

Domain-specific and Context-aware Approaches to Sentiment Analysis

Shamsuddeen Hassan Muhammad

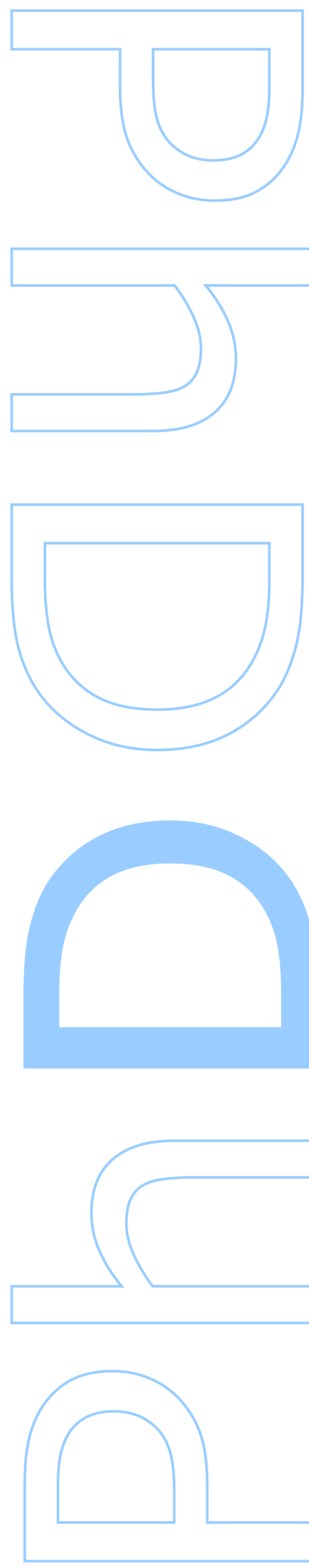
Doctoral Program in Computer Science - Mapi joint programme
Department of Computer Science
2023

Supervisor

Professor Pavel Bernard Brazdil ,
Faculty of Economics of University of Porto

Co-supervisor

Professor Alípio Mário Jorge,
Faculty of Sciences of the University of Porto

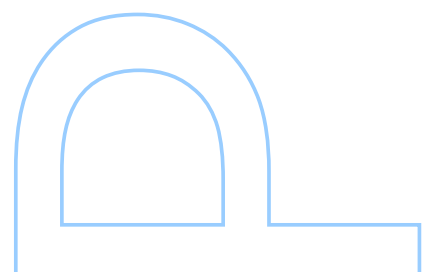
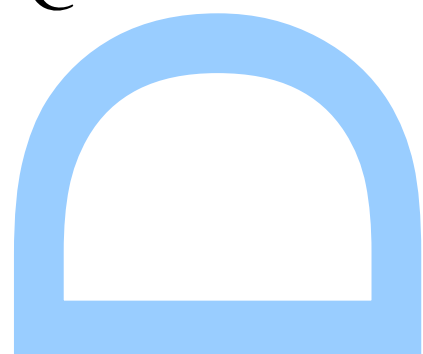




Universidade do
Minho

U. PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO



UNIVERSITY OF PORTO

DOCTORAL THESIS

Domain-specific and Context-aware Approaches to Sentiment Analysis

Author:

Shamsuddeen Hassan
MUHAMMAD

Supervisor:

Professor Pavel Bernard Brazdil

Co-supervisor:

Professor Alípio Mário Jorge

*A thesis submitted in fulfilment of the requirements
for the degree of PhD. Computer Science
at the*

Faculty of Science University of Porto
Department of Computer Science.

September 26, 2023

**Dedicated to
My parents**

Acknowledgments

The Ph.D. is a process of discovery, not only of new knowledge but also of oneself. I would like to thank all those who helped me discover my potential and pursue my passion.

Pursuing a Ph.D. involves facing many obstacles, unexpected events, doubts, and challenges. This thesis is the culmination of years of effort that involved the contribution of many people, in various ways. First and foremost, I am deeply grateful to my supervisors Pavel Bernard Brazdil and Alípio M. Jorge for their support, advice, and encouragement throughout my Ph.D. years. I appreciate the freedom they gave me to pursue my interests and curiosity.

I would like to thank my collaborators at MasaKhaneNLP and HausaNLP community for the collaborative work we did that contributed to this thesis. I am particularly grateful to Ibrahim Said Ahmad, Bello Shehu Bello, Idris Abudlmumin, David Adelani, Sebastian Ruder, Saif Mohammad, Nedjma Ousidhoum, and Seid Muhie Yimam. I would like to thank my friends in the Department of Computer Science (DCC) with whom I shared and discussed many ideas.

I would like to express my gratitude to the funding bodies that have supported my Ph.D. I acknowledge funding from the Tertiary Education Trust Fund (TETFund) and the Ph.D. scholarship I received from the Faculty of Computer Science at the University of Porto. I also acknowledge the financial support received from the project funded by the National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. Lastly, I acknowledge the funding our team received from Lacuna Fund to create a sentiment corpus for Nigerian languages, a hate speech corpus for African languages, and to conduct a shared-task on sentiment analysis for African languages.

Last but not least, I am extremely grateful to my parents who have shown me the value of education and patience, which have fueled my PhD journey and kept me going. I would

like to thank my wife for her patience and affection during this challenging journey. To my daughters Hannah and Ameenah, I love you both.

Abstract

Sentiment Analysis (SA) is a crucial field that enables inference of the general sentiment of texts, thereby facilitating decision-making processes. Various methods have been proposed, ranging from those that leverage sentiment lexicons, which can be further enhanced with contextual rules, to the more recent deep-learning techniques. In this thesis, we have explored these areas and introduced several enhancements.

Lexicon-based sentiment analysis relies on a sentiment lexicon. However, generating a sentiment lexicon for every domain manually is impractical. Therefore, we have investigated automatic methods for this task. The first one uses statistical/machine learning-like techniques to construct the lexicon based on the distribution of word ratings in training data. The second method uses a pre-trained and fine-tuned language model to assign sentiment values or class to words. These approaches have shown better performance compared with the existing approaches. We have also proposed a method for extending a domain-specific sentiment lexicon with a general-purpose one. Furthermore, it is well known that certain words sometimes referred to as shifters, can alter (shift) the intensity of the sentiment of other words. Therefore, we have investigated contextually aware methods that improve the performance of lexicon-based approaches.

We have also explored different approaches for fine-tuning pre-trained language models. While these approaches achieve high performance, their predictions are not accompanied by explanations of how they were derived. This limitation motivated us to investigate neuro-symbolic approaches combining symbolic and deep learning methods. Our findings demonstrate that a specific variant of this approach achieves reasonable performance while simultaneously maintaining the ability to explain how predictions were obtained.

Despite progress in sentiment analysis, African languages have received little attention due to the scarcity of freely available datasets. To overcome this limitation, we developed new corpora and established a benchmark for sentiment analysis in African languages. To encourage new research in sentiment analysis for these languages, we organized a SemEval shared-task on sentiment analysis for African languages that attracted several teams worldwide.

Resumo

A Análise de Sentimento (SA) é um campo crucial que permite inferir o sentimento geral dos textos, facilitando assim os processos de tomada de decisão. Vários métodos foram propostos, desde aqueles que aproveitam léxicos de sentimento, que podem ser aprimorados com regras contextuais, até as técnicas mais recentes de aprendizagem profunda. Nesta tese, exploramos essas áreas e introduzimos diversas melhorias.

A análise de sentimento baseada em léxico depende de um léxico de sentimento. No entanto, gerar manualmente um léxico de sentimento para cada domínio é impraticável. Portanto, investigamos métodos automáticos para esta tarefa. O primeiro usa técnicas estatísticas/de aprendizado de máquina para construir o léxico com base na distribuição de classificações de palavras em dados de treinamento. O segundo método usa um modelo de linguagem pré-treinado e ajustado para atribuir valores de sentimento ou classe às palavras. Essas abordagens mostraram melhor desempenho em comparação com as abordagens existentes. Também propusemos um método para estender um léxico de sentimento específico de domínio com um de uso geral. Além disso, é bem sabido que certas palavras, por vezes referidas como shifters, podem alterar (mudar) a intensidade do sentimento de outras palavras. Portanto, investigamos métodos contextualmente conscientes que melhoram o desempenho de abordagens baseadas em léxico.

Também exploramos diferentes abordagens para ajustar modelos de linguagem pré-treinados. Embora estas abordagens alcancem um elevado desempenho, as suas previsões não são acompanhadas de explicações sobre como foram derivadas. Essa limitação nos motivou a investigar abordagens neuro-simbólicas combinando métodos de aprendizagem simbólica e profunda. Nossas descobertas demonstram que uma variante específica desta abordagem atinge um desempenho razoável, ao mesmo tempo que mantém a capacidade de explicar como as previsões foram obtidas.

Apesar dos progressos na análise de sentimentos, as línguas africanas têm recebido pouca atenção devido à escassez de conjuntos de dados disponíveis gratuitamente. Para superar esta limitação, desenvolvemos novos corpora e estabelecemos uma referência para análise de sentimento em línguas africanas. Para encorajar novas pesquisas em análise de sentimento

para estas línguas, organizamos uma tarefa partilhada SemEval sobre análise de sentimento para línguas africanas que atraiu várias equipas em todo o mundo.

Contents

List of Tables	xvii
-----------------------	-------------

List of Figures	xxi
------------------------	------------

1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement and Research Questions	3
1.3 Research Contributions	4
1.4 Publications	4
1.5 Dissertation Structure	6
2 Literature Review	9
2.1 Introduction	9
2.2 Types of Sentiment Analysis Levels	10
2.3 Lexicon-Based Approaches to Sentiment Analysis	11
2.3.1 Sentiment Lexicons	12
2.3.2 Sentiment lexicon generation methods	14
2.3.3 Fully automatic approaches	14
2.3.4 Challenges of Lexicon-based SA	17
2.4 Machine learning approaches to sentiment analysis	17
2.5 Deep Learning Approaches to Sentiment Analysis	18

2.6	Hybrid Approaches to Sentiment Analysis	19
2.6.1	ML approaches in conjunction with sentiment lexicons	19
2.6.2	DL approaches in conjunction with sentiment lexicons: Neuro-symbolic approaches	19
2.7	Comparative Study of Lexicon-Based, ML, and DL Approaches	20
2.8	Sentiment Analysis for African Languages	21
3	Corpora	23
3.1	Film reviews (IMDB) dataset	23
3.2	EcoSentiPT: A Financial Sentiment Analysis Corpus	24
3.3	NaijaSenti: A Nigerian Twitter Sentiment Corpus	25
3.3.1	Data Collection and Cleaning	26
3.3.2	Dataset Annotation	28
3.3.3	Human Evaluation	31
3.3.4	NaijaSenti Statistics and Analysis	32
3.4	AfriSenti: Sentiment Corpus for African Languages	33
3.5	Summary of Corpora	36
3.6	Conclusions	37
4	Automatic Lexicon Generation	39
4.1	Methodology of Lexicon Creation	39
4.1.1	Manual generation of sentiment lexicon using NaijaSenti dataset	39
4.1.2	Automatic lexicon generation using a statistical/ML-like method	41
4.1.3	Sentiment lexicon generation using Bert-based language model	44
4.1.4	Combining domain-specific and general purpose lexicon	46
4.2	Using Lexicons for Sentiment Classification	47
4.2.1	Evaluation the Quality of Generated Predictions	47
4.3	Experiments and Results on Different Datasets	47

4.3.1	Experiments with the Amazon IMDB dataset	47
4.3.2	Experiments with the EcoSentiPT dataset	48
4.4	Conclusions	50
5	Improving Lexicon-based Approach with Contextual Shifters	51
5.1	Shifters	51
5.2	Shifter Pattern	51
5.3	Different types of shifter patterns	53
5.3.1	Intensification	53
5.3.2	Attenuation/downtoning	54
5.3.3	Reversal/inversion	55
5.4	Combining sentiment lexicons with shifter patterns	56
5.4.1	Explanations provided by the system	57
5.4.2	Dealing with shifters that affect multiple words in their scope	57
5.4.3	Processing phrases with multiple shifters	58
5.5	Experimental Setup and Results	59
5.5.1	Lexicons used	59
5.5.2	Experiments involving sentence fractions	60
5.5.3	Experiments involving sentences	60
5.6	Limitations and Future work	61
5.7	Conclusions	63
6	Deep Learning for Sentiment Analysis	65
6.1	Methodology	65
6.1.1	Pre-Trained Language Models	65
6.1.2	Fine-tuning Language Model	67
6.2	Experiments with EcoSentiPT Dataset	68

6.2.1	Experimental Setup	68
6.2.2	Experimental Results	69
6.3	Experiments with NaijaSenti Datase	70
6.3.1	Experimental Setup	70
6.4	Experiments with AfriSenti Dataset	73
6.4.1	Experimental Setup	74
6.4.2	Experimental Results	74
6.5	AfriSenti-SemEval	75
6.6	A Neuro-Symbolic Approach to Sentiment Analysis	76
6.6.1	Methodology	76
6.6.2	Experiments	77
6.7	Conclusions and Future Work	78
7	Conclusions and Future Work	81
7.1	Contributions and Findings	81
7.2	Future work	83
	References	85

List of Tables

2.1	Calculating sentiment value for "bad" using Almatarneh's Approach	16
3.1	Examples of positive and negative sentences in IMDB movie review dataset .	24
3.2	Examples of some sentences and ratings in EcoSentiPT dataset	25
3.3	Summary of datasets used in six existing datasets on sentiment analysis in four major Nigerian languages in comparison to ours. The symbol "*" indicates that the provided URL is no longer accessible. Code-mixed means that the data set contains mixed language content.	26
3.4	3-class and 5-class annotation and inter-Annotator agreement. hau= hausa, ibo = Igbo, yor = Yoruba and pcm = Nigerian-Pidgin	30
3.5	Fleiss kappa agreement among each class	30
3.6	Human performance result using micro-F1 and Matthew's correlation coefficient.	32
3.7	Key stats of NaijaSenti: #mono-lingual tweets, #code-mix tweets, #token . .	32
3.9	Inter-annotator agreement scores using the free marginal multi-rater kappa [135] for the different languages.	34
3.10	Sizes and splits of the AfriSenti datasets. We do not allocate training splits to Oromo (orm) and Tigrinya (tir) due to the limited size of the data and only evaluate them in a zero-shot transfer setting	35
3.11	Summary of the four datasets used in our work. The symbol * indicates an existing corpus.	37
4.1	Examples of a manually generated sentiment lexicon for three Nigerian languages	40
4.2	Sample of lexicon entries generated using statistical-based approach	43
4.3	Distribution of four different lexical classes in EcoLex	43

