good for small temporal gaps, they show a good capacity to persist better over time for larger temporal gaps. A look at longer temporal gaps, beyond the 5-year gap considered in our experiments, would be an interesting avenue for future work, e.g. to assess the capacity of 2TA and 2TE to persist further.

In addition, our experiments help us assess the impact of three parameters (see Table 5.1):

Embedding type: we show that the use of an incremental aggregation of embeddings (ITE, 2TE, ITA, 2TA) improves over the use of discrete embeddings (DTE). This is consistent across datasets and temporal gaps.

Learning strategy: our results indicate that the best learning strategy is the use of diachronic alignment (ITA, 2TA), in our case tested using *compass*. With a few exceptions, we observe that these methods generally outperform methods that perform incremental model updates (2TE, ITE), and consistently outperform the lack of a learning strategy by relying on discrete embeddings (DTE).

Data source: the worst performance is for the embedding method solely using the source year (DTE), a baseline method that one would naturally use with static classifiers. Other methods considering additional years lead to improved performance. We observe two main patterns: (1) use of all years preceding the target year (ITE, ITA) lead to improved performance over the use of source and target years (2TE, 2TA), however with a larger performance drop for longer temporal gaps, and (2) use of source and target years only leads to lower performance in short temporal gaps, however with a substantially lower performance drop showing a promising trend towards achieving model persistence.

5.7 Conclusion

Our work demonstrates the substantial impact of temporal evolution on stance classification in social media, with performance drops of up to 18% in relative macro-F1 scores in only five years. We investigate temporal adaptation of word embeddings used to thrain the classifier to mitigate this drop in performance, showing that incrementally aligning embedding data for all years (ITA) leads to the best performance. However, we also find that consideration of only source and target years in the alignment leads to the smallest performance drop with promising trends towards longer term persistence.

Building on this research, we aim to investigate the extent to which different factors (e.g., opinion change, social media use) impact performance. Additionally, we will explore the potential of using few-shot learning to quantify the benefits of labeling small amounts of target data. Future enhancements could also involve exploring temporal adaptation using contextual language models like BERT, RoBERTa and GPT. Incorporating these models could provide deeper insights and more robust performance, leveraging their advanced contextual understanding capabilities despite the increased resource requirements. This would further extend the practical benefits of our incremental method to more sophisticated language models, ensuring their effectiveness in dynamic and evolving linguistic landscapes.

Chapter 6

Comparative Analysis of Classifier Temporal Persistence

To enable a thorough comparison in our research, we've identified a noteworthy pattern: ***text classification models often see a drop in performance when applied to newer data compared to test data [Alkhalifa et al., 2021, Alkhalifa et al., 2023]sampled from within the same time frame of the training data; moreover, performance over time can be seen as a fluctuation based on the dataset characteristics, including the level of familiarity between the train and test set [Alkhalifa et al., 2023]***. To address this, we've launched **LongEval shared task** with a primary focus on sentiment analysis of text from social media platforms over time. Our aim is to encourage a collaborative exploration of how text evolves and how we can maintain text classifier performance over time.

In this chapter, we delve into the methodology of LongEval 2023, which focuses on the development and evaluation of text classifiers submitted by participants. The primary objective is to introduce text classifiers that outperform the RoBERTa model, provided as a baseline classifier in the shared task, in terms of temporal persistence. Our goal is to find classifiers that maintain their performance over time more effectively than the baseline model. Through this collaborative effort, we aim to foster a forward-looking perspective, encouraging broader engagement within the NLP community with the temporal aspects explored in this thesis. These temporal aspects have been relatively underexplored until

now.

As part of this initiative, we have thoughtfully selected a representative subset from the longitudinal sentiment dataset known as LongEval Temporal English Sentiment Analysis (LE-TESA). Detailed information on this data set is available in Chapter 3, Section 3.2. We have initiated joint efforts to assess the persistence of sentiment classifiers using various machine learning techniques chosen by participants. In total, 25 teams registered for the task, and four teams ultimately participated. In the following sections, we provide a more comprehensive analysis and insight into the shared task.

6.1 Shared Task Overview

As the meanings of words and phrases evolve over time, sentiment classifiers may struggle to accurately capture the changing linguistic landscape [Alkhalifa and Zubiaga, 2022], resulting in decreased effectiveness in capturing sentiments expressed in text. Recent research shows that this is particularly the case when dealing with social media data [Alkhalifa et al., 2023]. Understanding the extent of this performance drop and its implications is crucial for maintaining accuracy and reliable sentiment analysis models in the face of linguistic drift. The objective of this task was to quantitatively measure the performance degradation of sentiment classifiers over time, providing insights into the impact of language evolution on sentiment analysis tasks and identifying strategies to mitigate the effects of temporal dynamics. Participants of this task were invited to submit classification outputs of their systems that attempted to mitigate the temporal performance drop.

The aim of Task 2 was ultimately to answer the following research questions:

- **6.RQ1:** ***What types of models offer better short-term temporal persistence?***
- **6.RQ2:** ***What types of models offer better long-term temporal persistence?***
- **6.RQ3:** ***What types of models offer better overall temporal persistence?***

6.2 Description of the task

In this section, we introduce the task of temporal persistence classification, as the focus of a recent shared task [Alkhalifa et al., 2023a]. The goal of this task was to develop classifiers that can effectively mitigate performance drops over short and long periods of time compared to a test set from the same time frame as the training data.

In this task, we employed a lighter version to simplify the analysis. This approach focuses

on within-time splits and short-term and long-term intervals. The short-term splits cover periods of a few months to a couple of years, and the long-term splits span several years to decades. This lighter version serves as an initial step towards addressing the broader problem discussed in Chapter 3 and deeply analysed in the following chapters, allowing us to gauge model performance over different temporal spans with reduced complexity and computational resources. This lighter version involves basic model training and evaluation using a smaller subset of data, employing simpler evaluation metrics to assess model performance, and concentrating on key aspects of temporal variations, such as major shifts in language use or trending topics.

The shared task was in turn divided into two sub-tasks:

Sub-Task 1: Short-Term Persistence: In this sub-task, participants were asked to develop models that demonstrated performance persistence over short periods of time. Specifically, the performance of the models was expected to be maintained within a temporal gap of two years between the training and test data.

Sub-Task 2: Long-Term Persistence: This sub-task focused on developing models that demonstrated performance persistence over a longer period of time. The classifiers were expected to mitigate performance drops over a temporal gap of five years between the training and test data.

By providing a comprehensive training dataset, two practice sets, and three testing sets, the shared competition aimed to stimulate the development of classifiers that can effectively handle temporal variations and maintain performance persistence over different time distances. Participants were expected to submit solutions for both sub-tasks, showcasing their ability to address the challenges of temporal variations in performance.

6.3 Dataset

The **LE-TESA** dataset was used in this work. To assess the extent of the performance drop of models in shorter and longer temporal gaps, we provided training data pertaining to a specific year (2016), as well as test datasets pertaining to a close (2018) and a more distant (2021) year. In addition to measuring performance in each of these years separately, this setup enabled evaluating relative performance drops by comparing performance across years. A comprehensive analysis of the **LE-TESA** dataset is provided in Chapter 3, Section 3.2.

6.4 Evaluation Metrics

The submissions were ranked primarily based on the macro-averaged F1-score. This ranking approach emphasized the overall performance of the sentiment classification models on the testing set. The higher the macro-averaged F1-score, the higher the ranking of the submission. **Relative Performance Drop (RPD):** was introduced in previous work by Alkhalifa et al. [2021] discussed in Chapter 5, and detailed as evaluation measure in Chapter 3, Section 3.1.3. This metric quantified the difference in performance between the "within-period" data and the short- or long-term distant testing sets. RPD was computed as the difference in performance scores between two sets. A negative RPD value indicated a drop in performance compared to the "within-period" data, while a positive value suggested an improvement.

6.5 Results

Our shared task consisted of two subtasks: Short-term persistence (Sub-task A) and Long-term persistence (Sub-task B). Sub-task A focused on developing models that demonstrated performance persistence within a two-year gap from the training data, while Sub-task B required models that exhibited performance persistence over a longer period, surpassing the two-year gap. Additionally, an unlabeled corpora covering all periods of training, development, and testing was provided to teams interested in data-centric approaches. Along with the data, participating teams received python-based baseline code, and evaluation scripts [1]. The shared task progressed through two phases and results are discussed in the following subsections.

6.5.1 Proposed Text classifiers

In the CLEF-2023 LongEval Task 2, only two papers were submitted for comparison to the baseline model. The first paper, by Medina-Alias and Şimşek [2023] presented a text classifier model that explores the longitudinal generalization capabilities of large generative Pre-trained Language Models (PLMs) like GPT3 and T5 for sentiment analysis in social media. They conducted a thorough investigation into various factors affecting classifier performance, including temporal variations in data, model size, and fine-tuning. Their results showed that large generative models outperformed the baseline model (RoBERTa), and limited exposure to training data enhanced temporal robustness. The study also delved into the impact of model size and potential reasons for performance drops, providing valuable insights into building temporally robust sentiment classifiers.

[1] `https://clef-longeval.github.io/`

The second paper, by Ninalga [2023] introduced an innovative approach to text classification by utilizing date-prefixing and a novel data augmentation strategy. Date-prefixing involved including the year of the timestamp as a prefix in the input text, effectively conditioning the language model's output on the text's temporal context. The authors also applied a semi-supervised learning approach, which involved training a teacher model on labeled data to generate pseudo-labels for unlabeled data, followed by training a student model using these pseudo-labels and fine-tuning it on the original labeled data. This strategy aimed to improve performance in the context of temporal data and demonstrated a different approach to addressing temporal issues in text classification.

These two papers provide distinct methodologies for handling temporal aspects in text classification, with Medina-Alias and Şimşek [2023] focusing on large generative PLMs and Ninalga [2023] introducing unique temporal conditioning techniques and data augmentation strategies. The comparison to the baseline model highlights their contributions to the field.

6.5.2 Practice phase

The initial phase was the practice phase, where participants received three distantly annotated sets, training set, within time practice set and short-term practice set. The training set was used for model training, while the two labeled practice set allowed participants to refine their systems before the subsequent phase. Moreover, we limited the sharing practice sets to within-time (Practice-2016) and single distance practice sets the short-term set (Practice-2018). This decision was made because participants were requested to take part in both sub-tasks and reduce over-fitting. The results of this phase were not considered in final models ranking.

Team Name	Model Name	F1 Score Within	F1 Score Short	Overall Drop	Overall Score
Medina-Alias and Şimşek [2023]	**GPT3**	0.8244 (1)	0.7976 (1)	-0.0325 (2)	0.811
saroyehun	unknown	0.8170 (2)	0.7917 (2)	-0.0310 (1)	0.8043
Alkhalifa et al. [2023]	**Baseline-RoBERTa**	*0.7879 (3)*	*0.7611 (3)*	*-0.0340 (3)*	*0.7745*

Table 6.1: Performance comparison for the practice set

As it can be seen from Table 6.1, the fine-tuned version of **GPT3** introduced by Medina-Alias and Şimşek [2023] exhibited more temporal persistence, surpassing the **Baseline-RoBERTa** model with the highest scores in F1 Score Within (0.8244) and F1 Score Short (0.7976), as well as the highest Overall Score (0.811). **saroyehun** also demonstrated remarkable performance achieving the lowest Overall Drop (-0.0310), as well as outperforming the **RoBERTa-Baseline** model in F1 Score Within (0.8170) and F1 Score Short (0.7917). The results highlight the potential of both **GPT3** and **saroyehun**'s submissions for en-

hancing the baseline model's results. Though, it was not known which model used for **saroyehun** submission.

6.5.3 Evaluation phase

During the evaluation phase, participants were provided with three human-annotated testing sets, namely Test-within, Test-short and Test-long (See Chapter 3, Section 6.3 for dataset details). The performance of participants on this phase was used to determine the overall rankings on the task.

Team Name	Model Name	F1 Score Within	F1 Score Short	F1 Score Long	RPD Within-Short	RPD Within-Long	Overall Drop	Overall Score
Medina-Alias and Şimşek [2023]	**T5**	0.7377 (2)	0.6739 (3)	0.6971 (1)	-0.0866 (5)	-0.0550 (3)	-0.0708 (4)	0.7029
Alkhalifa et al. [2023]	**Baseline-RoBERTa**	*0.7459* (1)	*0.6839* (1)	*0.6549* (4)	*-0.0830* (4)	*-0.1220* (5)	*-0.1025* (5)	*0.6949*
Ninalga [2023]	**Bernice**	0.7246 (3)	0.6771 (2)	0.6751 (3)	-0.0656 (1)	-0.0683 (4)	-0.0669 (3)	0.6923
saroyehun	unknown	0.7203 (4)	0.6674 (4)	0.6874 (2)	-0.0735 (2)	-0.0457 (2)	-0.0596 (2)	0.6917
pakapro	unknown	0.5033 (5)	0.4648 (5)	0.4910 (5)	-0.0765 (3)	-0.0243 (1)	-0.0504 (1)	0.4863

Table 6.2: Performance comparison for the evaluation set.

Within Time Performance: From Table 6.2, the **Baseline-RoBERTa** model achieved the highest F1 score within the Test-within dataset (0.7459), indicating strong immediate performance on the dataset used for training. This suggests that the model has a good understanding of the patterns in the training data.

Short-term Temporal Persistence: The **Baseline-RoBERTa** model attained the highest short-term F1 Score (0.6839) among all the teams, demonstrating its ability to capture short-term patterns effectively. However, **Bernice**, introduced by *Ninalga [2023]*, recorded the lowest short-term Relative Performance Drop (RPD) value (-0.0656), indicating a smaller drop in performance between the Test-within and Test-short datasets. This suggests that while **Bernice** may not have the highest short-term F1 score, it maintains more consistent performance over short-term intervals, offering better short-term temporal persistence.

Long-term Temporal Persistence: The **T5** model by *Medina-Alias and Şimşek [2023]* achieved the highest F1 score on the Test-long dataset (0.6971), indicating strong performance in capturing long-term patterns. However, the model by *pakapro* recorded the smallest long-term RPD value (-0.0243), suggesting a smaller drop in performance over long-term intervals. This indicates that while **T5** has the highest long-term F1 score, *pakapro* offers better long-term temporal persistence despite its lower F1 score.

Overall Temporal Persistence: Considering the overall scores, the **T5** model by *Medina-Alias and Şimşek [2023]* achieved the highest overall score (0.7029) with a moderate overall

RPD (-0.0708), indicating better overall temporal persistence compared to other models. However, *pakapro* exhibited the best performance in terms of overall temporal persistence based on the Overall Drop metric, indicating that *pakapro*'s approach maintains more consistent performance over time despite its lower F1 scores.

Systems Temporal Ranking: The **Baseline-RoBERTa** model ranks first in within-time and short-term F1 scores but drops to fourth place in long-term F1 scores. The **T5** model by *Medina-Alias and Şimşek [2023]* and **Bernice** by *Ninalga [2023]* interchange the second and third positions in the within-time F1 score and short-term F1 score categories, suggesting a relatively consistent ranking between these two models within these specific categories. *saroyehun* consistently ranks fourth in both within-time F1 score and short-term F1 score. *pakapro* shows the lowest performance among all models, ranking fifth in all three F1 score categories but demonstrating consistent performance across different timeframes.

In summary, the best model overall is **T5** by *Medina-Alias and Şimşek [2023]*, showing a better overall F1 score and greater temporal persistence than the **Baseline-RoBERTa** model. Additionally, the **Baseline-RoBERTa** model performed best in short-term temporal persistence, and *pakapro* shows promise for long-term temporal persistence despite not having the highest long-term F1 score. This comprehensive analysis highlights the strengths and weaknesses of each model in terms of both immediate and temporal performance.