4.4	Sample of Lexicons generated using statistical-based approach	44
4.5	Examples of some lexicon entries in DeepEcoLex and EcoLex	45
4.6	DeepNaijaLex: Lexicon and their translation created using a pre-trained language model from the NaijaSenti Dataset.	46
4.7	Examples of combining domain-specific lexicon (EcoLex) and general purpose lexicon (Sentilex').	46
4.8	Experiments using IMDBLex and AlmatarLex.	48
4.9	Numbers of elements in the train and test sets.	48
4.10	Lexicon size used in each setup	49
4.11	Weighted F1-score results on sentence fractions	49
4.12	Weighted F1-score results on Long sentence	50
5.1	Dependencies between tokens provided by the dependency parser	52
5.2	Some shifter patterns that include intensification	54
5.3	Some shifter patterns that include attenuation/downtoning	55
5.4	Some shifter patterns that include reversal/inversion	56
5.5	Weighted F1-score results on sentence fractions	60
5.6	Weighted F1-score results on long sentences	60
6.1	Weighted F1-score results on sentence fractions	69
6.2	Weighted F1-score results on long sentences	69
6.3	Weighted F1 evaluation of different Multilingual PLMs. Average and standard deviation over 5 runs. Numbers with “*” are within the standard deviation of the best model.	71
6.4	Weighted F1 evaluation of different Models using LAPT. The LAPT is trained on either general corpus or Twitter.	71
6.5	Weighted F1 evaluation of different Models using Multi-task.	72
6.6	Transfer learning experiments. PLMs are trained on English SemEval 2017 and evaluated on Nigerian languages in a zero-shot setting	72

6.7	Accuracy scores of monolingual baselines for AfriSenti on the 12 languages with training splits. Results are averaged over 5 runs.	74
6.8	Weighted F1-score results on sentence fractions	77
6.9	Weighted F1-score results on long sentences	78

List of Figures

1.1	Google Trends for sentiment analysis search hits (2004-2023)	2
2.1	Areas of research in sentiment analysis	10
3.1	Star ratings distribution in the IMDB movie review corpus.	24
3.2	Distribution of ratings in EcoSentiPT dataset	25
3.3	Inter-annotators progress over thirty batches—one-thousand tweets per batch.	31
3.4	Language Family in the AfriSenti datasets.	35
3.5	Countries and languages represented in the AfriSenti data collection (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá).	36
5.1	Two examples with dependency links between some tokens	52
5.2	Example of tokens linked by dependencies	58
6.1	Classical transfer learning with pre-training and fine-tuning language model	67
6.2	Adaptive or continual pre-training and fine-tuning of language models for improved performance on new data.	68
6.3	Sample efficiency on <code>hau</code> , <code>ibo</code> , and <code>yor</code> using the AfriBERTa model.	73
6.4	Architecture of the proposed neuro-symbolic approach to sentiment analysis	77

List of Algorithms

1	Automatic Lexicon Generation [12]15
---	-------------------------------------

Chapter 1

Introduction

"Bad impressions and bad stereotypes are quicker to form and more resistant to disconfirmation than good ones."

Baumeister

This chapter introduces my Ph.D. dissertation titled "Domain-specific and Context-aware Approaches to Sentiment Analysis." It presents the motivation, problem statement, and main contributions of the research. It also explains how the thesis is organized and structured.

1.1 Motivation

The exponential growth of social media and dynamic websites has allowed users to express their opinions on various topics such as politics and business on the Internet. This digital revolution and paradigm shift have enabled users to share their sentiments online, both positive and negative. However, there is a tendency for people to focus more on negative information or experiences than on positive ones. [17]. In psychology, this phenomenon is known as

negative bias [140], which explains how humans tend to pay more attention to negative information. This negative information can have a detrimental impact on the reputation and performance of businesses and other entities. For example, a single negative review or comment can dissuade potential customers from buying a product or service or influence the public perception of a policy or a leader. Therefore, it is crucial that an entity monitors and analyzes the feelings of its customers, stakeholders, and audiences and responds appropriately to their feedback and concerns [149, 90, 45, 83].

Traditionally, companies conduct surveys and opinion polls to find user opinions about their products or services to improve services and avoid losing customers. Similarly, governments use surveys to find the reaction and acceptance of people toward new or existing policies. However, with the rapid increase in user-generated text online, manually analysing large amounts of text would be ineffective. As a result, automated approaches are required to extract sentiment from text, effectively facilitating understanding of user sentiment or opinion [149, 90, 45, 83].

Sentiment analysis (SA) is a subfield of natural language processing that aims to computationally determine sentiment, opinion, emotion, and attitude from written text [83]. According to Pang et al. [123], the interest in sentiment analysis was initiated around 2001 and identified three factors that motivated interest in this field: the rise of machine learning methods in natural language processing and information retrieval; the availability of datasets on which machine learning algorithms to be trained on; and the realization of both fascinating intellectual challenges and practical applications. Today, sentiment analysis is commonplace and widely used in many domains and applications such as recommender systems, user reviews, and politics [149, 90, 45]. The Google Trends data for “sentiment analysis” search hits, shown in Figure 1.1, demonstrates a growing interest in the field over time. However, such progress has concentrated on a few high-resource languages such as English. For example, African languages have received little attention in sentiment analysis research due to the lack of sentiment analysis resources such as corpora and lexicons.

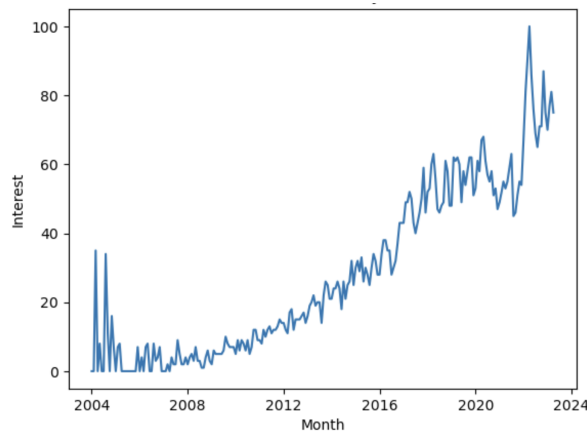


Figure 1.1: Google Trends for Sentiment Analysis Search Hits (2004-2023). A value of 100 is the peak popularity and 0 means that there was not enough data for the SA search.)¹

Approaches to sentiment analysis have evolved from the early days of lexicon-based methods [165, 152, 98] to more advanced machine learning [5, 81], deep learning-based methods [181, 178], and hybrid approaches that combine lexicon and machine learning-based approaches [52, 72]. Today, pre-trained language models (PLMs) provide state-of-the-art performance for sentiment classification.

¹[Trends.google.com:Sentiment Analysis Search Trends \(2004-2019\)](https://trends.google.com/trends/explore?q=sentiment+analysis&hl=en)

This dissertation focuses on improving lexicon-based sentiment analysis, also known as symbolic approaches. These approaches leverage sentiment lexicons and adhere to the principle of explainable AI. They provide explainable sentiment predictions that are useful in different domains, such as the medical domain [182]. In contrast to black-box deep learning-based approaches, we also explored a neuro-symbolic approach that combines the advantages of symbolic (explanation capabilities) and deep learning techniques (performance). In the next section, we define the specific problems addressed in our research.

1.2 Problem Statement and Research Questions

Lexicon-based sentiment analysis approaches depend on sentiment lexicons that contain words and phrases conveying positive or negative sentiment. These lexicons are domain-dependent—words can have different connotations across domains. For example, the word *high* can have a positive connotation in economics (e.g., *He has a high salary*) and a negative connotation in medicine (e.g., *He has high blood pressure*).

To improve performance in lexicon-based sentiment analysis, domain-specific lexicons are used. However, manually generating domain-specific sentiment lexicons for each domain can be tedious and time-consuming. Therefore, an approach is needed to generate domain-specific sentiment lexicons.

Existing approaches for sentiment lexicon generation focus on creating a lexicon from a single corpus and do not support automatic extension of the lexicon. However, there are several reasons to extend an existing lexicon, including an insufficient number of entries in the existing domain-specific lexicon, the need to update an existing lexicon to another domain (e.g., general-purpose to domain-specific), and language evolution (new words and meaning changes). Additionally, the classical lexicon-based sentiment analysis approach relies on unigrams, and it ignores the effect of sentiment shifters—words or expressions that can modify the polarity of another word or expression through intensification, downtoning/attenuation or reversal/inversion.

Africa has more than 2,000 languages from six language families and has the highest linguistic diversity among all continents. This includes 75 languages with at least one million speakers each. However, Africa has seen little research on sentiment analysis for its languages due to the lack of high-quality annotated datasets for low-resource African languages. Therefore, high-quality resources for sentiment analysis in African languages, including sentiment corpora, lexicons, and models are needed to bridge this gap.

In summary, this thesis aims to answer the following four research questions:

1. Can we improve the existing approaches for automatic lexicon generation?

2. Can we extend lexicon-based methods to capture some contextual information?
3. Can we combine deep learning and lexicon-based approaches in a neuro-symbolic model to achieve high performance and interpretability?
4. Can we advance sentiment analysis research for under-resourced African languages?

1.3 Research Contributions

To address the research questions outlined above, the main contributions of this dissertation are as follows:

1. Development of novel methods to automatically create sentiment lexicons from a labelled corpus [22].
2. Development of a novel method that extends an existing sentiment lexicon [21].
3. Improvement of lexicon-based approaches by incorporating contextual information, such as negation and intensification [22].
4. Proposal of a new neuro-symbolic approach for sentiment analysis that combines deep learning and lexicon-based methods [160].
5. Advancing sentiment analysis for under-resourced African languages by developing novel methods and resources that address the challenges of data scarcity and a comprehensive benchmark [161, 159, 154]. To promote research in these languages, I also co-organized a SemEval shared-task on sentiment analysis for African languages that attracted several teams worldwide.

1.4 Publications

Publications directly related to the thesis are:

1. **Muhammad, Shamsuddeen Hassan.** An overview of sentiment analysis approaches. In *MAP-i Seminar Proceedings*, pages 65–70, 2019 [155]
2. **Muhammad, Shamsuddeen Hassan,** Pavel Brazdil, and Alípio Jorge. Incremental approach for automatic generation of domain-specific sentiment lexicon. In *Proceedings of Advances in Information Retrieval, LNCS*, volume 12036, pages 619–623. Springer, 2020 [156]

3. Pavel Brazdil and **Shamsuddeen, Muhammad** et.al. Semi-automatic approaches for exploiting shifter patterns in domain-specific sentiment analysis. *Mathematics*, 10(18), 2022 [22]
4. Pavel Brazdil, Purificação Silvano, Fátima Silva, **Muhammad, Shamsuddeen**, Fátima Oliveira, João Cordeiro, and António Leal. Extending general sentiment lexicon to specific domains in (semi-)automatic manner. In Sara Carvalho and Renato R. Souza, editors, *SALLD-1: Proceedings of the Workshop on Sentiment Analysis & Linguistic Linked Data*. CEUR, Vol. 3064, 2021 [21]
5. **Shamsuddeen Hassan Muhammad**, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. Naijasenti: A Nigerian twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference*, pages 590–602, Marseille, France, June 2022. European Language Resources Association [161]
6. **Muhammad, Shamsuddeen Hassan** et al. AfriSenti: a Twitter sentiment analysis benchmark for African languages. *arXiv preprint arXiv:2302.08956*, 2023 [159]
7. **Muhammad, Shamsuddeen** et al. SemEval 2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada, July 2023 [154]
8. **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio Jorge. Combining symbolic and deep learning approaches for sentiment analysis. In Pascal Hitzler, Md Kamruzzaman Sarker, and Audrey Eberhart, editors, *Frontiers in Artificial Intelligence and Applications*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 506–521. IOS Press, August 2023 [157]
9. **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio M. Jorge. Leveraging symbolic and deep learning techniques for explainable sentiment analysis. In *22nd EPIA Conference on Artificial Intelligence*. EPIA, 2023 [158]

Conference paper accepted and awaiting publication:

1. **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio M. Jorge. Leveraging symbolic and deep learning techniques for explainable sentiment analysis. In *22nd EPIA Conference on Artificial Intelligence*. EPIA, 2023 [158]

Other collaborative works conducted during the PhD work but not directly related to the thesis work are:

1. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022 [77]
2. A few thousand translations go a long way! leveraging pre-trained models for African news translation. *In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics [4]
3. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022 [143]
4. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France, June 2022. European Language Resources Association [2]
5. Masakhaner: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021 [65]
6. Participatory research for low-resourced machine translation: A case study in African languages. *In Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics [109]
7. Bibletts: a large, high-fidelity, multilingual, and uniquely African speech corpus. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2383–2387. ISCA, 2022 [93]
8. Adapting to the low-resource double-bind: Investigating low-compute methods on low-resource African languages. *arXiv preprint arXiv:2303.16985*, 2022 [82]
9. A survey on machine learning techniques in movie revenue prediction. *SN Computer Science*, 1:1–14, 2020 [7]
10. Deep sequence models for text classification tasks. *In 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6. IEEE, 2021 [1]

1.5 Dissertation Structure

The thesis is organized into seven chapters.

Chapter 1: provides an introduction to the thesis, the problem, and the overall contribution of the thesis.

Chapter 2: presents a literature review on sentiment analysis, including the generation of domain-specific sentiment lexicons and current research directions and gaps in the automatic generation of domain-specific lexicons. It also discusses related literature on deep learning for sentiment analysis.

Chapter 3: describes the corpora used in this work and our contributions to creating new corpora for sentiment analysis in low-resource African languages

Chapter 4: introduces two new approaches for automatically generating domain-specific sentiment lexicons and presents an experimental study using the generated lexicons. This chapter also discusses the manual construction of lexicons for three low-resource Nigerian languages.

Chapter 5: presents an improved lexicon-based method that leverages a sentiment lexicon and a manually developed set of shifter patterns and rules.

Chapter 6: discusses deep learning methods for sentiment analysis and neuro-symbolic approaches.

Chapter 7: concludes the thesis, summarizes the work presented, and discusses possible future research directions.

Chapter 2

Literature Review

"You don't know where you're going
until you know where you've
been.—James Baldwin."

James Baldwin

This section presents the background and literature on sentiment analysis. We discuss sentiment lexicons, approaches for generating lexicons, and approaches for sentiment analysis. Finally, we present some comparative studies on approaches to sentiment analysis.