6.6 Discussion

Only two out of the four teams have submitted technical reports for their used models. In the following, we delve into the discussion and interpretation of the findings concerning the three research questions we raised in relation to our classification task. These interpretations are solely based on the evaluation matrix, which is further explained in Section 6.4.

- Regarding **6.RQ1**, which aimed to identify the types of models offering better short-term temporal persistence, we observed distinct differences between the practice and evaluation phases. During the evaluation phase, the **Baseline-RoBERTa** model achieved the highest short-term F1 Score (0.6839), indicating its strong performance in maintaining consistency over a shorter time frame. Additionally, the model by *Ninalga [2023]* using **Bernice**, which is specifically pre-trained on Twitter data, demonstrated the smallest short-term RPD (-0.0656). This suggests that domain adaptation plays a crucial role in enhancing short-term temporal persistence.

  In the practice phase, the **GPT3** model introduced by *Medina-Alias and Şimşek [2023]* showed superior performance with the highest short-term F1 Score (0.7976)

and the highest Overall Score (0.811). This result underscores the potential of generative AI models like **GPT3** in short-term contexts, emphasizing the effectiveness of robust pre-training and fine-tuning strategies for short-term performance.

The comparison between **Baseline-RoBERTa** and **Bernice** highlights the significance of domain adaptation in improving model performance generalizability. While **Baseline-RoBERTa**'s strong initial performance underscores the effectiveness of robust pre-training and fine-tuning strategies, **Bernice**'s domain-specific pre-training showcases the benefits of tailoring models to the specific characteristics of the target data. This adaptation not only improves short-term performance but also enhances the model's ability to generalize across different temporal contexts.

In summary, the findings from **RQ1** demonstrate that while general-purpose models like **Baseline-RoBERTa** perform well in the short term, domain-adapted models like **Bernice** and robustly pre-trained generative models like **GPT3** offer enhanced temporal persistence and generalizability, particularly in specialized domains such as social media.

- Regarding **6.RQ2**, which investigated the models offering better long-term temporal persistence, we observed that the **T5** model by *Medina-Alias and Şimşek [2023]* achieved the highest long-term F1 Score (0.6971) in the evaluation phase. This indicates its superior ability to maintain performance over an extended period. The **T5** model's architecture, which includes strong sequence-to-sequence processing capabilities, aligns well with the nature of long-term data evolution. This model's ability to process and generate sequences effectively contributes to its high performance in long-term contexts.

  In terms of long-term RPD, the *pakapro* model demonstrated the smallest value (-0.0243), suggesting its potential for maintaining performance stability over time. This indicates that while the **T5** model excels in long-term F1 performance, models like *pakapro*, with their minimal performance drop, offer more consistent long-term temporal persistence.

- Regarding **6.RQ3**, which aimed to identify the models offering better overall temporal persistence, we found that the **T5** model by *Medina-Alias and Şimşek [2023]* ranked as the top-performing system in the evaluation phase, achieving the highest overall score (0.7029) and a relatively low overall RPD (-0.0708). This indicates that the **T5** model provides a balanced approach, maintaining high performance across various temporal datasets.

In the practice phase, the **GPT3** model again demonstrated its effectiveness, achieving the highest overall score (0.811) with strong F1 scores across both within-time and short-term datasets. This further supports the idea that robustly pre-trained generative models like **GPT3** and **T5** are well-suited for maintaining consistent performance over time.

Interestingly, the *pakapro* model demonstrated promising results for long-term temporal persistence in the evaluation phase, despite not achieving the highest long-term F1 Score. The *pakapro* model exhibited the best performance in terms of overall temporal persistence based on the Overall Drop metric, indicating that *pakapro*'s approach maintains more consistent performance over time despite its lower F1 scores.

The evaluation results highlight that different models excel in various aspects of temporal persistence. The **RoBERTa** model is best suited for tasks requiring strong short-term performance, while the **T5** model excels in long-term pattern recognition and overall temporal persistence. Additionally, the practice phase results underscore the effectiveness of robust pre-training and fine-tuning strategies, as evidenced by the performance of the **GPT3** model.

By delving into the evaluation matrix results, we provided valuable insights into the performance trends observed among the participating systems. However, it is important to acknowledge that the absence of submissions from certain systems may have influenced the overall interpretation of our findings. To address this limitation and ensure a more comprehensive analysis, we have made our leaderboard available for future submissions on Codalab [2]. This initiative aims to facilitate a more robust and unbiased assessment of the temporal persistence of text classifiers within the research community, allowing for continuous updates and improvements based on new data and models.

6.7 Conclusion

Overall, these findings emphasize the importance of evaluating temporal persistence in model performance. The identified models showcase varying levels of persistence in both short-term and long-term scenarios. These insights provide valuable guidance for future research and development efforts aimed at improving text classifier temporal persistence.

For future shared tasks, we aim to incorporate evolving training sets and extend our investigation of temporal persistence to additional tasks, including stance detection and topic

[2] `https://codalab.lisn.upsaclay.fr/competitions/12762`

categorization. This expansion will enable a broader understanding of how different models perform over time across diverse applications, fostering the development of more robust and adaptable text classification solutions.

Chapter 7

Conclusions and Further Work

In this thesis, we embarked on a comprehensive exploration of various aspects related to text classification, with a primary focus on understanding the temporal persistence of text classifiers and language use dynamics. The thesis makes a comprehensive and unique contribution in this understudied research challenge, establishing a new benchmark evaluation methodology, producing new datasets and resources, quantifying the extent of the problem, assessing the potential for mitigation and widening participation of the scientific community through a shared task.

The following subsections provide an overview of these contributions, chapter by chapter, summarizing the key findings from each chapter, followed by a concise summary of results addressing the initial research questions presented in Chapter 1, Section 1.2. The chapter concludes by outlining avenues for future research directions.

7.1 Overview of Chapters

Chapter 2 delves into the theoretical foundations of text classification, covering essential concepts and methodologies that underpin our investigation. It further sets out our aim of achieving an in-depth understanding of the area of temporal data analysis using textual data.

Chapter 3 introduces the methodologies employed in this thesis, including data preprocessing, feature engineering, and machine learning techniques used as a foundation for our

experiments in the following chapters. We discussed our dataset characteristics, highlighting their dynamic features to comprehend the impact of dynamic language use in building persistent text classifiers. We also outlined the tools and strategies used in our analysis. In the same chapter, we presented our approach to data collection and annotation, introducing the TESA dataset as a valuable resource for our thesis. We outlined the steps we took to preprocess and sample the data, ensuring its quality and representativeness for our experiments as a baseline longitudinal dataset containing both large-scale distantly annotated texts and human-annotated evaluation sets.

Chapter 4 sets the stage for understanding the challenges of temporal persistence in text classifiers and provides a comprehensive methodology for evaluating their performance across different temporal scenarios. The use of various metrics and experimental setups enables a deeper analysis of the factors influencing model performance over time. Our assessment consists of a series of five experiments with the following findings:

- Performance of text classification models tends to degrade over time due to changes in data.
- The ability to estimate a model's performance over time based on its performance over a restricted time period and the linguistic characteristics of the dataset.
- RoBERTa presents best persistence text classifier due to its enhanced generalizability across datasets, potentially attributed to its extended vocabulary.
- Datasets exhibiting low temporal vocabulary growth demanding less effort to achieve temporal persistence as in ABRR where testing older models on future data (even spanning an 18-year gap) yielded superior performance.
- In our selected datasets, our findings showed that a robust text classifier design could be guided by the model-algorithm pairing that show high performance at year 0. However, the same approach did not hold true in Chapter 4, where the winning text classifier outperformed the baseline model in the long term but failed to persist within and across shorter-term evaluation datasets.