2.1 Introduction

Sentiment analysis is a prominent subfield of Natural Language Processing that aims to automatically identify sentiments or opinions expressed in various forms of online content, including posts on social networks, blogs, and product reviews [84, 97]. This area of research has numerous applications, including literary analysis, commercial use (e.g., tracking opinions toward products), and research in psychology and social science [136, 55, 39]. According to Pang et al. [124], sentiment analysis research began around 2001 and has since rapidly evolved. As a result, various research areas on sentiment analysis have emerged over the years, as shown in Fig. 2.1.

Early work in sentiment analysis started with lexicon-based approaches [165, 152, 98] and continued to more advanced machine learning [5, 81], deep learning-based approaches [181, 178], and hybrid approaches that combine lexicon and machine learning-based approaches [52, 72]. Today, pre-trained language models (PLMs) provide state-of-the-art performance for sentiment classification in different languages [36]. However, these approaches rely on deep learning methods, which offer better performance, but are often considered as *black boxes*

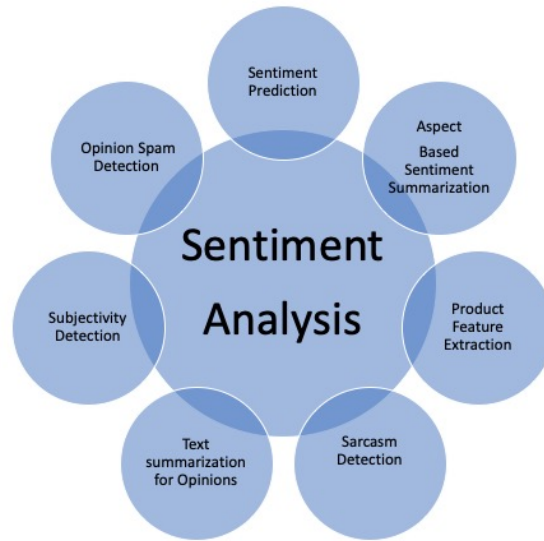


Figure 2.1: Areas of research in sentiment analysis

because they do not provide clear explanations for their predictions [178]. As users often want to understand why a specific polarity was attributed to a text in sentiment analysis tasks, developing solutions that provide such interpretability is useful. Therefore, recent work on sentiment analysis focused on explainable sentiment analysis [27].

The remainder of this chapter is structured as follows: Section 2.2 discusses the levels of sentiment analysis. Subsequent sections explore various approaches to sentiment analysis, including lexicon-based approaches in Section 2.3, machine learning-based approaches in Section 2.4, and deep learning-based approaches in Section 2.5. Section 2.6 examines hybrid approaches, while Section 2.6 presents comparative studies. Finally, Section 2.8 discusses sentiment analysis for low-resource African languages.

2.2 Types of Sentiment Analysis Levels

Sentiment analysis can be performed at different levels of granularity, depending on the scope and goal of the task. This section reviews three levels: sentence, document, and aspect-based sentiment analysis.

Sentence level sentiment analysis Sentence-level sentiment analysis aims to identify the polarity of individual sentences within a document as positive, negative, or neutral. This process is typically performed after *subjectivity classification*—which distinguishes between objective and subjective sentences. Subsequently, the sentiment orientation of subjective sentences is determined [85]. However, some objective sentences may convey opinions [85, 78], indicating that sentiment analysis can be applied to subjective and objective sentences.

Sentence-level sentiment analysis may not always yield optimal results, particularly in the case of complex and compound sentences that express multiple opinions. As such, previous studies have often assumed that each sentence expresses a single opinion [85].

Document level sentiment analysis Document-level sentiment analysis involves summarizing the entire content of a document into a single opinion [122, 124]. This approach assumes that each document expresses an opinion about a single entity and may not be suitable for documents containing sentiments about multiple entities [85].

Aspect level sentiment analysis Aspect-level sentiment analysis is more complex than sentence-level and document-level sentiment analysis [85]. This approach aims to identify sentiments about specific entities and their aspects within a document or sentence. For instance, the sentence *The new Tesla design is superb, but its battery is terrible* evaluates two aspects of the Tesla car: its design and battery life. The sentiment towards the Tesla design is positive, while the sentiment toward its battery is negative. This level of sentiment analysis recognizes that opinions without targets are meaningless. In the example above, the opinion targets are the *Tesla design* and *battery life* [85].

In the following sections, we discuss different approaches for sentiment analysis, including lexicon-based, machine-learning, and deep-learning methods. We also discuss a hybrid approach that combines the advantages of both lexicon-based and machine-learning methods [149, 13].

2.3 Lexicon-Based Approaches to Sentiment Analysis

The lexicon-based approach to sentiment analysis is based on using a sentiment lexicon. It involves identifying sentiment-carrying words in the text and using a sentiment lexicon to assign a sentiment value to each word. These values are then summed to obtain the overall sentiment score of the text. This method is particularly useful for analyzing sentiment when computational resources are limited. For example, consider the following text: *"I love an iPhone. It is an excellent phone."* To determine the sentiment of this text using a lexicon-based approach, we search for and retrieve the sentiment values of each word in the text from a sentiment lexicon. In this case, let us assume that our lexicon contains sentiment values for only two words: *love* (score of 2) and *excellent* (score of 3). By summing the sentiment scores for each word in the text, we can calculate the overall sentiment score, which is 5 (2 + 3). If the aim were to determine the polarity, the output would be a categorical value (e.g., "1" representing positive polarity). Variations of this approach exist based on what value is assigned to sentiment words in the sentiment lexicon and how negation and context are handled [25].

2.3.1 Sentiment Lexicons

A sentiment lexicon is a lexical resource or dictionary that consists of lexical items with corresponding semantic orientations. The lexical items convey a single meaning and can include words (e.g., *good*, *bad*, *etc.*), phrases (e.g., *I am over the moon*), and idiomatic expressions (e.g., *over the moon*). Semantic orientation can take many forms, such as categorical values (positive, negative, or neutral), ordinal values (strongly positive, mildly positive, and strongly negative), and values on a numeric scale indicating sentiment strength from highly negative to highly positive (e.g., -3 to +3) [6, 152]. Sentiment lexicons can be general-purpose or domain-specific [104, 54, 11, 76, 125]. Sentiment lexicons are mostly adjectives and adverbs. However, Chesley et al. [33] and Subrahmanian et al. [151] have shown that some classes of verbs can indicate sentiment. For example, *appraising verbs* usually indicate positive orientation, while *doubting verbs* indicate negative orientation. Therefore, we can identify sentiment in sentences like *I wholeheartedly **support** the president*. In chapter 3, we create a sentiment lexicon that contains four classes: *adjective*, *adverb*, *verb*, and *noun*, which improve the performance of the lexicon. Sentiment lexicons can be classified into two types: general-purpose and domain-specific.

General-purpose lexicons are developed for sentiment analysis in various domains. They are nonspecific and can be used to identify opinions in various domains such as movies, social media, health, and government. The following are examples of general-purpose lexicons:

WordNet is an online English general lexical resource database. It groups adjectives, nouns, and verbs into synonyms set through semantic relations [94].

SentiWordNet is a lexical resource with a high level of coverage developed by Esuli and Sebastiani [14]. It uses three sentiment orientations (positive, negative, and neutral) developed from WordNet. In SentiWordNet, words may have different meanings and, therefore, different polarities. For example, the word *cold* may mean having a low temperature, as in *cold beer*, or lacking human warmth or emotion, as in *cold person*. SentiWordNet uses glosses for each word entry to distinguish one meaning from another [14].

SenticNet is a publicly available lexical resource for concept-level sentiment analysis. The lexicon includes both semantic and affective lexical units and provides over 30,000 multi-word expressions to enable a fine-grained analysis of natural language opinions. SenticNet uses sentiment orientation between -1 and +1 (-1 being extremely negative and +1 extremely positive) [28].

General Inquirer is a general lexical system developed at Harvard for content analysis research in behavioural sciences. The system uses two dictionaries: a psycho-sociological dictionary and an anthropological dictionary used for studying themes in the folktales of

many traditions and cultures. During sentence analysis, the general inquirer looks up these dictionaries to find the category to which a word belongs if it exists [148].

Bing Liu’s Opinion Lexicon is a freely available sentiment lexicon developed by Bing Liu. It consists of an English opinion lexicon that is continuously being developed. The lexicon contains a list of positive and negative words totalling approximately 6,800 [63].

One advantage of general-purpose lexicons is their comprehensive coverage. However, their precision may suffer because sentiment words are often domain dependent [131, 104, 11, 45, 125]. A word may have a positive polarity in one domain and a negative polarity in another. Thus, the semantic orientation of a word is defined by the domain and context in which it is used. For example, the word *suck* can have negative and positive orientations depending on the context. In the sentence *The camera we bought sucks*, it has a negative orientation. In contrast, in the sentence *the vacuum cleaner we bought sucks very well* it has a positive orientation. In addition to having different meanings across domains, sentiment-bearing words can have different meanings within the same domain. This adds to the complexity of sentiment analysis. For example, in the camera domain, the word *long* appears twice in the sentence *The battery life of the camera is long, but it takes a long time to focus*, with different orientations. The first instance indicates a positive opinion, while the latter indicates a negative one [148]. Therefore, more recent attention has focused on generating a domain-specific sentiment lexicon.

Domain-specific lexicons are developed for specific domains and carry sentiment information that depends on the domain. Using domain-specific lexicons can improve the accuracy in sentiment analysis tasks. However, manually curating domain-specific sentiment lexicons for each domain is difficult, so they are often automatically generated from domain-specific corpora or extended from general-purpose lexicons. However, creating a high-quality domain-specific lexicon requires large-scale corpora, often unavailable for each domain. One solution to this challenge is to generate domain-specific lexicons and use them to extend general-purpose lexicons. As a result, recent research has focused on developing automatic methods for generating domain-specific lexicons [11, 71, 89, 54, 80, 67]. The following are examples of general-purpose lexicons.

SPLM is a domain-focused sentiment lexicon generated from the IMDB movie reviews corpus by Almatarneh et al. [11]. It contains only *adjective* and *adverb* and a total of 532 negative words and 612 positive words. The lexicon showed higher performance compared to manually generated lexicons (SO-CAL [152] and SentiWords [?]).

SocialSent is a sentiment lexicon for computational social science research. The authors induced the lexicon from 250 Reddit forums to show that the sentiment lexicon varies between communities [54].

2.3.2 Sentiment lexicon generation methods

Sentiment lexicons can be generated using manual, automatic, or semi-automatic approaches. Semi-automatic approaches combine elements of both manual and automatic generation to create sentiment lexicons.

The manual approach involves using an existing dictionary or corpus to manually select lexical items with sentiment orientation (e.g., good and bad) and annotating them with sentiment strength. For example, a 5-class sentiment strength scale (-2 to +2), -2 and -1 can be considered *negative sentiment*, 0 as *neutral sentiment*, and +1 and +2 as *positive sentiment*. Because humans annotate each lexical item, this method achieves good performance in sentiment analysis. However, it is time-consuming and challenging due to the inherent nature of the manual activity involved. It also has limited coverage compared to automatically generated lexicons [6]. Moreno-Ortiz et al. [100] developed a domain-specific lexicon for English called SentiEcon. This lexicon was manually annotated with semantic orientation and intensity from the financial/economic domain.

The manual process can also be used to extend the lexicon. Forte et al. [45] manually extended an existing general-purpose sentiment lexicon with certain domain-specific terms, which led to a marked improvement in predictive performance. Similarly, Silva et al. [147] used a manual approach and extended an existing lexicon to the economic and financial domain. The disadvantage of manual approaches is that domain-specific extensions require considerable manual effort, which has motivated others to investigate automatic approaches to lexicon generation.

The automatic approach to the generation of sentiment lexicons uses computational methods to overcome the time and difficulty associated with manual approaches. This approach can involve extending an existing set of sentiment-bearing terms using automatic methods or constructing a new lexicon from scratch using a corpus. This thesis presents an enhanced, fully automatic approach to generate sentiment lexicons. In the following section, we review related studies on automatic lexicon generation.

2.3.3 Fully automatic approaches

Almatarneh et al. [11] proposed an approach to automatically generate domain-specific sentiment lexicons from the IMDB corpus [129]. The corpus is labeled with star ratings from the most negative to the most positive (+1 to +10). Their approach identifies sentiment-bearing words (adjectives and adverbs) from the corpus and automatically classifies them as either negative or positive. The method first categorizes the star ratings of the corpus (+1 to +10) into three groups: negative (+1 to +4), neutral (+5 to +6), and positive (+7 to +10). The corpus is then tagged with part-of-speech tags, and adjectives and adverbs are

Algorithm 1 Automatic Lexicon Generation [12]

```

input : Annotated corpus with star ratings
output: A sentiment lexicon with Positive and Negative Polarity

1  Set a category set  $c$  based on the ratings;
2  POS-tagged the corpus ;
3  Adjectives and adverbs are selected and stemmed ;
4  for every word  $w$  in each category  $c$  do
5      calculate its Average Relative Frequency using  $Total_c$  (total number of
        word tokens in each Category);
6       $RF_c(w) \leftarrow \frac{freq(w,c)}{Total_c}$ ;
7      calculate its Average Positive Scores;
8       $A_p(w) \leftarrow \frac{\sum_{c=1}^p RF_c(w)}{p}$ ;
9      calculate its Average Negative Scores;
10      $A_n(w) \leftarrow \frac{\sum_{c=(p+n+1)}^q RF_c(w)}{p}$ ;
11     assign each  $w$  to either positive or negative category;
12      $D(w) \leftarrow A_p(w) - A_n(w)$ ;
13     if  $D(w) < 0$  then
14         |  $w$  is assign to negative category;
15     else
16         |  $w$  is assign to positive category;
17 end

```

selected and stemmed. Finally, the stemmed words are classified as positive or negative (see Algorithm 1).

To demonstrate this approach using an example of the word "*bad*", the authors calculated the average of the relative frequencies for positive ratings (7,8,9,10) and negative ratings (1,2,3,4) of this word. as shown in Table 2.1. The average for the positive category is 0.0009 and for the negative category is 0.0033. Since the difference between the two values is negative ($0.0009 - 0.0033 = -0.0024$), the sentiment value of this word is negative.

Almatarneh et al. [11] uses only *adjective* and *adverbs* as sentiment-bearing words from the corpus. However, as pointed out earlier, some verbs can indicate sentiment [33, 151]. Therefore, in Chapter 3, we proposed an extended approach that includes more lexical classes (verbs and nouns), which marked improved performance.

Guang et al. [132] proposed a double propagation approach to generate domain-specific lexicons. The technique uses a seed word to extract sentiment-bearing words and features. These extracted lexical items are then used iteratively to find new sentiment words and features until all sentiment-bearing words are exhausted. Similarly, Kanayama et al. proposed an approach for generating domain-specific lexicons.

Kanayama et al. [71] developed a novel approach that uses the concept of context coherency to construct a domain-focused lexicon for sentiment analysis automatically. Context coherency

N.Stars	Freq.	$Total_c$	R.Freq.	Ave. of Neg and Pos.	polarity of bad
1	122232	25395214	0.005		
2	40491	11755132	0.003		
3	37787	13995838	0.003	0.0033	
4	33070	14963866	0.002		
5	39205	20390515	0.002		
6	43101	27420036	0.002		-0.0024
7	46696	40192077	0.001		
8	42228	48723444	0.001		
9	29588	40277743	0.001	0.0009	
10	51778	73948447	0.001		

Table 2.1: Calculating sentiment value for "bad" using Almatarneh's Approach

refers to the phenomenon where words with the same polarity tend to appear adjacent within a context. This approach has a reported accuracy of 94% and has been shown to be effective and easily adaptable to different domains.

Jijkoun et al. [67] developed a bootstrapping approach that automatically generates topic-focused lexicons from general-purpose lexicons. A recent study by Hang et al. [56] introduced a new method for generating domain-specific lexicons from unlabelled review data. This approach is divided into two parts. The first task involves labeling training reviews with polar values (negative and positive). Lexical units with higher ranking scores are selected and used as training data. The second task uses the selected training data to obtain a new domain-focused sentiment lexicon. Their approach has been shown to offer better performance compared to other domain-specific lexicons.

Using embeddings for lexicon generation With the recent advancements in deep learning and neural network applications in natural language processing, techniques such as word embeddings are used to induce domain-specific lexicons [76]. Hamilton et al. [54] combined domain-specific word embedding with label propagation to automatically generate domain-dependent sentiment lexicons using a small seed of words. Their framework, called SENTPROP, learns sentiment lexicons from seed words and domain-specific corpora and has been shown to give state-of-the-art performance competitive with manual approaches. This approach has the advantage of being able to work with even small-sized domain-specific corpora (107 tokens). In Chapter 3, we use a pre-trained language model to create sentiment lexicons automatically.

2.3.4 Challenges of Lexicon-based SA

There are several general challenges that are difficult to solve with lexicon-based sentiment analysis. One challenge is determining implicit sentiment. For example, in the sentence *The food is too expensive*, the concept of price is implicitly assumed but not explicitly stated. Another challenge is handling counterfactual expressions such as conditional sentences and questions, which may contain sentiment words but do not express sentiment. For example, the sentences *Can you tell me which camera is good?* and *If I can find a good camera in the store, I will buy it* do not express any opinion or sentiment, even though they contain sentiment words. On the contrary, some sentences do not contain sentiment words, but may convey a negative or positive opinion. For example, the sentence *The car consumes a lot of petrol* gives a negative opinion because most people prefer cars with economical petrol consumption. Sarcasm is another common challenge found in political discussions because it may not contain any positive or negative words. These challenges can also be overcome with machine learning and deep learning-based approaches.

Dealing with contextual information by employing shifters Lexicons that include only unigrams do not capture contextual information. For example, the sentiment conveyed by the sentence *inverts the growth of the economy* is negative, despite the presence of a positive word *growth* in our lexicon. Here, the word *inverts* shifts the semantic attribution from positive to negative. Therefore, contextually aware methods are needed to improve the performance of lexicon-based approaches to sentiment analysis. One way to solve this problem is to employ *sentiment shifters*, sometimes referred to as *valence shifters*. A shifter is a word or expression that changes the polarity of another word or expression [163], such as *invert*. Sentiment shifters can be created manually [163, 21], but can also be automatically generated using *seed shifters*. In Chapter 6, we proposed a context-aware approach that exploits shifter patterns.

2.4 Machine learning approaches to sentiment analysis

A significant amount of research in sentiment analysis employs machine-learning approaches for sentiment classification. These approaches can be divided into two categories: supervised and unsupervised machine learning methods for sentiment analysis.

Supervised sentiment classification methods use existing learning algorithms trained with sentiment-labeled data to classify unseen documents. It uses features such as n-grams or lexicons for sentiment classification tasks. Pang et al. [124] classified review documents based on overall sentiments, whether a review is positive or negative. The authors used three standard machine learning algorithms: Naive Bayes, Maximum Entropy, and Support Vector

Machine, and found that these learning algorithms outperformed a human-produced baseline. However, the performance of these machine learning methods on sentiment classification was suboptimal compared to traditional topic-based categorization, indicating that sentiment classification is more complex than topic classification.

In unsupervised sentiment classification, unlabelled documents are preprocessed and probabilistic modeling methods are used to detect sentiment [26]. Turney proposed one of the pioneering unsupervised learning methods [165]. This simple unsupervised learning algorithm classifies reviews based on the average semantic orientation of phrases that contain adjectives and verbs. The semantic orientation of a phrase is positive if it has good associations and negative if it has bad associations. A review is classified as recommended if it is positive and not recommended if it is negative.

However, with the availability of more data and computational resources, deep learning approaches have become increasingly popular in sentiment analysis and often provide better performance.

2.5 Deep Learning Approaches to Sentiment Analysis

Classical machine learning and lexicon-based methods use hand-crafted features such as n-grams or lexicons for sentiment classification tasks [155]. On the contrary, deep learning approaches employ neural network architectures to automatically learn features and provide state-of-the-art performance [161, 95, 40, 22]. However, despite providing better performance, the models obtained by deep learning approaches are less interpretable.

Deep learning architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNNs), and recently transformer-based models have been widely used for sentiment analysis [178]. Sharfuddin et al. [146] used the RNN-based model to classify binary sentiments in the Bengali language and achieved an accuracy of 85.67%. Similarly, Muhammad et al. [105] used word2vec and LSTM to classify hotel reviews in Indonesia and achieved an accuracy of 85.9%. Although originally developed for computer vision tasks such as image classification [50], CNN can be used for natural language processing tasks. Ouyang et al. [118] used word2vec with CNN and provided competitive performance for sentiment classification.

State-of-the-art deep learning approaches use the transformer architecture and exploit the idea of self-attention and positional encoding [176]. Transformers can handle longer word sequences, making them capable of learning from larger contexts. Transformer-based pre-trained models such as Bert [38] and Roberta [87] are used in many NLP tasks, such as text classification and machine translation. The multilingual variants of these models, including mBERT [38], enable users to adapt to other languages that were not included in training

[51]. More details of our work in this area are given in Chapter 6.

2.6 Hybrid Approaches to Sentiment Analysis

2.6.1 ML approaches in conjunction with sentiment lexicons

Researchers have explored a hybrid approach to sentiment analysis that combines the strengths of both learning-based and lexicon-based approaches for optimal accuracy [103, 131, 149]. In the work of Prabowo et al. [131], the authors leveraged the strength of rule-based classification and supervised learning. The combined approach achieved higher accuracy when applied to movie reviews, product reviews, and Myspace comments. Sommar et al. [149] performed Twitter sentiment analysis with a combination of a lexicon-based approach and a trained classifier. Their results achieve competitive performance. As deep learning approaches outperform traditional machine learning methods, recent studies explore hybrid approaches that combine deep learning with other techniques to improve performance further.

2.6.2 DL approaches in conjunction with sentiment lexicons: Neuro-symbolic approaches

Recent approaches have combined deep learning and lexicon-based approaches. This research line is sometimes called a neurosymbolic approach [60]. Various studies exploited Concept Net in their studies that involved deep learning [86]. Concept Net is a semantic network or knowledge graph consisting of triples (C_i, r, C_j) , where C_i and C_j are concepts and r a relation between them (e.g., has property, is used for, etc.). Ghosal et al. [47] addressed the problem of cross-domain sentiment analysis. They explored the role of external commonsense knowledge and in particular, ConceptNet to enrich the given text by providing both information about domain-general and domain-specific concepts. The proposed approach involves two steps. The first one creates a subgraph of ConceptNet based on source and target domains and trains a graph-convolutional autoencoder to learn concept embeddings. In the second step, the trained graph is used to extract document-relevant features for each instance, which were used in adversarial training. The authors used the Amazon review benchmark dataset that covers four domains (books, DVDs, electronic and kitchen appliances) for the experiments on domain adaptation in SA. The experimental results on tasks that involved one of the domains as the source (e.g., Books) and another domain as the target (e.g., DVD9) demonstrated the advantages of the proposed approach when compared to other alternative approaches.

Kocon et al. [73] used information from different knowledge bases to feed into their DL models. One of the approaches they considered is based on the Tailored Kepler model [171], which incorporated knowledge from a knowledge graph (KG). The results showed that

for small and medium-sized datasets, this approach led to significant gains in performance and more stable results. However, for large datasets, additional knowledge did not lead to improvements in most cases. Chapter 6, explores a neurosymbolic approach that combines an improved lexicon-based approach with a lexicon generated automatically using a pre-trained language model.

2.7 Comparative Study of Lexicon-Based, ML, and DL Approaches

This section discusses some work that compares symbolic, machine learning, and deep learning-based approaches to sentiment analysis. Atteweldt et al. [168] compared different approaches for sentiment analysis using a Dutch corpus of economic domains. The authors used off-the-shelf sentiment lexicons and showed that the lexicon did not perform well compared to machine learning (ML) methods, particularly deep learning methods such as Convolutional Neural Networks (CNN).

Tavares et al. [153] explored sentiment classification from economic news headlines in Portuguese newspapers. The authors have compared various approaches using a sentiment lexicon, four ML, and a rule-based approach. Among the ML approaches, SVM obtained the best results using lemmatized data (accuracy of 77.0%). Furthermore, the authors discuss a rule-based approach developed manually and achieved better performance than the ML approaches (accuracy of 86.3%).

Catelli et al. [31] also compared lexicon-based and deep learning-based approaches. They used a network of local grammars to compute the sentiment values of words surrounded by the contexts in which they occur. The sentence annotation was performed using embedded local grammar as a finite-state automaton. Local grammars are algorithms that, through grammatical, morphological, and lexical instructions, formalize linguistic phenomena, including text parsing. The results of both studies showed that while the deep learning approach was superior to the enhanced lexicon-based approach, the differences were not so high. For example, F1 of the DL reported by Catelli et al. achieved 84%, while the enhanced lexicon-based approach achieved 81%. The fact that the enhanced lexicon-based approach provided competitive performance to deep learning encouraged us to seek further improvements in this line of research, as discussed in Chapter 6.

2.8 Sentiment Analysis for African Languages

Africa has a rich linguistic history, with language contact, expansion, development of trade languages, language shifts, and language death occurring on several occasions. The continent is home to over 2000 languages, including 75 languages with at least one million speakers each. Africa also has a rich tradition of storytelling, poems, songs, and literature [29, 15]. In recent years, digital and social media have facilitated a proliferation of communication in which code-switching is the norm. Code switching, the practice of alternating between two or more languages within a single conversation, is particularly prevalent in Africa and has become increasingly common in these new forms of communication [162]. However, despite this linguistic richness, African languages have been comparatively underrepresented in sentiment analysis research. This is primarily due to the absence of a freely accessible dataset in these languages. Most existing work on sentiment analysis in African languages relied on automatically generated data. For example, Ogbuju et al. [112] translated an English sentiment lexicon [64] and manually added Igbo native words to create IgboSentiLex. Orimaye et al. [117] built a Yorùbá YouTube corpus and applied a translated SentiWordNet to the sentiment analysis task. Similarly, Oyewusi et al. [120] built a Pidgin tweet corpus and used a translated VADER English lexicon for sentiment analysis.

This low-resourceness problem is not unique to sentiment analysis, but affects NLP research as a whole in Africa[68]. Recent studies [109, 65] have examined how sociocultural factors hinder NLP in low-resource languages and can result in economic inequities [173]. Chapter 3 presents our contributions to developing sentiment analysis datasets and lexicons for African languages.

Chapter 3

Corpora

This chapter describes the existing corpora used in this work and its contributions to creating new corpora for sentiment analysis in African languages.

- **Film reviews (IMDB) dataset:** Movie review Dataset
- **EcoSentiPT:** Domain-specific sentiment corpus in European Portuguese
- **NaijaSenti:** Sentiment corpus for four Nigerian languages
- **AfriSenti:** Sentiment corpus for 14 low-resource African languages

For the corpora we built (NaijaSenti and AfriSenti), we propose text collection, filtering, processing, and labelling methods that enable us to create datasets for these low-resource languages. More details about each corpus are given in the following sections.

3.1 Film reviews (IMDB) dataset

The IMDB review corpus has been extensively studied in sentiment analysis and contains 50,000 reviews, split into 25,000 for training and testing. The corpus is labeled with star ratings ranging from the most negative to the most positive (+1 to +10) and is divided into three categories: negative (+1 to +4), neutral (+5 to +6), and positive (+7 to +10). The distribution of star ratings in the corpus is presented in Fig. 3.1.

We used the IMDB review dataset to generate a sentiment lexicon and perform sentiment classification tasks in Chapter 4.

Review	Sentiment
Don't know how this missed award nominations. Great film. Certainly of the calibre of "Beautiful Mind" Great acting, photography, script and drama. I can't imagine anyone not being entranced by this film...	9 (positive)
Admittedly, I watched this piece with already VERY low expectations. Dieter Bohlen is a rather untalented composer parvenu whose lack of talent is only surpassed by the size of his ego. This was the first cartoon movie that I watched that was 100 per cent humor free. It is rude, offensive, redneck and blatantly anti- women. As such, it is a creation befitting Bohlen, but the average viewer will be rather put off by it. No wonder that it was never shown in a cinema theater: It would've bombed BIG time! Not even the expense of 6.5 MegaEuros were able to save this utter piece of crap. Save your time... and money!.	1 (negative)

Table 3.1: Examples of positive and negative sentences in IMDB movie review dataset

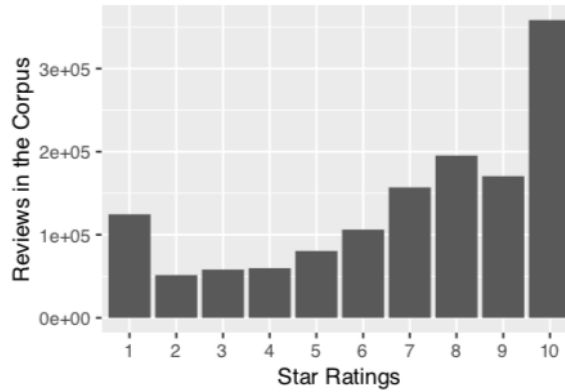


Figure 3.1: Star ratings distribution in the IMDB movie review corpus.