  In summary, Chapter 4 introduces a comprehensive methodology for assessing the performance of text classifiers under varying temporal text classifier models and datasets. Followed by conducting an in-depth analysis of the factors that influence model performance over time.

Chapter 5 presents a novel solution to enhance text classifier performance by proposing a

new approach to address the problem of performance degradation over time. In this chapter, we focus on incorporating temporal information into word embeddings and presented findings from experiments comparing different methods across various temporal gaps. Key findings from this chapter include:

- Stance classification performance deteriorates as the temporal gap between training and testing data increases, emphasizing the need to consider temporal dynamics in social media data analysis.
- Among the proposed methods, Incremental Temporal Alignment (ITA) performs the best for same-period experiments but experiences a more significant performance drop for longer temporal gaps.
- Methods that rely on source and target years (2TE and 2TA) demonstrate modest performance for same-period experiments but exhibit a smaller performance drop for longer temporal gaps, indicating their potential for achieving model persistence over time.
- Incremental update of embeddings generally improves performance compared to using discrete embeddings, and methods employing diachronic alignment methods that outperform those using static word embedding tuned for the time-restricted dataset.
- The choice of data source for building embedding models also affects performance, with methods considering all years preceding the target year leading to improved performance but a larger performance drop for longer temporal gaps.

In summary, Chapter 5 highlights the challenges posed by temporal dynamics in social media data and suggests that a combination of incremental embedding updates and diachronic alignment methods can help mitigate these challenges for better model persistence over time.

Chapter 6 is dedicated to the evaluation of text classifier performance as a shared task to allow a collaborative effort by the NLP research community and aimed to compare our proposed baseline with existing technology proposed by the participants. The shared task focused on measuring the temporal persistence of sentiment analysis models when applied to social media data over different timeframes. Our evaluation metrics, including F1-score and Relative Performance Drop (RPD), which are consistent with the experiment settings introduced in Chapter 4 and the implementation of methods to improve text classifier performance introduced in Chapter 5, allowed us to assess the performance of submitted models and answer three critical research questions. In this chapter, we present the results

of our shared task. We discuss the performance of different models in terms of short-term, long-term, and overall temporal persistence. Notably, the baseline model demonstrates strong short-term persistence, while models such as Medina-Alias and Şimşek [2023] exhibited improved long-term persistence. We also highlight the promising performance of models like *Pakapro* in terms of overall temporal persistence.

In conclusion, our research contributes to the field of text classification by emphasizing the importance of evaluating text classifiers' temporal persistence. The findings from our shared task shed light on which types of classifiers are more suitable for maintaining persistent performance over time.

7.2 Summary of key findings

RQ1. How does the dynamics of language use impact text classifiers? or, in other words, can we capture the temporal dynamics of language use so as to adapt text classifiers?

Answer: The dynamic nature of language profoundly affects text classifiers, leading to a degradation in performance over time. Providing a theoretical background about this field in Chapter 2, Chapters 3 and 4 provide essential insights into capturing these temporal dynamics, highlighting the challenges and implications in more practical terms.

RQ2. How can we assess the influence of time on text classifier performance using a longitudinal dataset?

Answer: Chapter 4 details our assessment of the influence of time on text classifier performance using longitudinal datasets. Key findings include the tendency of text classification models to degrade over time, the ability to estimate performance changes based on a restricted time period, and the impact of linguistic characteristics on performance. These insights provide a comprehensive understanding of the factors influencing text classifier performance over time.

RQ3. What machine learning and statistical methods can facilitate the maintenance of classifier accuracy in the context of evolving language use?

Answer: Addressing this question, the evaluation (Chapter 4) and comparison (Chapter 6) chapters provide insights into machine learning and statistical methods for maintaining classifier accuracy amid evolving language use. These chapters rank text classifiers based on their top performance over the past and future, short and long term, as well as within and overall performance. This approach provides a comprehensive understanding of the effectiveness of different classifiers under diverse temporal scenarios.

RQ4. How can we maintain our classifiers' performance by circumventing the impact of evolving language and patterns?

Answer: Addressing RQ4, Chapter 5 emphasizes the challenges posed by evolving vocabulary and knowledge. The findings highlight the importance of considering temporal dynamics in social media data analysis, and the proposed methods, such as Incremental Temporal Alignment (ITA), demonstrate their potential to enhance accuracy over time. This chapter provides insights into maintaining classifiers' accuracy in the context of evolving language use.

7.3 Limitations

The scope of this study, while extensive, is marked by constraints that stem primarily from the reliance on specific datasets covering periods of time that could have ideally been longer, like TESA. These datasets, although rich in data, may not fully encapsulate the broader and more rapid linguistic variations seen in more localized or evolving dialects or entire languages. Such a dataset-centric approach, while informative, has its limitations in capturing the entire spectrum of language evolution, which could affect the generalizability of our findings.

Further, the adaptation model introduced in Chapter 5 is not intended to achieve state-of-the-art performance. Rather, it aims to demonstrate that significant improvements can be made by adapting language embeddings to changes in language use over time, showing the impact of evolving updates in word embedding on text classifier performance and highlighting the practical benefits of temporal adaptation for text classifiers in dynamic environments. Consequently, leveraging contemporary language models, such as contextual language modeling (BERT, RoBERTa, GPT), to apply adaptation approaches was not examined and is left for future work.

Another notable limitation is the dynamic and unpredictable nature of online communication. The languages and terminologies on digital platforms evolve rapidly, posing a challenge for the models developed in this thesis to adapt promptly. For example, models trained on existing datasets might not readily anticipate the emergence of new Internet jargon or trending phrases, potentially limiting their predictive accuracy.

The computational demands of implementing advanced machine learning and deep learning models like RoBERTa and ITA are also significant. These complex models require substantial computational power and resources, which may not be accessible in all research environments. This limitation poses a challenge, particularly for extensive model training

and optimization processes.

Furthermore, the reliance on advanced classification models often results in a lack of interoperability. Their black box nature, while powerful in predictive capabilities, limits the depth of our understanding of their internal processes. This aspect is particularly crucial in applications that demand clarity and accountability.

These limitations, however, pave the way for numerous future research opportunities. By addressing these limitations, we aim to contribute further to the field of NLP, enhancing the sustainability and persistence of text classifiers in the longitudinal evolution of natural language processing.

Longitudinal analyses of sentiment can be valuable for longstanding topics or events, however labelling data can be unaffordable if done through manual labelling. Our methodology can enable the longitudinal analysis of events, for example studying how sentiment has changed over events such as Brexit [Hürlimann et al., 2016].

Recent research is showing that machine learning models for classification of social media posts fade over time in terms of performance [Florio et al., 2020, Alkhalifa et al., 2021]. Where one uses labelled data from a particular point in time for training machine learning models, performance of these models drops when they are applied on temporally distant data, among others because language and communication in social media changes over time and models need to be adapted to capture that change. Availability of longitudinally labelled datasets can help further study this problem in the context of sentiment analysis.

Given that our methodology is not restricted to a specific language, it can be applied to a wide range of languages to enable multilingual analyses of social media, as well as development of multi- [Dashtipour et al., 2016] and cross-lingual [Zhou et al., 2016] models for text classification. While we haven't tested our methodology on other social media platforms beyond Twitter (apart from the ABRR dataset obtained from Amazon), we anticipate that our approach can be effectively applied to other platforms to enable broader investigation of text classification over time and across platforms.

Tweets provided in our dataset are labelled through distant supervision, as has been validated in previous work, but being an imperfect method it also leads to some noisy labels. Manually labelled, longitudinal datasets would be a gold mine to further enable and extend this research, but can be very costly and unaffordable for most researchers.

7.4 Future Research Directions

Our interdisciplinary approach, integrating insights from linguistics, sociology, and other relevant NLP fields, opens up several promising avenues for future research. The field of NLP is indeed a fascinating and dynamic area of research that needs to be explored more. As language use constantly evolves due to cultural shifts, technological advancements, and societal changes, there are several intriguing research avenues within this domain.