3.2 EcoSentiPT: A Financial Sentiment Analysis Corpus

We have used a domain-specific sentiment analysis corpus in the economic and financial domain in Portuguese called EcoSentiPT. This corpus was prepared by the linguists of the Faculty of Arts and Humanities of the University of Porto (FLUP). It involved 23 selected texts from online Portuguese newspapers that publish articles on finance and economics. Each text contained a varying number of sentences, ranging from 2 to 81 sentences, with a median of 12. The total number of sentences is 398, each with a different length ranging from 2 to 30 tokens. Table 3.2 shows examples of some sentences in this corpus. For the annotation task, the annotators followed predefined annotation guidelines, which consisted of sentences (or phrases) with appropriate sentiment values on a scale of -3 to +3. In such instances of disagreement, the other linguists adjudicated such cases and came up with a single sentiment value. Table 3.2 shows examples of some sentences/phrases and their ratings.

Sentence	Translation	Val.
Portugal não pode estar a governar só para os mercados , ou seja, para tentar demonstrar que o défice está melhor .	Portugal cannot govern only for the markets , that is, try to demonstrate that the deficit is better .	-1
Governo mostra-se mais inseguro e débil , e sem um rumo definido .	Government appears more insecure and weak , and without a defined direction .	-2
Quem lida com a exportação de serviços sabe que a falta de qualificação dos portugueses é uma falsa questão .	Anyone who deals with the export of services knows that the lack of qualifications of the Portuguese is a false issue .	1
O saldo positivo das nossas trocas compensa largamente o financiamento das atividades do país.	The positive balance of our exchanges largely compensates for the financing of the country's activities.	2

Table 3.2: Examples of some sentences and ratings in EcoSentiPT dataset

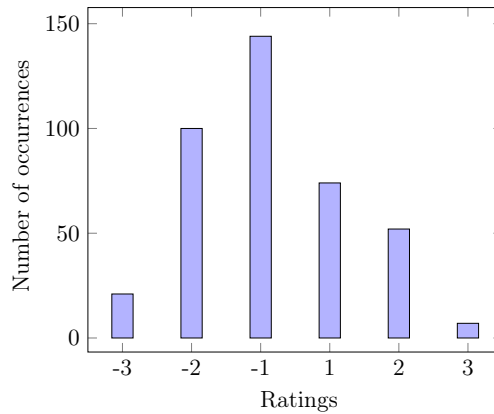


Figure 3.2: Distribution of ratings in EcoSentiPT dataset

The distribution of the ratings, as shown in Fig. 3.2, is skewed towards negative values. This probably reflects the general tendency observed in some studies, suggesting that people tend to report on various negative aspects more frequently than on the positive ones [144].

3.3 NaijaSenti: A Nigerian Twitter Sentiment Corpus¹

Sentiment analysis for low-resource languages has recently gained popularity [180, 177, 69] due to the availability of relatively large amounts of tweets in such languages. Several studies have investigated the use of Twitter for sentiment analysis—by automatically building a Twitter corpus or manually annotating one. Notable studies that automatically built Twitter corpora include [48, 121, 174]. More recently, [79] automatically built an Arabic Twitter sentiment analysis corpus using distant supervision and self-training. In contrast, other

¹I collaborated with AfricaNLP researchers from MasaKhaneNLP and HausaNLP to develop this dataset. The creation of the data set was funded by Lacuna Fund and I served as principal investigator for the project

studies, such as [137, 24, 102, 107, 101], employed native speakers to annotate the corpus manually. NaijaSenti corpus creation involves both the use of emoji as a distantly supervised approach to tweet extraction and the use of a translated sentiment lexicon to filter tweets before manual annotation [107]. Our approach is closely related to [8] and [79].

Despite advances in sentiment analysis, indigenous Nigerian, low-resource languages have received little attention. This is mainly due to the lack of a freely accessible dataset in these languages. Table 3.3 summarizes the existing datasets for Nigerian languages; only two datasets are freely available, indicating that more work is needed to make the indigenous datasets accessible to stimulate research in these languages. To our knowledge, NaijaSenti is the first publicly available large-scale manually annotated dataset for sentiment analysis research in the following Nigerian languages: Hausa, Igbo, Yorùbá, and Nigerian Pidgin. Nevertheless, there are a few works in Nigerian code-mixed English [111, 115, 119, 74, 134, 114, 116, 61].

Dataset	Language	Open-source	Annotated/translated	Code-mixed	Source
[3]	Hausa	✗	annotated	✓	Twitter
[112]	Igbo	✗	translated	✗	General
[166]	Igbo	✗	annotated	✗	General
[120]*	Pidgin	✗	annotated/translated	✓	Twitter
[117]	Yorùbá	✓	annotated	✓	Youtube
[66]	Yorùbá	✗	annotated	✗	General
Ours	Hausa, Igbo, Yorùbá, Pidgin	✓	annotated	✓	Twitter

Table 3.3: Summary of datasets used in six existing datasets on sentiment analysis in four major Nigerian languages in comparison to ours. The symbol "*" indicates that the provided URL is no longer accessible. Code-mixed means that the data set contains mixed language content.

3.3.1 Data Collection and Cleaning

Twitter’s Limited Support for African Languages Twitter provides easy access to a large amount of domain-independent and topic-independent public opinionated user-generated data. We collected tweets using the Twitter Academic API ², which provides historical and real-time tweet data. The Twitter API supports retrieving tweets in 70 languages (including Amharic as the only African language) using language parameters. This makes it easy to extract a tweet in these languages. The API supports none of the languages considered in this work. Therefore, we consider different heuristic approaches to crawling tweets. This disparity in language coverage highlights the need for further research and development in NLP for low-resource languages.

²<https://developer.twitter.com/en/products/twitter-api/academic-research>

Stopwords, emoji, and sentiment words Caswell et al. [30] have shown that token-based filtering is a useful processing step for automatic language identification. Therefore, we automatically built lists of common words (stopwords), which were verified by native speakers and used them to query the Twitter API to retrieve tweets in each language. Emoticons and emojis are used as a distance-supervised approach to automatically classify subjective tweets as positive or negative [48, 79]. Using a similar approach, we used *happy* and *sad* emojis [75] in combination with stopwords to query the Twitter API to extract tweets containing stopwords and emojis. In addition, we used the Google Language API to translate the Affin lexicon [183] into each of the languages (Hausa, Igbo, Yorùbá), except Pidgin. We then filtered the tweets using the verified translated Affin sentiment lexicon to improve the likelihood of annotating sentiment-bearing tweets [167].

Hashtags and Handles We used Twitter hashtags to crawl tweets from trending topics (e.g., #Yorubaday) to collect sufficient tweets that are expected to be in the language under consideration. We also collect tweets from news handles (e.g., @bbchausa) which are expected to be factual and non-subjective. We selected the handles that tweet frequently in each language from the Indigenous Tweets³ website.

One downside of this approach is that Twitter conversations with a popular Twitter handle may dominate the dataset and may introduce bias toward certain topics. For example, a Hausa Twitter conversation that involves the handle @bbchausa and another conversation involving the handle @Rahmasadau make up 54% and 14%, respectively, of collected tweets associated with Hausa handles. Limiting the number of tweets per conversation mitigates this problem.

Language Detection Some stopwords co-occur across indigenous languages in a multilingual society such as Nigeria [30]. This results in tweets being collected in a language that differs from the query language. For example, using the stopword “*nke*” to crawl tweets in Igbo produces tweets in Hausa, such as “*amin ya rabbi godiya nke*”. To mitigate this, we collected tweets based on locations where a language is spoken predominantly, using location, longitude, latitude and radius parameters (25 miles) to specify a circular geographic area.

We also used Google CLD3⁴ and Natural Language API⁵ to detect the language of the collected tweets. Pidgin is not supported by the API, so we used the stopword list to build an n-gram language detection tool to detect Pidgin. Before annotation, we cleaned the tweets. Retweets and duplicates were removed. We removed URLs and mentions, as well as trailing and redundant white spaces, converted all tweets to lowercase, and removed tweets with less

³<http://indigenoustweets.com/>

⁴<https://github.com/google/cld3>

⁵<https://cloud.google.com/natural-language/docs>

than three words as they may contain insufficient information for sentiment analysis [179].

3.3.2 Dataset Annotation

We developed an annotation guideline that focuses on the classification of subjective tweets. A subjective tweet has a positive or negative emotion, opinion, or attitude [138].

We adopt a sentiment annotation guide from [96] and define five classes: positive (POS), negative (NEG), neutral (NEU), mixed (MIX) and indeterminate (IND).

- **Positive (POS) Sentiment** This occurs if a tweet implies a positive sentiment, attitude and emotional state. For example, a tweet implies a positive opinion or sentiment (e.g., *"I love iPhone"*), positive emotional state (e.g., *"we won the game last night"*), expression of support (e.g., *"I will vote for PDP"*), thankfulness (e.g., *"thank god she has not been kidnapped"*), success (e.g., *"I passed all my exams"*), or positive attitude.
- **Negative (NEG) Sentiment** This occurs if a tweet implies negative sentiment or emotion. For example, if a tweet implies negative sentiment (e.g., *"The iPhone camera is bad"*), negative emotional states such as failure, anger, and disappointment.
- **Neutral (NEU)** This occurs if the user's tweet does not imply any positive or negative language directly or indirectly. These are usually factual tweets, such as news.
- **Mixed (MIX)** This occurs if the user's tweet implies both negativity and positivity directly or indirectly. For example, *"I love an iPhone 10, but its camera is bad"*.
- **Indeterminate (IND):** This occurs if the user's tweet does not fall into the categories discussed above, or if the annotator can only guess the class of a tweet, especially in the case of proverbs or sarcasm without sufficient context. We additionally use this class to label tweets in a different language (not code-mixed).

Annotators training and preparation For each language, three native speakers were recruited as annotators. Annotators are men and women between the ages of 20 and 45. We also recruited a coordinator for each language to supervise and ensure the quality of the annotation task. Annotators and coordinators have backgrounds in computer science or linguistics and were trained on the annotation task using the LightTag annotation tool [126]. Data annotation is not a one-off process; it requires an agile approach with many iterations, collecting feedback from the annotators during the pilot stage, and refining the annotation guide to ensure that the annotators can achieve reasonable performance before moving to the next stage. We performed three iterations of the training and annotation practice of 100

tweets. For the first two iterations, the agreement among the annotators was poor. We asked the annotators for feedback and adapted a simplified sentiment questionnaire annotation guide [96].

Tweets annotation The dataset was annotated in batches of 1,000 tweets by the annotators. For each batch, we adjudicated the cases in which the three annotators assigned a different label to a tweet. Annotators discuss these tweets, which allows them to address ambiguities and peculiar issues, and recommend ways to improve the annotation guidelines. We excluded these ambiguous tweets from the dataset. We iteratively update our annotation guide based on adjudication reports. Overall, the annotators annotated the following number of tweets: Hausa (35,000), Igbo (29,000), Pidgin (30,000) and Yorùbá (33,000).

Determining the gold standard label People often disagree on subjective concepts [18]. For example, person A, who has been using Apple products, says, “The Apple iPhone camera is better than the Samsung camera”, while person B says, “The Samsung camera is better”. This is an example of subjective disagreement in contrast to objective disagreement. Therefore, different from the simple majority vote approach [37], we introduced a new form of majority vote that involves an independent annotator who adjudicates subjective disagreement cases as follows:⁶

- **Three-way agreement:** Similar to the majority vote approach, if the three annotators agree on a label, we consider the agreed sentiment class the gold standard.
- **Three-way disagreement:** When all annotators disagree on a label, we discard the tweet.
- **Two-way partial disagreement:** This occurs if two of the annotators agree on a label, and the third annotator has a partial disagreement. For example, if two annotators classify a tweet as POS (or NEG), and the other annotator classifies it as a non-contradicting class such as NEU, we consider the POS (or NEG) classification the gold standard.
- **Two-way total disagreement:** This occurs if two annotators agree on a label, and the third annotator has a total disagreement. For example, if two annotators identify a tweet as POS and another as NEG or vice versa, the majority vote is not the final class (in this case, POS). To resolve such subjective disagreement, independent annotators review the disagreement and assign a final label.

⁶We determine a single gold standard label for sentiment analysis in accordance with prior work. Future work may alternatively leverage annotator disagreement [44].

		Sentiment datasets			
	sent.	hau	ibo	yor	pcm
5-class	POS	9,235	5,621	9,839	7,038
	NEG	9,033	4,726	5,003	11,774
	NEU	12,826	14,877	14,356	2,205
	IND	8	1,909	1,754	2,651
	MIX	1,466	19	622	1,696
	Total	32,568	27,152	31,574	29,837
	IAA (κ)	0.487	0.488	0.555	0.347
3-class	POS	8,019	5,395	9,391	5,839
	NEG	8,119	4,513	4,638	9,400
	NEU	11,122	13,380	13,367	2,004
	Total	27,260	23,288	27,396	17,243
	IAA (κ)	0.607	0.516	0.600	0.457

Table 3.4: 3-class and 5-class annotation and inter-Annotator agreement. hau= hausa, ibo = Igbo, yor = Yoruba and pcm = Nigerian-Pidgin

corpus				
Class	hau	ibo	yor	pcm
POS	0.626	0.542	0.626	0.347
NEG	0.518	0.521	0.553	0.416
NEU	0.442	0.404	0.491	0.310
MIX	0.297	0.020	0.242	0.130
IND	0.045	0.591	0.764	0.679

Table 3.5: Fleiss kappa agreement among each class

Inter-Annotator Agreement (IAA) We used the Fleiss kappa (κ) reliability measure [43] to determine the inter-annotator agreement between the three annotators. The IAA for the 5-class and adjusted 3-class agreements are shown in Table 3.4. The agreement between the five classes was not particularly high (e.g., (κ) = 0.35) for Pidgin. However, according to the Fleiss classification [43], an agreement greater than 0.40 is considered reasonable (moderate) and beyond chance.