7.4.1 Temporal Adaptation Approaches

In the future, we intend to expand our investigation in the future to include **evolving human-annotated training sets** as well as explore methods for **tracking and understanding textual shifts** in online conversations, social media, or news articles, enabling us to better simulate the dynamics of real-world data. Additionally, we plan to extend our analysis to other text classification tasks, such as **hate speech detection** and **fake news detection**, to gain a more comprehensive understanding of temporal persistence in machine learning models.

The **cross-domain and cross-linguistic efficacy** of the models developed in this study is another area that remains to be thoroughly explored. For example, their applicability and adaptability to specialized domains such as legal or medical texts, or to languages other than those tested, need further investigation. Other future research directions include exploring the impact of other factors on **performance drop** and investigating the potential of **few-shot learning** for this problem. Broader directions include investigating the **ethical implications** of using evolving language in NLP applications. Issues related to bias, fairness, and inclusion in models that adapt to linguistic changes are critical areas for exploration.

Moreover, assessing more statistical approaches for **automatically updating lexicons** and **expanding vocabularies** to adapt to emerging words, phrases, and linguistic trends is crucial. This could involve exploring word embeddings or other representation models that capture semantic shifts, including the use of increasingly popular **large language models (LLMs)**. LLMs, like GPT3 and BERT, offer significant potential for enhancing data annotation processes and adapting to new linguistic trends. However, they also share challenges such as bias, computational resource demands, and the need for continual updates to remain relevant.

7.4.2 Generative AI Models

In [Pangakis et al., 2023], the authors demonstrated that while generative large language models (LLMs) like GPT4 show significant potential for augmenting text annotation tasks, their performance is highly variable and must be validated against human-generated labels due to persistent challenges such as prompt quality, text data idiosyncrasies, and conceptual difficulty.

This variability highlights the importance of a robust validation workflow for effective deployment, leading to more work involving humans in the loop [Wang et al., 2024]. This requirement is particularly critical in scenarios involving **temporal changes to language use over time**, making **temporal adaptation methods crucial**. As language evolves, **models trained on older data struggle with new terminologies and context shifts**, such as those related to COVID-19 developments post-2020.

With the advent of generative AI models, future research can **bridge the gap between traditional NLP methods and the advancements brought by state-of-the-art generative AI technologies**. Leveraging LLMs can significantly enhance the robustness and adaptability of text classifiers. However, despite their potential, **generative AI models face challenges such as bias and explainability**. These issues are particularly pertinent in stance detection, where the model's output can be highly sensitive to the inherent biases present in training data annotated to represent events at a particular time period. Future research should focus on developing methods to mitigate these biases, ensuring that generative models provide fair and balanced outputs across different contexts.

Additionally, **integrating multimodal data**, such as text, images, and videos, offers an opportunity to enhance the contextual understanding of text classifiers. By utilizing the capabilities of generative AI, researchers can **create more nuanced and comprehensive models that better capture the complexities of real-world data**.

In conclusion, our work serves as a foundation for improving the temporal persistence of text classifiers, which is crucial in the dynamic landscape of natural language processing. We hope that our findings will inspire further research and innovation in this area, leading to more sustainable, reliable, and persistent text classifiers. Furthermore, a better understanding of dataset complexity and its influence on the performance of text classifiers will enhance the development of future models.

References

Aseel Addawood and Masooda Bashir. "What Is Your Evidence?" A Study of Controversial Topics on Social Media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11. Association for Computational Linguistics, 2016. 10.18653/v1/W16-2801. URL http://aclweb.org/anthology/W16-2801.

Aseel Addawood, Amirah Alshamrani, Amal Alqahtani, Jana Diesner, and David Broniatowski. Women's driving in saudi arabia–analyzing the discussion of a controversial topic on twitter. In *2018 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction and Behavior Representation in Modeling and Simulation, BRiMS 2018*, 2018.

Mahmoud Al-Ayyoub, Abdullateef Rabab'ah, Yaser Jararweh, Mohammed N. Al-Kabi, and Brij B. Gupta. Studying the controversy in online crowds' interactions. *Applied Soft Computing Journal*, 66:557–563, 2018. ISSN 15684946. 10.1016/j.asoc.2017.03.022. URL https://doi.org/10.1016/j.asoc.2017.03.022.

Abeer Aldayel and Walid Magdy. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20, 2019. ISSN 25730142. 10.1145/3359307. URL https://dl.acm.org/citation.cfm?id=3359307.

Rabab Alkhalifa and Arkaitz Zubiaga. Qmul-sds@ sardistance2020: Leveraging network interactions to boost performance on stance detection using knowledge graphs. *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020*, page 198, 2020.

Rabab Alkhalifa and Arkaitz Zubiaga. Capturing stance dynamics in social media: open challenges and research directions. *International Journal of Digital Humanities*, pages 1–21, 2022.

Rabab Alkhalifa, Adam Tsakalidis, Arkaitz Zubiaga, and Maria Liakata. QMUL-SDS @ DIACR-ITA evaluating unsupervised diachronic lexical semantics classification in italian. *CoRR*, abs/2011.02935, 2020a. URL https://arxiv.org/abs/2011.02935.

Rabab Alkhalifa, Theodore Yoong, Elena Kochkina, Arkaitz Zubiaga, and Maria Liakata.

QMUL-SDS at checkthat! 2020: Determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions. *CoRR*, abs/2008.13160, 2020b. URL https://arxiv.org/abs/2008.13160.

Rabab Alkhalifa, Iman Bilal, Hsuvas Borkakoty, Jose Camacho-Collados, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, et al. Extended overview of the clef-2023 longeval lab on longitudinal evaluation of model performance. In *CEUR Workshop Proceedings*, volume 3497, pages 2181–2203. CEUR-WS, 2023a.

Rabab Alkhalifa, Iman Bilal, Hsuvas Borkakoty, Jose Camacho-Collados, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, Elena Kochkina, Maria Liakata, Daniel Loureiro, Harish Tayyar Madabushi, Philippe Mulhem, Florina Piroi, Martin Popel, Christophe Servan, and Arkaitz Zubiaga. LongEval: Longitudinal evaluation of model performance at CLEF 2023. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in information retrieval*, pages 499–505, Cham, 2023b. Springer Nature Switzerland. ISBN 978-3-031-28241-6.

Rabab Alkhalifa, Hsuvas Borkakoty, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Tobias Fink, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, David Iommi, Maria Liakata, Harish Tayyar Madabushi, Pablo Medina-Alias, Philippe Mulhem, Florina Piroi, Martin Popel, Christophe Servan, and Arkaitz Zubiaga. Longeval: Longitudinal evaluation of model performance at clef 2024. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part VI*, page 60–66, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-56071-2. 10.1007/978-3-031-56072-9_8. URL https://doi.org/10.1007/978-3-031-56072-9_8.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, pages 27–32, 2021.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60 (2):103200, 2023. ISSN 0306-4573. https://doi.org/10.1016/j.ipm.2022.103200. URL https://www.sciencedirect.com/science/article/pii/S0306457322003016.

Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, 71:100663, 2021.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, AS Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. Cats Rule and Dogs Drool!: Classifying Stance in Online Debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, WASSA '11, pages 1–9. Association for Computational Linguistics, 2011. URL `http://www.aclweb.org/anthology/W11-1701`.

Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, ..., and Paolo Rosso. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter. *Data and Knowledge Engineering*, 124:101738, 2019. ISSN 0169023X. 10.1016/j.datak.2019.101738. URL `https://doi.org/10.1016/j.datak.2019.101738`.

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, pages 226–234. Association for Computational Linguistics, 2009. ISBN 9781617382581. 10.3115/1687878.1687912. URL `http://www.cs.pitt.edu/mpqa.`

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290, 2018.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, 2016.