We further computed the IAA (κ) (see Table 3.5) of each class with other classes to determine which classes the annotators find confusing or difficult and frequently disagree with. Table 3.5 indicates that annotators generally have the lowest overall agreement in the MIXED class, which includes elements of both the positive and negative classes and some annotators identify it as either negative or positive. This highlights the subtlety of annotating mixed sentiment on social media and is in contrast to reviews where the annotation of mixed sentiment is clearer [130]. To address this, we introduced an adjusted 3-class IAA agreement.

In the adjusted 3-class agreement, we considered only positive, negative, and neutral as valid

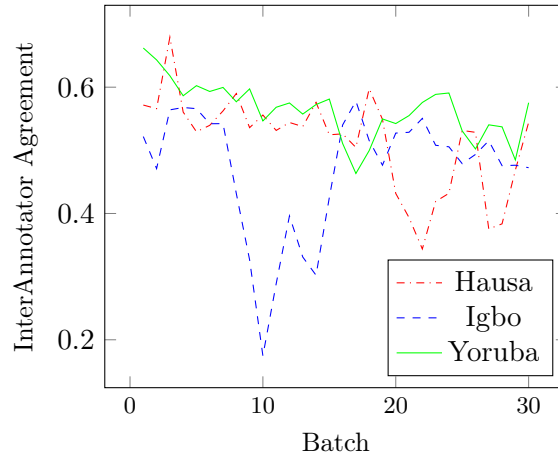


Figure 3.3: Inter-annotators progress over thirty batches—one-thousand tweets per batch.

sentiment classes. We selected only tweets with at least two labels in the valid classes and discarded the rest. For the selected tweets, where the label between two annotators is valid and the third label is in the invalid sentiment (Indeterminate or Mixed), we changed the label to the agreed valid label. For instance, given three annotation labels of a tweet as POS, POS, and MIX, the third label is changed to POS, whereas the annotation labels of POS, POS, and NEU are left unchanged. Table 3.4 shows the final statistics of at least two agreed tweets of the various datasets after converting to the 3-class annotation, and their corresponding inter-annotator agreements (IAA) using the Fleiss’ Kappa (κ) metric.

To determine the performance of the IAA over time, we have shown the IAA in three languages over 30 batches in Figure 3.3. We hypothesised that their annotation quality would improve as the annotators became more experienced with the task. However, the overall performance of the IAA deteriorates over time. Igbo has the lowest performance drop. This suggests that familiarity with the task does not necessarily improve IAA. Only Yorùbá annotators have some level of consistency that is not below 0.5. Therefore, it is essential to monitor the IAA as the annotation progresses and use some form of random quality check.

3.3.3 Human Evaluation

We assess human performance by re-annotating 200 random sample tweets by three different annotators [172, 108]. We take the majority vote as the final class. Human performance offers us an idea of the upper bound performance of the automatic approach and the reproducibility of the first three annotators [172]. Table 3.6 shows the micro-F1 and Matthew’s correlation coefficient (MCC) [70]. The human performance result validates the reliability of the corpus and is in accordance with previous literature [139].

Measures	hau	ibo	yor	pcm
Micro-F1	0.75	0.76	0.85	0.78
MCC	0.63	0.68	0.77	0.69

Table 3.6: Human performance result using micro-F1 and Matthew’s correlation coefficient.

3.3.4 NaijaSenti Statistics and Analysis

Table 3.7 shows the summary of the NaijaSenti dataset, including the number of monolingual and code-mixed tweets in each dataset. The percentage of code-mixed tweets highlights the highly multilingual setting in Nigeria. Code-mixing is more prevalent in Igbo (43%) than in Hausa (23%) and Yorùbá (19%). Code-mixing between English and a native language is more common than between native languages but can also occur between more than two native languages.

Hausa does not have diacritics and therefore has an insignificant number of indeterminate cases (only 8), unlike Yorùbá and Igbo where the absence of diacritics may render a tweet incomprehensible and therefore lead to labelling it as indeterminate. Pidgin has the highest number of indeterminate cases. This is because some tweets appear to be pidgin, but they are Nigerian English and, therefore, we consider them indeterminate.

Languages	mono-lingual	#code-mixed	token	neg words	pos words
Hausa (hau)	21,039	6,426	3,493,92	1,008	1,214
Igbo (ibo)	8,688	6,561	1,830,02	1,180	904
Naija (pcm)	—	—	3,669,68	—	—
Yorùbá (yor)	18,662	4,457	40,897,6	2,185	2,228

Table 3.7: Key stats of NaijaSenti: #mono-lingual tweets, #code-mix tweets, #token

Tone in Yorùbá helps to give meaning to words in a specific context, especially words that have the same orthographic representation. For instance, the sentence “*Awon omo fo abo*” does not have a meaning without diacritics, and the annotators classify it as indeterminate (IND). However, the same sentence with diacritics can have two opposite meanings: *Àwọn ọmọ fọ abọ* (The children washed the dishes) has a positive meaning, while *Àwọn ọmọ fọ abọ* (The children broke the dishes) is negative.

Similarly, tonality is heavily used in Igbo. Many Twitter users do not write Igbo with diacritics. One reason is the lack of an Igbo keyboard that accepts and shows diacritics. Even if such a keyboard exists, it is not used by many. Although it may be easy to understand the sentiment of Igbo tweets on Twitter—either due to the presence of *emojis* or the context of the surrounding discourse, it is difficult to annotate the tweets correctly. The example below highlights the impact of tone and punctuation marks on the same Igbo tweets but

with different sentiments:

- ò nwèkwàrà mgbe i naenwe sense ? – Will you ever be able to talk sensibly? – You’re a fool.
- ò nwèkwàrà mgbe i naenwe sense – Sometimes you act with great maturity – I’m impressed.

NaijaSenti corpus is the largest sentiment corpus in four major Nigerian languages. In the next section, we present AfriSenti- an extension of the NaijaSenti corpus that includes 14 African languages, which were subsequently used in the Africentric SemEval-Shared Task—SemEval Shared Task 12: Sentiment analysis for African languages (AfriSenti-SemEval).

3.4 AfriSenti: Sentiment Corpus for African Languages⁷

The AfriSenti dataset was developed for the AfriSenti-SemEval Shared Task. SemEval is a workshop that focuses on advancing natural language processing research by improving semantic analysis and creating high-quality annotated datasets. SemEval hosts workshops that feature shared tasks to showcase and compare computational semantic analysis systems developed by different teams⁸. Despite the prominence of the sentiment classification task in the SemEval competition, the previous tasks were mainly in English and other high-resource languages. To motivate research on African languages, particularly sentiment analysis, we extended the NaijaSenti dataset to the AfriSenti dataset using a similar methodology.

The main goal of the AfriSenti-SemEval shared task is to facilitate sentiment analysis research in African languages. Some of the important benefits of the task include the following:

- (1) Giving researchers the opportunity to solve sentiment analysis tasks for African languages;
- (2) Establishing a performance benchmark for future research comparisons in African languages;
- (3) Providing publicly available corpora that offer opportunities to investigate sentiment analysis and foster the creation of more low-resource African datasets; and
- (4) Exploring lexical and semantic differences in African languages.

AfriSenti covers 14 African languages, each with unique linguistic characteristics and writing systems, which are shown in Figure 3.4. We use the same annotation guides and strategy

⁷I conceived the SemEval project idea and collaborated with African NLP researchers from HausaNLP and MasaKhaneNLP. The creation of the dataset was partially funded by LacunaFund and I serve as the Principal Investigator for the project.

⁸<https://semeval.github.io/>

Lang.	Tweet	Sentiment
amh	ያ ጤካኝ አረመኔ ታስሮ ይኸው ካቴና ገብቶለታል ይሉናል። ቆይ አስረው የጀበና ቡና አየጋበዙት ነው እንዴ?	negative
arq	@user الشروق هذه من خرجت وهي نتاج تبديل، مستوى منحنط وشعبوي	negative
ary	واش بغيتوهم يبدوا يتكرفسو على العادي والبادي عاد تبقاو أمتا على خاطر خاطركم	negative
ary	rabi ykhali alhbiba makayn ghir nachat o chi machat	positive
hau	@USER Aunt rahma i luv u wallah irin totally dinnan	positive
ibo	akowaro ya ofuma nne kai daalu nwanne mmadu	positive
kin	@user Ariko akokanu ngo inyebebe unyujijemo sisawa wangu	negative
orm	@user Jawaar Kenya OMN haala akkamiin argachuu dandeenya	neutral
por	Honestidade é algo que não se compra. Infelizmente a humanidade esqueceu disso por causa das suas ambições.	positive
pcm	E don tay wey I don dey crush on this fine woman ...	positive
swa	Asante sana watu wa Sirari jimbo la Tarime vijijini Huu ni Upendo usio na Mashaka kwa Mbunge wenu John Heche	positive
tir	@user ከመክረኩም እንተተኾይነ፡ንሐውሓት ነፃም ውሑድ ቁጽሮም እባ ምጥፋሊ ይሕሽ ኩም!	negative
tso	@user @user Yu , tindzava ? Tsika mbangui mpfana e nita ku desprogramara	negative
twi	messi saf den check en bp na wo kwame danso wo di twe da kor aaa na wawu	negative
yor	onírèégbè aláádúgbò ati olójúkòkòrò	negative

Table 3.8: Examples of tweets and their sentiments in the different AfriSenti Languages. Note that the collected tweets in Moroccan Arabic (ary) are written in both Arabic and Latin scripts.

Lang.	3-way											2-way		
	arq	ary	hau	ibo	kin	pcm	pt-MZ	swa	tso	twi	yor	ama	orm	tir
κ	0.41	0.62	0.66	0.61	0.43	0.60	0.50	—	0.50	0.51	0.65	0.47	0.20	0.51

Table 3.9: Inter-annotator agreement scores using the free marginal multi-rater kappa [135] for the different languages.

in NaijaSenti for AfriSenti and report the free marginal multi-rater kappa scores [135] in Table 3.9 since chance-adjusted scores such as Fleiss- κ can be low despite a high agreement due to the imbalanced label distributions [135, 42, 91]. We obtained intermediate to reasonable levels of agreement (0.40 – 0.75) in all languages except Oromo, where we obtained a low agreement score due to the annotation challenges faced by annotators despite multiple annotation training.

Writing Systems and Language Family African languages, with a few exceptions, use Latin script, written from left to right, or Arabic script, written from right to left [46, 92], with Latin script being the most widely used in Africa [41]. Ten languages of the fourteen in AfriSenti are written in Latin script, two in Arabic script, and two in Ethiopic (or Geez) script. On social media, people may write Moroccan Arabic and Algerian Arabic in both

	ama	arq	hau	ibo	ary	orm	pcm	pt-MZ	kin	swa	tir	tso	twi	yor
train	5,985	1,652	14,173	10,193	5,584	-	5,122	3,064	3303	1,811	-	805	3,482	8,523
dev	1,498	415	2,678	1,842	1,216	397	1,282	768	828	454	399	204	389	2,091
test	2,000	959	5,304	3,683	2,962	2,097	4,155	3,663	1027	749	2,001	255	950	4,516
Total	9,483	3,062	22,155	15,718	9,762	2,494	10,559	7,495	5,158	3,014	2,400	1,264	4,821	15,130

Table 3.10: Sizes and splits of the AfriSenti datasets. We do not allocate training splits to Oromo (**orm**) and Tigrinya (**tir**) due to the limited size of the data and only evaluate them in a zero-shot transfer setting

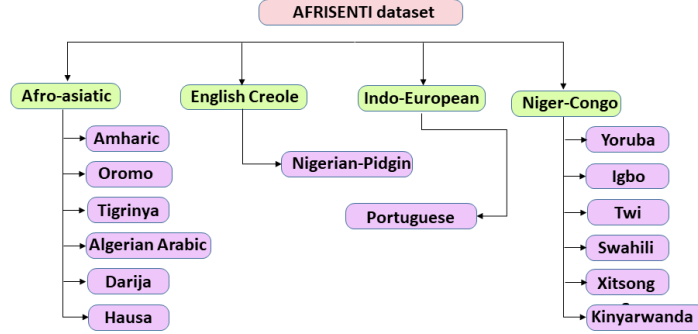


Figure 3.4: Language Family in the AfriSenti datasets.

Latin and Arabic characters for various reasons, including access to technology, i.e., the fact that Arabic keyboards were not easily accessible on commonly used devices for many years, code-switching, and other phenomena. This makes Algerian and Moroccan Arabic digraphic, i.e., their texts can be written in multiple scripts on social media. Similarly, Amharic is digraphic and written in Latin and Geez script [19]. This presents an additional challenge to the processing of these languages in NLP. Despite our efforts to detect the right languages, it is worth mentioning that, since multilingualism is common in African societies, the final dataset contains many code-mixed tweets. As shown in Figure 3.4, the AfriSenti dataset includes six languages of the Afroasiatic family, six languages of the Niger-Congo family, one from the English Creole family, and one from the Indo-European family.

Geographic Representation and Annotation Challenges AfriSenti covers the majority of African subregions. Many African languages are spoken in neighbouring countries within the same sub-regions. For example, Hausa variations are spoken in Nigeria, Ghana, and Cameroon, while Swahili variants are widely spoken in East African countries, including Kenya, Tanzania, and Uganda. AfriSenti also includes datasets in the top three languages with the most African speakers (Swahili, Amharic, and Hausa). We show the geographic distribution of the languages in AfriSenti in Figure 3.5. Sentiment annotation for African languages presents some challenges [161]. For example, a significant portion of tweets in *Twi* was ambiguous, making it difficult to accurately categorize sentiment. Some tweets

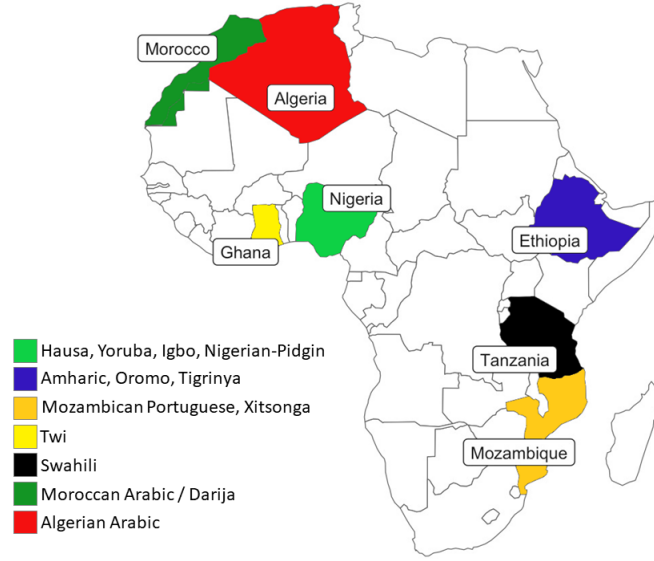


Figure 3.5: Countries and languages represented in the AfriSenti data collection (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá).

contained symbols not in the Twi alphabet, which is a frequent occurrence due to the lack of support for certain Twi letters on keyboards [142]. Similarly, a significant challenge for the Mozambican Portuguese and Xitsonga data annotators was the presence of code-mixed and sarcastic tweets. Code-mixing in tweets made it challenging for annotators to determine the intended meaning of the tweet, as it involved multiple languages spoken in Mozambique that some annotators did not understand. Similarly, the presence of two variants of Xitsonga spoken in Mozambique (Changana and Ronga) added to the complexity of the annotation task.

3.5 Summary of Corpora

This section summarizes the information about each corpus and where it is used in subsequent chapters (see Table 3.11). The column "*language*" indicates which languages the data set was created from, and the column "Lexicon creation" shows whether the sentiment lexicon was created from the corpus manually or automatically. The table also presents the source of each corpus, the chapter in which it is used, and whether the corpus is used for a sentiment classification experiment.

Corpora	Language	Lexicon creation	Sentiment classification	Chapter used	Source
IMDB*	en	automatic	✓	4	Reviews
EconSentiPT*	pt	manual/automatic	✓	4, 5, 6	News
NaijaSenti	ha/ig/yo/pg	manual/automatic	✓	6	Twitter
AfriSenti	14 lang.	-	✓	6	Twitter

Table 3.11: Summary of the four datasets used in our work. The symbol * indicates an existing corpus.

3.6 Conclusions

In this chapter, we introduce corpora used in this thesis, including pre-existing and newly created resources. Our main contributions include the development of the following:

- **NaijaSenti:** We created a new sentiment corpus for four Nigerian languages [161].
- **AfriSenti:** We created a new sentiment corpus for 14 low-resource African languages [159]. This dataset was used in the first Afrocentric-SemEval shared task we organized [154].⁹

We have provided a detailed description of the steps used in preparing these datasets and the challenges encountered in creating them. The following chapters will explain how the corpora discussed here generate a sentiment lexicon and prediction.

⁹<https://afrisenti-semeval.github.io/>

Chapter 4

Automatic Lexicon Generation

In this chapter, we introduce a novel approach for automatically generating sentiment lexicons. We present a manually constructed lexicon for three low-resource Nigerian languages and demonstrate the need for an automated solution. Next, we propose two new methods for generating domain-specific sentiment lexicons and present an experimental study using the generated lexicons. Our experimental study shows that our generated lexicon provides competitive performance in lexicon-based sentiment analysis.

4.1 Methodology of Lexicon Creation

Sentiment lexicons are essential for lexicon-based sentiment analysis, which provides a simple and efficient approach to identifying sentiment. Sentiment lexicons are particularly useful in low-resource settings—where data and computational resources are scarce. Importantly, lexicon-based sentiment analysis offers an interpretable solution that can be easily understood and explained to non-technical audiences.

We introduce two novel methods for constructing sentiment lexicons. We begin by presenting a manual approach to creating three lexicons using the NaijaSenti dataset discussed in Section 3.3. Next, we propose two automated methods for generating a sentiment lexicon: a statistical/ML-like method and a method that leverages pre-trained language models. Additionally, we present a method for extending existing domain-specific lexicons, allowing flexibility to extend and apply lexicon-based sentiment analysis in different domains.

4.1.1 Manual generation of sentiment lexicon using NaijaSenti dataset

Manually creating a sentiment lexicon involves compiling a list of sentiment-bearing words (positive, negative, and neutral) through annotation by annotators fluent in the language and

sometimes familiar with the domain in which the lexicon will be used. The lexicon is then reviewed to ensure accuracy and consistency among the annotators, and any inconsistencies are resolved through adjudication. In situations where a sentiment lexicon is not available, for example, in a low-resource language, some authors have attempted to translate a lexicon from a high-resource language (e.g., English) to the low-resource target language [112]. However, translation can alter the sentiment of words and affect the quality of sentiment prediction [99].

The NaijaSentiLex sentiment lexicon was developed manually using the NaijaSenti dataset for three major Nigerian languages [161]¹. Using an annotation guideline, three trained annotators were assigned to tag words that conveyed negative or positive sentiments in tweets. Due to the large number of tweets available for each language in the NaijaSenti dataset (approximately 30,000), tagging all tweets could lead to redundant annotation of the same words. Therefore, we use an incremental process as an alternative strategy to create a sentiment lexicon from this dataset. First, a sample of 1000 tweets was tagged by the trained annotators, who identified positive and negative words. This initial set of annotations was used to construct a small lexicon, which was then used to filter out untagged tweets that contained at least one previously tagged word. For example, if the word *good* were identified as positive in the current iteration, any tweet containing *good* would be removed from subsequent iterations to avoid redundant annotation. By iteratively expanding the lexicon in this way, we were able to create a sentiment lexicon for sentiment analysis in these languages. Table 4.1 shows a sample of lexicons in each language.

To determine the final sentiment of a given word, we use a simple majority vote method [37]. For example, if two annotators labelled a word as positive and one annotator labelled it as negative, the majority view was taken into account, and the word was considered positive. In cases of disagreement among annotators or instances where only one annotator provided a label, an independent annotator was called upon to adjudicate such cases and assign a final label. Approximately 5,000 tweets were tagged in each language to generate the final sentiment lexicon. Table 4.1 shows a sample of lexicons and translations in each language.

Hausa		Igbo		Yoruba	
HausaLex	Senti.	IgboLex	Senti.	YoroLex	Senti.
qarya (<i>lie</i>)	neg	igbu (<i>kill</i>)	neg	aisododo (<i>injustice</i>)	neg
wuya (<i>difficult</i>)	neg	nsogbu (<i>problem</i>)	neg	abirun (<i>disabled</i>)	neg
ciwo (<i>pain</i>)	neg	etigbu (<i>hit</i>)	neg	àisàn (<i>illness</i>)	neg
kyauta (<i>gift</i>)	pos	ekete (<i>congratulations</i>)	pos	kágbn (<i>wise</i>)	pos
soyayya (<i>love</i>)	pos	oganiru (<i>prosperous</i>)	pos	itelorun (<i>satisfaction</i>)	pos

Table 4.1: Examples of a manually generated sentiment lexicon for three Nigerian languages

¹I worked with researchers from HausaNLP to annotate this lexicon, and I am the lead researcher

The manual approach to creating sentiment lexicons is laborious and time-consuming. Additionally, the semantic orientation of a word can vary depending on the domain in which it is used, so manually generating a domain-specific sentiment lexicon for each domain is challenging. It is also impractical to generate a complete sentiment lexicon for a domain from a single corpus due to several reasons including insufficient entries, domain transfer, or language evolution. In the next section, we propose two methods for automatically generating domain-specific sentiment lexicons using statistical approaches and pre-trained language models. We also explore extending general-purpose lexicons to domain-specific ones.

4.1.2 Automatic lexicon generation using a statistical/ML-like method

This section presents a heuristic-based approach for automatically generating domain-specific sentiment lexicons. We use two datasets: (1) the EcoSentiPT dataset from the financial domain discussed in Section 3.2, and (2) the IMDB dataset from the movie review domain discussed in Section 3.1.

Our approach extends the work of Almatarnah et al. [11] discussed in Section 2.3.3. Their approach classifies sentiment-bearing words, considering only *adjectives* and *adverbs*, as either negative or positive. In contrast, our approach uses four lexical classes: *nouns*, *verbs*, *adjectives*, and *adverbs*. This is because words from the other classes can also convey sentiment. For instance, verbs such as *criticize* and *insult* have negative connotations. Additionally, we do not eliminate *stopwords* in this approach. This is because *stopwords* in Portuguese includes words such as “*não*” (*no*), “*mais*” (*more*), “*muito*” (*many*), which play an essential role in the shifter patterns discussed in the next chapter.

Another issue regarding the Almatarnah et al. [11] method is that it does not account for the length of reviews, which results in a sub-optimal lexicon. To illustrate the limitations of their approach, consider two sentence fragments that both include the word *taxes*:

- *low taxes* / 1
- *taxes* / -1

If the approach of Almatarnah et al. [11] were followed, the sentiment value of the word *taxes* would be calculated as the mean of the sentiment values of the two sentence fraction i.e., $(1 + (-1))/2$ giving 0. This does not seem right as the second phrase includes just one word (*taxes*) and hence should influence the final sentiment value more. Thus, we extended the work, and our approach considers the length of labelled sentences or fractions, as these can vary in length (when considering sentiment-bearing words). Our approach assigns different weights to different phrases/fractions while considering the number of sentiment-bearing words in

each phrase. The larger the number of sentiment-bearing words, the smaller the weight. To demonstrate why our approach is effective, consider the same two-sentence fractions again.

Since the first phrase/fraction includes two sentiment-bearing words (*taxes, low*), its weight is $1/2$. The second phrase is just one word, so its weight is 1. The final sentiment value SV_{ti} of word/term ti is calculated as the weighted mean of all values, as expressed by the following formula:

$$SV_{ti} = \frac{\sum R_{pj,ti} * W_{pj}}{W_0 + \sum W_{pj}} \quad (4.1)$$

where $R_{pj,ti}$ represents the rating of term ti in phrase/fraction pj and W_{pj} the respective weight of the phrase/fraction pj . Symbol W_0 represents a small constant (e.g., 0.5) that provides a kind of smoothing. When the term ti is relatively rare, the value of $\sum W_{pj}$ will be rather small and so W_0 will pull the sentiment value down. Hardly any effect can be observed for terms that are relatively frequent. Let us consider our previous example again. With $W_0 = 0.5$, we get :

$$SV_{imposto} = \frac{(1 * 0.5) + (-1 * 1)}{0.5 + (0.5 + 1)} = \frac{-0.5}{2} = -0.25 \quad (4.2)$$

As we can see, the weight of the shorter phrase (*taxes*) influences the final outcome, and therefore, the final sentiment value is negative (-0.25).

Splitting Long Sentences In our experiments to create lexicons from the EcoSentiPT dataset, we have observed that long texts represent a challenge when trying to induce high-quality sentiment lexicons. For example, when a sentence with 20 words is rated with negative sentiment (-1), it is difficult to identify which word(s) contributed to this negative rating. To address this issue, long sentences were split into shorter fractions, which were labelled with the appropriate new rating by linguists of the Faculty of Arts and Humanities of the University of Porto (FLUP)². The shorter fraction typically included only words that appeared to affect the overall sentiment value, while other irrelevant words were discarded.

To illustrate the process of splitting with an example, consider the sentence: "*Portugal não pode estar a governar só para os mercados, ou seja, para tentar demonstrar que o défice está melhor*(Portugal cannot govern only for the markets, that is, try to demonstrate that the deficit is better.)" This sentence has a negative sentiment orientation (-1). It was divided into the following fractions and words with appropriate ratings, as shown below:

- *governar só para os mercados* (govern only for the markets) / -2,

²Prof. P. Brazdil also collaborated on the creation of the corpus

- *défice* (deficit) / -2,
- *demonstrar que o défice está melhor* (demonstrate that the deficit is better) / 2.

The annotators have generated more than 1700 new sentence fractions by applying the splitting approach to the original long sentences. The new sentence fractions were used as source data for generating a domain-specific sentiment lexicon for the domain of finance and economics in Portuguese. Consequently, we have two versions of EcoSentiPT— a version with long sentences and the corresponding version with sentence fractions (used for generating lexicon).

EcoLex We used the Equation (4.1) and generated domain-specific lexicons dubbed EcoLex using the *EcoSentiPT* dataset—sentence fractions version. Table 4.2 shows a sample of the generated lexicons. The column labeled "*EcoLex*" shows the sentiment value generated for each word using our proposed approach, while the column labeled *SentiLexPT'* shows a unigram sentiment values from SentiPT—a general purpose sentiment lexicon that is a Portuguese adaptation of Vader sentiment lexicon. As we can observe, the general-purpose lexicon entries (*SentiLexPT'*) do not contain a specific lexicon from the financial domain, such as the word "*competitivo (competitive)*", which may limit its performance. Also, Table 4.3 shows the distribution of four lexical classes (N, V, ADJ, ADV) in Ecolex.

Word	class	EcoLex	SentiLexPT'
competitivo (competitive)	Adj	1.33	-
adequado (adequate)	Adj	0.40	1
consolidação (consolidation)	N	0.58	-
alto (high)	Adj	-0.39	-
ativo (active)	Adj	0.33	1
capital (capital)	N	-0.52	-
absurdo (absurd)	Adj	-	-
agressivo (aggressive)	Adj	-0.93	-

Table 4.2: Sample of lexicon entries generated using statistical-based approach

Class	Count	%
N (noun)	595	47.8
V (verb)	329	26.5
ADJ	251	20.1
ADV	64	5.1
PRON	3	0.2
Total	1246	100

Table 4.3: Distribution of four different lexical classes in EcoLex

IMDBLex Similarly, we used the same formula in Equation (4.1) to generate a movie review lexicon called IMDBLex from the IMDB review dataset. Table 4.2 shows a sample of generated lexicon entries. The column labeled "*IMDBLex*" shows the sentiment value generated for each word using our approach, while the column *AlmatarLex* shows sentiment values generated using the Almatarneh et al. [11] approach. As can be observed, our approach assigns sentiment values to words with different precision compared to Almatarneh et al.’s approach. In Section 4.3.1, we present a performance comparison that uses our sentiment analysis system (SA) with the two lexicons for sentiment classification.

Word	class	IMDBLex	AlmatarLex
bad	Adj	-1.27	-0.08
excellent	Adj	1.27	0.06
stupid	N	-0.72	-0.04

Table 4.4: Sample of Lexicons generated using statistical-based approach

Recent advances in NLP, particularly transformer architecture, have led to state-of-the-art performance on various tasks. The following section presents an alternative approach for automatically creating a domain-specific sentiment lexicon by leveraging a transformer-based pre-trained language model.

4.1.3 Sentiment lexicon generation using Bert-based language model

Transformer-based pre-trained deep neural networks have been shown to achieve good performance on various tasks, including sentiment analysis [170, 161]. Despite their success, no previous work explored creating a sentiment lexicon using these models. Our proposed method fills this gap and offers a promising way to build sentiment lexicons this way.