R. Kelly Aune and Toshiyuki Kikuchi. Effects of language intensity similarity on perceptions of credibility relational attributions, and persuasion. *Journal of Language and*

Social Psychology, 12(3):224–238, 1993. ISSN 15526526. 10.1177/0261927X93123004. URL http://journals.sagepub.com/doi/10.1177/0261927X93123004.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518, 2017.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, 2018a.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, 2018b.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, 2017.

Pierpaolo Basile and Barbara McGillivray. Exploiting the web for semantic change detection. In *International Conference on Discovery Science*, pages 194–208. Springer, 2018.

Douglas Biber. Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2):97–116, 2006.

Miroslav Blšták and Viera Rozinajová. Machine learning approach to the process of question generation. In *International Conference on Text, Speech, and Dialogue*, pages 102–110. Springer, 2017.

Francesco Bodria, André Panisson, Alan Perotti, and Simone Piaggesi. Explainability methods for natural language processing: Applications to sentiment analysis (discussion paper). In *Proceedings of the Italian Symposium on Advanced Database Systems*, 2020.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word

vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Tabitha Bonilla and Cecilia Hyunjung Mo. The evolution of human trafficking messaging in the united states and its effect on public opinion. *Journal of public policy*, 39(2): 201–234, 2019.

Luís Borges, Bruno Martins, and Pável Calado. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26, 2019.

J. Bruin. newtest: command to compute new test @ONLINE, February 2011. URL `https://stats.oarc.ucla.edu/stata/ado/analysis/`.

Michael Burgoon, Stephen B Jones, and Diane Stewart. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity1. *Human Communication Research*, 1(3):240–256, 1975.

Mats Byrkjeland, Frederik Gørvell de Lichtenberg, and Björn Gambäck. Ternary twitter sentiment classification with distant supervision and sentiment-specific word embeddings. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, 2018.

K. Chakraborty, S. Bhattacharyya, and R. Bag. A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems*, 7(2):450–464, 2020.

Li-Chen Cheng, Kuanchin Chen, Ming-Chu Lee, and Kua-Mai Li. User-defined swot analysis–a change mining perspective on user-generated content. *Information Processing & Management*, 58(5):102613, 2021.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1724. Association for Computational Linguistics, 2014.

Yun-Shiuan Chuang, Rheeya Uppaal, Yi Wu, Luhang Sun, Makesh Narsimhan Sreedhar, Sijia Yang, Timothy T Rogers, and Junjie Hu. Evolving domain adaptation of pretrained language models for text classification. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

Ryan L Claassen and Benjamin Highton. Does policy debate reduce information effects in

public opinion? analyzing the evolution of public opinion on health care. *The Journal of Politics*, 68(2):410–420, 2006.

Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. Fake news types and detection models on social media a state-of-the-art survey. In *Asian Conference on Intelligent Information and Database Systems*, pages 562–573. Springer, 2020.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online, July 2020. Association for Computational Linguistics. 10.18653/v1/2020.acl-main.157. URL `https://www.aclweb.org/anthology/2020.acl-main.157`.

Michael D Conover, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Partisan asymmetries in online political activity. *EPJ Data Science*, 1(1):6, 2012.

Paul Cook and Suzanne Stevenson. Automatically Identifying Changes in the Semantic Orientation of Words. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, 2010.

Pierre Cossette. The study of language in organizations: A symbolic interactionist stance. *Human Relations*, 51(11):1355–1377, 1998.

Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M Almeida, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing & Management*, 58(3):102481, 2021.

Eleonora D'Andrea, Pietro Ducange, Alessio Bechini, Alessandro Renda, and Francesco Marcelloni. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116:209 – 226, 2019. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2018.09.009. URL `http://www.sciencedirect.com/science/article/pii/S0957417418305803`.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771, 2016.

William Deitrick and Wei Hu. Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks. *Journal of Data Analysis and Information Processing*, 01 (03):19–29, 2013. ISSN 2327-7211. 10.4236/jdaip.2013.13004. URL `http://dx.http://www.scirp.org/journal/jdaip`.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, 2019.

Lei Deng, Bingying Xu, Lumin Zhang, Yi Han, Bin Zhou, and Peng Zou. Tracking the evolution of public concerns in social media. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 353–357, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. Training temporal word embeddings with a compass. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 6326–6334, 2019.

Haim Dubossarsky, Eitan Grossman, and Daphna Weinshall. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, 2017. ISBN 9781945626838. 10.18653/v1/d17-1118.

Dimitrios Effrosynidis, Alexandros I. Karasakalidis, Georgios Sylaios, and Avi Arampatzis. The climate change Twitter dataset. *Expert Systems with Applications*, 204(October 2021):117541, oct 2022. ISSN 09574174. 10.1016/j.eswa.2022.117541. URL `https://linkinghub.elsevier.com/retrieve/pii/S0957417422008624`.

Heba Elfardy and Mona Diab. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 434–439, 2016.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Robert Englebretson. Stancetaking in discourse: An introduction. In *Rice Linguistics Symposium*. John Benjamins Publishing Company, 2007.

William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168. Association for Computational Linguistics, 2016. ISBN 9781941643914. 10.18653/v1/N16-1138. URL `http://aclweb.org/anthology/N16-1138`.

Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180, 2020.

Lea Frermann and Mirella Lapata. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016.

Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *ICWSM*, 2007.

David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. Ideological and temporal components of network polarization in online political participatory media. *Policy & internet*, 7(1):46–79, 2015.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. Every colour you are: Stance prediction and turnaround in controversial issues. In *12th ACM Conference on Web Science*, WebSci '20, page 174–183, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379892. 10.1145/3394231.3397907. URL `https://doi.org/10.1145/3394231.3397907`.

Pedro Guerra, Roberto Nalon, Renato Assunçao, and Wagner Meira Jr. Antagonism also

flows through retweets: The impact of out-of-context quotes in opinion polarization analysis. In *Proceedings of the International AAAI conference on web and social media*, pages 536–539, 2017.

Kristina Gulordava and Marco Baroni. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus, 2011. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.220.7680`.

Phong Ha, Shanshan Zhang, Nemanja Djuric, and Slobodan Vucetic. Improving word embeddings through iterative refinement of word-and character-level models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1204–1213, 2020.

Quang-Thuy Ha, Thi-Ngan Pham, Van-Quang Nguyen, Thi-Cham Nguyen, Thi-Hong Vuong, Minh-Tuoi Tran, and Tri-Thanh Nguyen. A new lifelong topic modeling method and its application to vietnamese text multi-label classification. In *Asian Conference on Intelligent Information and Database Systems*, pages 200–210. Springer, 2018.

Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415, 2019.

Mark A Hamilton. Language intensity as an expression of power in political messages. In *The exercise of power in communication*, pages 233–265. Springer, 2015.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, 2016a. ISBN 9781945626258. 10.18653/v1/d16-1057.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, 2016b. ISBN 9781945626258. 10.18653/v1/d16-1229.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, 2016c.

Harith Hamoodat, Firas Aswad, Eraldo Ribeiro, and Ronaldo Menezes. A longitudinal

analysis of vocabulary changes in social media. In Hugo Barbosa, Jesus Gomez-Gardenes, Bruno Gonçalves, Giuseppe Mangioni, Ronaldo Menezes, and Marcos Oliveira, editors, *Complex Networks XI*, pages 212–221, Cham, 2020. Springer International Publishing. ISBN 978-3-030-40943-2.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis-and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, 2022.

Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 751–762. Association for Computational Linguistics, 2014. ISBN 9781937284961. 10.3115/v1/d14-1083. URL `http://aclweb.org/anthology/D14-1083`.

Yu He, Jianxin Li, Yangqiu Song, Mutian He, and Hao Peng. Time-evolving text classification with deep neural networks. In *Proceedings of IJCAI, the International Joint Conference on Artificial Intelligence*, pages 2241–2247, 2018.

Tomáš Hercig, Peter Krejzl, and Pavel Král. Stance and Sentiment in Czech. *Computación y Sistemas*, 2018. accepted.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Valentin Hofmann, Hinrich Schütze, and Janet B. Pierrehumbert. The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(Icwsm):1259–1267, may 2022. ISSN 2334-0770. 10.1609/icwsm.v16i1.19377. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/19377`.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138881. 10.1145/1014052.1014073. URL `https://doi.org/10.1145/1014052.1014073`.

Manuela Hürlimann, Brian Davis, Keith Cortis, André Freitas, Siegfried Handschuh, and Sergio Fernández. A twitter sentiment gold standard for the brexit referendum. In *Proceedings of the 12th international conference on semantic systems*, pages 193–196, 2016.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. Emotion semantics show both cultural variation and universal structure. *Science (New York, N.Y.)*, 366(6472):1517–1522, 2019. ISSN 1095-9203. 10.1126/science.aaw8160. URL `http://www.ncbi.nlm.nih.gov/pubmed/31857485`.

Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

Kristen Johnson and Dan Goldwasser. Identifying Stance by Analyzing Political Discourse on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 66–75. Association for Computational Linguistics, 2016. 10.18653/v1/W16-5609. URL `http://aclweb.org/anthology/W16-5609`.

Christian Kahmann, Andreas Niekler, and Gerhard Heyer. Detecting and assessing contextual change in diachronic text documents using context volatility. In *IC3K 2017 - Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 1, pages 135–143, 2017. ISBN 9789897582714. 10.5220/0006574001350143. URL `https://git.informatik.uni-leipzig.de/mam10cip/KDIR.git`.

Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P Rana, Pushp Patil, Yogesh K Dwivedi, and Sridhar Nerur. Advances in social media research: Past, present and future. *Information Systems Frontiers*, 20(3):531–558, 2018.

Herbert C Kelman. Processes of opinion change. *Public opinion quarterly*, 25(1):57–78, 1961.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, 2020.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language pro-

cessing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, 2014.

Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Comput. Surv.*, 53(1): 12:1–12:37, 2020. 10.1145/3369026. URL `https://doi.org/10.1145/3369026`.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, 2015.

S Kumar. Weakly Supervised Stance Learning Using Social-Media Hashtags, 2018. URL `https://www.ml.cmu.edu/research/dap-papers/f18/dap-kumar-sumeet.pdf`.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, 2018.

Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. Friends and enemies of clinton and trump: using context for detecting stance in political tweets. In *Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23–28, 2016, Proceedings, Part I 15*, pages 155–168. Springer, 2017a.

Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Extracting graph topological information and users' opinion. In Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 112–118, Cham, 2017b. Springer International Publishing. ISBN 978-3-319-65813-1.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. Multilingual stance detection in social media political debates. *Computer Speech & Language*, page 101075, 2020.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 15–27. Springer, 2018.

John Lawrence and Chris Reed. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818, 2020.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363, 2021.

Jianping Li, Yimou Xu, and Huaye Shi. Bidirectional lstm with hierarchical attention for text classification. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 456–459, 2019. 10.1109/IAEAC47372.2019.8997969.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, 2019.

Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005.

Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li. TASC: Topic-adaptive sentiment classification on dynamic tweets. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1696–1709, 2015. ISSN 10414347. 10.1109/TKDE.2014.2382600.

Xiaohua Liu, Furu Wei, Shaodian Zhang, and Ming Zhou. Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology*, 4(1), 2013. ISSN 21576904. 10.1145/2414425.2414428. URL `http://doi.acm.org/10.1145/2414425.2414428`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland, May 2022. Association for Computational Linguistics. 10.18653/v1/2022.acl-demo.25. URL `https://aclanthology.org/2022.acl-demo.25`.

Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. Stance prediction for russian: data and analysis. In *International Conference in Software Engineering for Defence Applications*, pages 176–186. Springer, 2018.

Jan Lukes and Anders Søgaard. Sentiment analysis under temporal shift. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 65–71, 2018.

Nikolaos Lykousas, Costantinos Patsakis, Andreas Kaltenbrunner, and Vicenç Gómez. Sharing emotions at scale: The Vent dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(Icwsm):611–619, jan 2019. ISSN 2334-0770. 10.1609/icwsm.v13i01.3361. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/3361`.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. A demographic analysis of online sentiment during Hurricane Irene. In *Proceedings of the 2012 Workshop on Language in Social Media*, pages 27–36, 2012. URL `http://dev.twitter.com`.

Pablo Medina-Alias and Özgür Şimşek. The temporal persistence of generative language models in sentiment analysis. In *CEUR Workshop Proceedings*, volume 3497, pages 2458–2468. CEUR-WS, 2023.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Adam J Mills. Virality in social media: the spin framework. *Journal of public affairs*, 12 (2):162–169, 2012.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, 2016.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

Mitra Mohtarami, James Glass, and Preslav Nakov. Contrastive Language Adaptation for Cross-Lingual Stance Detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4433–4442. Association for Computational Linguistics, 2019. 10.18653/v1/D19-1452.

Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. Mitigation of diachronic bias in fake news detection dataset. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 182–188, 2021.

Le T Nguyen, Pang Wu, William Chan, Wei Peng, and Ying Zhang. Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12*, volume 12, pages 1–8. ACM Press, 2012. ISBN 9781450315432. 10.1145/2346676.2346682. URL `http://dl.acm.org/citation.cfm?doid=2346676.2346682`.

Thi-Cham Nguyen, Thi-Ngan Pham, Minh-Chau Nguyen, Tri-Thanh Nguyen, and Quang-Thuy Ha. A lifelong sentiment classification framework based on a close domain lifelong topic modeling method. In Ngoc Thanh Nguyen, Kietikul Jearanaitanakij, Ali Selamat, Bogdan Trawiński, and Suphamit Chittayasothorn, editors, *Intelligent Information and Database Systems*, pages 575–585, Cham, 2020. Springer International Publishing. ISBN 978-3-030-41964-6.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing (EMNLP-IJCNLP), pages 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018.

Dean Ninalga. Keeping in time: Adding temporal context to sentiment analysis models. *arXiv preprint arXiv:2309.13562*, 2023.

Kyosuke Nishida, Takahide Hoshide, and Ko Fujimura. Improving tweet stream classification by detecting changes in word probability. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 971–980, 2012.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010. ISBN 9781577354451. URL http://www.sca.isr.umich.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer, 2006.

Daniel Preoţiuc-Pietro and Trevor Cohn. A temporal model of text periodicities using gaussian processes. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 977–988, 2013.

Matthew Purver and Stuart Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491, 2012.

Irina V Pustokhina, Denis A Pustokhin, RH Aswathy, T Jayasankar, C Jeyalakshmi, Vicente García Díaz, and K Shankar. Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. *Information Processing & Management*, 58(6):102706, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ashwin Rajadesingan and Huan Liu. Identifying users with opposing opinions in twitter debates. In William G. Kennedy, Nitin Agarwal, and Shanchieh Jay Yang, editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 153–160, Cham, 2014. Springer International Publishing. ISBN 978-3-319-05579-4.

Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, 2020.

Neetu Rani, Prasenjit Das, and Amit Bharadwaj. Rumour detection in online social networks: Recent trends. *Available at SSRN 3564070*, 2020.

Amal Rekik, Salma Jamoussi, and Abdelmajid Ben Hamadou. Violent Vocabulary Extraction Methodology: Application to the Radicalism Detection on Social Media. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11684 LNAI, pages 97–109, 2019. ISBN 9783030283735. 10.1007/978-3-030-28374-2_9. URL `https://doi.org/10.1007/978-3-030-28374-2{_}9`.

Alexander Robertson, Walid Magdy, and Sharon Goldwater. Self-representation on twitter using emoji skin color modifiers. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.

Alexander Robertson, Farhana Ferdousi Liza, Dong Nguyen, Barbara McGillivray, Scott Hale, et al. Semantic journeys: quantifying change in emoji meaning from 2012-2018. In *Workshop Proceedings of the 15th International AAAI Conference on Web and Social Media*. AAAI Press, 2021.

Leonardo Rocha, Fernando Mourão, Adriano Pereira, Marcos André Gonçalves, and Wagner Meira. Exploiting temporal contexts in text classification. *International Conference on Information and Knowledge Management, Proceedings*, pages 243–252, 2008. 10.1145/1458082.1458117.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

Guy D Rosin, Eytan Adar, and Kira Radinsky. Learning word relatedness over time. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1168–1178, 2017.

Paul Röttger and Janet Pierrehumbert. Temporal adaptation of bert and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, 2021.

David Rozado, Ruth Hughes, and Jamin Halberstadt. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE*, 17(10):e0276367, oct 2022. ISSN 1932-6203. 10.1371/journal.pone.0276367. URL `http://dx.doi.org/10.1371/journal.pone.0276367`.

Sebastian Ruder. Word embeddings in 2017: Trends and future directions, 2017. URL `https://ruder.io/word-embeddings-2017/`.

Anna Rumshisky, Mikhail Gronas, Peter Potash, Mikhail Dubov, Alexey Romanov, Saurabh Kulshreshtha, and Alex Gribov. Combining network and language indicators for tracking conflict intensity. In *International Conference on Social Informatics*, pages 391–404. Springer, 2017.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111. Association for Computational Linguistics, 2009.

Giancarlo Salton and John Kelleher. Persistence pays off: Paying attention to what the lstm gating mechanism persists. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1052–1059, 2019.

Robin Schaefer and Manfred Stede. Improving implicit stance classification in tweets using word and sentence embeddings. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 299–307. Springer, 2019.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. A wind of change: Detecting and evaluating lexical semantic change across times and do-

mains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, 2019.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. 10.18653/v1/W17-5203. URL `https://www.aclweb.org/anthology/W17-5203`.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. Technical report, . Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University„ 1999.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, 2019.

Vasiliki Simaki, Carita Paradis, Maria Skeppstedt, ..., Magnus Sahlgren, Kostiantyn Kucher, and Andreas Kerren. Annotating Speaker Stance in Discourse: The Brexit Blog Corpus. *Corpus Linguistics and Linguistic Theory*, 2017. ISSN 16137035. 10.1515/cllt-2016-0060. URL `https://doi.org/10.1515/cllt-2016-0060`.

Vasiliki Simaki, Carita Paradis, and Andreas Kerren. Evaluating stance-annotated sentences from the Brexit Blog Corpus: A quantitative linguistic analysis. *ICAME Journal*, 42(1):133–166, 2018. ISSN 1502-5462. 10.1515/icame-2018-0007.

V Srinidhi Skanda, M Anand Kumar, and KP Soman. Detecting stance in kannada social media code-mixed text using sentence embedding. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 964–969. IEEE, 2017.

Kenny Smith. The evolution of vocabulary. *Journal of theoretical biology*, 228(1):127–142, 2004.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. Exploring deep neural networks for multitarget stance detection. *Computational Intelligence*, 35(1):82–97, 2019.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, pages 672–675, 2017. ISBN 9781577357889. URL `https://pymorphy2.readthedocs.io/en/latest/`.

Suhavi, Asmit Kumar Singh, Udit Arora, Somyadeep Shrivastava, Aryaveer Singh, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Twitter-STMHD: An Extensive User-Level Database of Multiple Mental Health Disorders. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(Icwsm):1182–1191, may 2022. ISSN 2334-0770. 10.1609/icwsm.v16i1.19368. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/19368`.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. Exploring various linguistic features for stance detection. In *Natural Language Understanding and Intelligent Applications*, pages 840–847, Cham, 2016. Springer International Publishing. ISBN 978-3-319-50496-4.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1), 2021.

Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661, 2015.

Mariona Taulé, Francisco Rangel, M. Antònia Martí, and Paolo Rosso. Overview of the task on multimodal stance detection in Tweets on catalan #1Oct referendum. In *CEUR Workshop Proceedings*, volume 2150, pages 149–166, 2018. URL `http://clic.ub.edu`.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods, 2010. ISSN 0261927X. URL `http://jls.sagepub.com`.

Kevin Tian, Teng Zhang, and James Zou. Cover: Learning covariate-specific vector repre-

sentations with tensor decompositions. In *International Conference on Machine Learning*, pages 4926–4935. PMLR, 2018.

Michele Tomaiuolo, Gianfranco Lombardo, Monica Mordonini, Stefano Cagnoni, and Agostino Poggi. A survey on troll detection. *Future Internet*, 12(2):31, Feb 2020. ISSN 1999-5903. 10.3390/fi12020031. URL `http://dx.doi.org/10.3390/fi12020031`.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, et al. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860, 2023.

Adam Tsakalidis and Maria Liakata. Autoencoding word representations through time for semantic change detection. *arXiv preprint arXiv:2004.13703*, 2020.

Adam Tsakalidis, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray. Mining the UK web archive for semantic change detection. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1212–1221, 2019.

Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010. ISSN 10769757. 10.1613/jair.2934. URL `http://www.natcorp.ox.ac.uk/`.

Jannis Vamvas and Rico Sennrich. X-stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th SwissText & 16th KONVENS Joint Conference 2020*, page 9. CEUR-WS. org, 2020.

Svitlana Volkova, Ilia Chetviorkin, Dustin Arendt, and Benjamin Van Durme. Contrasting public opinion dynamics and emotional response during crisis. In *Proceedings of the International Conference on Social Informatics*, pages 312–329. Springer, 2016.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across nlp tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, 2020.

Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant. Stance classification using dialogic properties of persuasion. In *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Proceedings of the Conference, pages 592–596, 2012a. ISBN 1937284204.

Marilyn A Walker, Pranav Anand, Jean E.Fox Tree, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 812–817, 2012b. ISBN 9782951740877. URL `http://nlds.soe.ucsc.edu/software`.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

Michael Wojatzki, Torsten Zesch, Saif Mohammad, and Svetlana Kiritchenko. Agree or Disagree: Predicting Judgments on Nuanced Assertions. In *Proceedings of *SEM*, pages 214–224, Stroudsburg, PA, USA, 2018. 10.18653/v1/S18-2026.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Brian Xu, Mitra Mohtarami, and James Glass. Adversarial domain adaptation for stance detection. *arXiv preprint arXiv:1902.02401*, 2019a.

Feng Xu, Zhenchun Pan, and Rui Xia. E-commerce product review sentiment classification based on a naïve bayes continuous learning framework. *Information Processing & Management*, 57(5):102221, 2020.

Kaiquan Xu, Jiexun Li, and Stephen Shaoyi Liao. Sentiment community detection in social networks. In *ACM International Conference Proceeding Series*, pages 804–805, 2011. ISBN 9781450301213. 10.1145/1940761.1940913.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, pages 907–916. Springer, 2016.

Zhenhui Xu, Qiang Li, Wei Chen, Yingbao Cui, Zhen Qiu, and Tengjiao Wang. Opinion-aware knowledge embedding for stance detection. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 337–348. Springer, 2019b.

Yilin Yan, Jonathan Chen, and Mei-Ling Shyu. Efficient Large-Scale Stance Detection in

Tweets. *International Journal of Multimedia Data Engineering and Management*, 9(3): 1–16, 2018. ISSN 1947-8534. 10.4018/ijmdem.2018070101.

Wenjie Yin, Rabab Alkhalifa, and Arkaitz Zubiaga. The emojification of sentiment on social media: Collection and analysis of a longitudinal twitter sentiment dataset. *arXiv preprint arXiv:2108.13898*, 2021.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.acl-main.291`.

Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Katsumi Tanaka. The past is not a foreign country: Detecting Semantically Similar Terms across Time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016. ISSN 10414347. 10.1109/TKDE.2016.2591008.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 247–256, 2016.

Arkaitz Zubiaga. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984, 2018.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989, 2016.