The generation of sentiment lexicons using this approach involves the following two steps: fine-tuning a pre-trained language model (PLM) and generating the lexicon by querying this model. We used a Bert-based pre-trained model from the HuggingFace transformer tool [176] for model training. The model was fine-tuned with domain-specific labeled training data using a batch size of 32, a maximum sequence length of 128, 20 epochs, and a learning rate of $5e-5$. In the lexicon generation step, we generated a list of words that were both frequent (more than ten occurrences) and informative (not stopwords) in the corpus. We then used the trained model to assign each word a sentiment score. We assign a sentiment score to each word if the model was fine-tuned with a dataset annotated with continuous values (e.g., SentiLex dataset). On the other hand, if the model was fine-tuned with a dataset annotated with classes (e.g., NaijaSenti dataset), we assigned a sentiment class to each word. Using this approach, we generate the following two sentiment lexicons:—*DeepEcoLex*, obtained from the EcoSentiPT dataset, and *DeepNaijaLex*, obtained from the NaijaSenti dataset.

DeepEcoLex We generated DeepEcoLex using a modified version of the EcoSentiPT dataset, annotated with continuous sentiment values ranging from -3 to +3. We fine-tuned a Portuguese Bert-based pre-trained language model called BERTimbau-based [150] with the EcoSentiPT sentence fractions dataset as a regression task. The trained model was then used to predict the sentiment values for individual words. Table 4.5 shows examples of sentiment values generated from both the statistical/ML-based and Bert-based pre-trained language model-based approaches. DeepEcoLex lexicon includes more entries than Ecolex.

Word	Class	SV DeepEcoLex	SV EcoLex
competitivo (competitive)	Adj	1.86	1.33
adequado (adequate)	Adj	1.84	0.40
consolidação (consolidation)	N	1.54	0.58
alto (high)	Adj	1.25	-0.39
ativo (active)	Adj	1.07	0.33
capital (capital)	N	1.03	-0.52
ajustamento (adjustment)	N	1.01	-
adorar (to love)	V	1.00	-
abuso (abuse)	N	-1.22	-
absurdo (absurd)	Adj	-1.53	-
agressivo (aggressive)	Adj	-2.05	-0.93

Table 4.5: Examples of some lexicon entries in DeepEcoLex and EcoLex

DeepNaijaLex We generated the DeepNaijaLex from the NaijaSenti dataset (annotated with three classes: positive, negative, and neutral), which comprises three sentiment lexicons for Hausa, Yoruba, and Igbo languages. For model-fine tuning, we used AfriBerta [113], a small and multilingual language model trained on only African languages that have demonstrated superior performance compared to large pre-trained language models such as BERT in African languages. Unlike BERTimbau [150], which was fine-tuned as a regression task, we fine-tuned AfriBerta as a classification task and used it to predict the sentiment class of individual words as positive, negative, or neutral. Table 4.6 shows the languages and examples of the lexicon created for each language.

The EcoLex and DeepEcoLex are both domain-specific lexicons, which is limited due to their number of entries, and we may need to extend them. In the next section, we discuss extending a domain-specific corpus by combining both the domain-specific lexicon and a general-purpose lexicon to improve performance.

Hausa		Igbo		Yoruba	
HausaLex	Translation	IgboLex	Translation	YoroLex	Translation
Qarya (neg)	Lie	igbu (neg)	Kill	aisododo (neg)	injustice
Wuya(neg)	Difficult	nsogbu (neg)	problem	abirun (neg)	disabled
Ciwo (neg)	Pain	Etigbu (neg)	hit	àisàn (neg)	illness
Matsala (neg)	Problem	ojo (neg)	bad	ainiṣ (neg)	unemployment
Kyauta (pos)	Gift	ekele (pos)	congratulations	kágbn (pos)	wise
soyayya(pos)	Love	oganiru (pos)	prosperous	itelorun (pos)	satisfaction

Table 4.6: DeepNaijaLex: Lexicon and their translation created using a pre-trained language model from the NaijaSenti Dataset.

4.1.4 Combining domain-specific and general purpose lexicon

One challenge in creating a domain-specific lexicon is the limited size of the domain-specific corpus used for lexicon generation. This often results in a lexicon of limited size that may not contain many entries from the general domain. To address this limitation, we employ a method to extend the domain-specific lexicon by combining both the domain-specific lexicon and the general-purpose lexicon [45, 147]. If a word appears in both lexicons, we give preference to the sentiment value in the domain-specific lexicon and include it in the final extended lexicon. We extended the domain-specific sentiment lexicon, EcoLex, by combining it with Sentilex' [45, 147], our adaptation of the general-purpose Portuguese sentiment lexicon. The resulting lexicon, which we refer to as *EcoSentiLex*, contains all entries from SentiLex that were missing in EcoLex. Table 4.7 shows the column "*EcoSentiLex*" with entries from both EcoLex and SentiLex.

Word	SV	SV	SV
	EcoLex	SentiLex	EcoSentiLex
competitivo (competitive)	1.33	-	1.33
adequado (adequate)	0.40	1	0.40
consolidação (consolidation)	0.58	-	0.58
alto (high)	-0.39	-	-0.39
ativo (active)	0.33	1	0.33
capital (capital)	-0.52	-	-0.52
ajustamento (adjustment)	-	-	-
adorar (to love)	-	-	-
abuso (abuse)	-	-	-
absurdo (absurd)	-	-1	-1
agressivo (aggressive)	-0.93	-1	-0.93

Table 4.7: Examples of combining domain-specific lexicon (EcoLex) and general purpose lexicon (Sentilex').

4.2 Using Lexicons for Sentiment Classification

In this section, we explain how the sentiment lexicons are used to generate predictions of sentiment values on new texts. This approach is simple and computationally efficient. It involves identifying sentiment-carrying words in the text and using a sentiment lexicon to assign a sentiment value to each word. These values are then summed to obtain the overall sentiment score of the text. This method is particularly useful for analyzing sentiment when computational resources are limited. For example, consider the following text: *"I love an iPhone. It is an excellent phone."* To determine the sentiment of this text using a lexicon-based approach, we search for and retrieve the sentiment values of each word in the text from a sentiment lexicon. In this case, let us assume that our lexicon contains sentiment values for only two words: *love* (score of 2) and *excellent* (score of 3). By summing the sentiment scores for each word in the text, we can calculate the overall sentiment score, which is 5 (2 + 3). If the aim were to determine the polarity, the output would be a categorical value (e.g., "1" representing positive polarity).

4.2.1 Evaluation the Quality of Generated Predictions

There are various measures that can be used to evaluate the quality of predictions, depending on whether the system generates numerical or categorical outputs. The mean absolute error (MAE) is often used for regression systems that produce numeric predictions. It is calculated as $MAE = |\widehat{SV}dj - SVdj|$, where $\widehat{SV}dj$ is the predicted sentiment value for the text dj and $SVdj$ is the true value. In this work, we focus on categorical predictions in terms of polarity. Common measures for this task include accuracy, precision, recall, and F1. As $F1$ is calculated for each class separately, the measures $microF1$, $macroF1$ and $wF1$ (weighted F1) combine the individual F1 measures [169]. Since our data is imbalanced, we use $wF1$ (weighted F1), which is calculated as follows:

$$wF1 = prop.pos * F1.pos + prop.neg * F1.neg \quad (4.3)$$

where $F1.pos$ ($F1.neg$) represents the $F1$ measure for the positive (negative) class and $prop.pos$ ($prop.neg$) the proportion of positive (negative) examples in the test set.

4.3 Experiments and Results on Different Datasets

4.3.1 Experiments with the Amazon IMDB dataset

Experimental setup To compare the performance of our generated lexicon with that of Almater et al. [11], we constructed lexicons using both approaches on a subset of the

Lexicon	Weighted-F1
IMDBLex	80.83
AlmatarLex	39.46

Table 4.8: Experiments using IMDBLex and AlmatarLex.

IMDB dataset. We used 10,000 reviews - 5,000 negatives and 5,000 positives -to generate the lexicon. The performance of the lexicon was evaluated using a lexicon-based sentiment analysis approach on a separate set of 2,000 reviews. Prior to creating the lexicon, we adjusted the rating scale from its original range of 1 to 10 to a new range of -3 to 3.

Experimental Results We measured the performance of the two lexicons generated—IMDBLex and SattamLex using weighted-F1. Table 4.8 shows the performance results. Our approach (IMDBLex) achieved a weighted-F1 score that was approximately twice as high as using AlmatarLex lexicon.

4.3.2 Experiments with the EcoSentiPT dataset

Experimental setup We use a form of cross-validation to divide the data into training and tests using N-fold (N=6). The data is partitioned into N folds, with one fold serving as the test set while the remaining folds are used for training. This process produced N pairs of training and test datasets and N separate evaluation results. However, our approach differs from that of classical cross-validation. The main difference is that we do not start with a single pool of data. Instead, we have six distinct sets of sentences and their corresponding annotations made by different annotators. As a result, the distribution of positive and negative cases varies across each set. For this reason, we prefer to avoid the term *fold j* and instead use the term *setup j*.

Setup	1	2	3	4	5	6	Mean
Train set (fractions)	1422	1452	1577	1570	1475	1246	1457.0
% of negative	51.0	49.9	48.3	48.1	48.7	53.4	49.9
Test set (fractions)	332	291	183	188	269	497	292.3
% of negative	44.6	49.1	62.3	63.8	55.8	40.4	52.7
Test set (sentences)	59	68	66	70	68	67	66.3
% of negative	66.1	75.0	69.7	67.1	85.3	35.8	66.5

Table 4.9: Numbers of elements in the train and test sets.

Table 4.9 provides further details about the six setups. The table shows the number of cases included and the proportion of negative and positive cases in each setup. For example, in setup 1, there are 1422 sentence fragments of which 51.0% are negative and 49% are positive.

We use these setups to automatically generate lexicons using both statistical and pre-trained language model approaches. All experiments were conducted on the same dataset split. The lexicon sizes used in different setups are shown in Table 4.10.

Setup	1	2	3	4	5	6	Mean
Sentilex'	5496	5496	5496	5496	5496	5496	5496
EcoLex	1250	1241	1283	1221	1240	1182	1227.5
EcoSentiLex	6518	6503	6542	6599	6515	6446	6503.8
DeepEcoLex	1461	1461	1461	1461	1461	1461	1461

Table 4.10: Lexicon size used in each setup

- **SentiLex'** Lexicon based on an off-the-shelf SentiPT
- **EcoLex** Lexicon generated using heuristic-based approach
- **EcoSentiLex** Extended Ecolex with combination of SentiLex.
- **DeepEcoLex** Lexicon generated using pre-train language model

Results for Sentence fractions Here, we generated lexicons from training data and evaluated the performance of short sentences. The results are shown in Table 4.11 where columns 1 to 6 show the performance for different setups. The mean value for each setup is shown in the last column. We used the abbreviation *SA* to denote sentiment analysis that uses a given lexicon for predictions. For example, SA-EcoLex is a sentiment analysis system that leverages the EcoLex lexicon.

Setup	1	2	3	4	5	6	Mean
SA-Sentilex'	51.1	47.0	66.4	60.5	57.6	46.3	55.0
SA-Ecolex	69.5	70.2	68.3	66.9	64.3	66.7	67.7
SA-EcoSentiLex	74.2	71.4	70.7	70.5	65.8	68.2	70.1
SA-DeepEcoLex	83.4	75.9	68.9	66.3	69.8	79.8	74.0

Table 4.11: Weighted F1-score results on sentence fractions

Among all generated lexicons, DeepEcoLex achieved the best results with a mean weighted F1 score of 73.7%. This indicates that using a pre-trained language model to generate the sentiment lexicon improves performance. The domain-specific EcoLex lexicon also outperformed the general-purpose SentiLex', demonstrating the usefulness of domain-specific lexicons. Furthermore, the extended EcoSentiLex lexicon, which combines domain-specific and general-purpose lexicons, performed better than using only the domain-specific lexicon with a 2.4% improvement. This suggests that combining domain-specific and general-purpose lexicons can be beneficial.

Results for Long Sentences In this set of experiments, we used the same sentiment lexicons as before, but we evaluated the performance on long sentences. The results are shown in Table 4.12. Columns 1 to 6 show the performance for different setups. The mean value for different setups is shown in the last column. The best result is obtained again with SentiLex’ with a mean weighted F1 of 66.2%. This value is substantially better (4.0% higher on average) than the performance of a similar EcoSenti-lex and EcoLex. Surprisingly, DeepEcoLex does not perform well on long sentences. This may be attributed to the fact that DeepEcoLex contains fewer lexicon entries (1461) than SentiLex(5496). We plan to investigate this further in future work. Overall, the lexicon-based method provides an explainable result which is discussed further in the next chapter in Section 5.5.4.

Setup	1	2	3	4	5	6	Mean
SA-SentiLex’	71.0	70.3	62.8	63.0	70.9	61.6	66.2
SA-EcoLex	53.5	68.1	62.0	66.7	58.9	64.9	62.3
SA-EcoSentiLex	55.0	64.3	64.7	68.1	58.1	63.3	62.3
SA-DeepEcolex	46.6	29.0	36.7	22.7	24.2	63.9	37.2

Table 4.12: Weighted F1-score results on Long sentence

4.4 Conclusions

In this chapter, we presented an approach for sentiment lexicon generations. Our main contributions are:

- **Automatic Lexicon Generation:** We proposed two new approaches for automatically generating sentiment lexicons and demonstrated their superior performance over an existing approach [21, 22].
- **NaijaLex and DeepNaijaLex:** For the first time, we created NaijaLex and DeepNaijaLex — manually and automatically generated sentiment lexicon for three major Nigerian languages (Hausa, Yoruba, and Igbo).

Chapter 5

Improving Lexicon-based Approach with Contextual Shifters

In this chapter, we present an improved lexicon-based method that leverages a sentiment lexicon along with a manually developed set of shifter patterns and rules. Shifter patterns capture intensification, attenuation/downtoning, and reversal/inversion strategies. Our experiments demonstrate that the use of shifter patterns improves the performance of our sentiment analysis system.

5.1 Shifters

We present an enhanced lexicon-based method for sentiment analysis that incorporates contextual information. Unlike previous methods that rely on unigrams and ignore the effect of sentiment shifters [163, 22, 163, 145], our enhanced approach leverages a sentiment lexicon along with a manually developed set of shifter patterns and rules. Sentiment shifters are words or expressions that can modify the polarity of another word or expression through *intensification*, *downtoning/attenuation* or *reversal/inversion* [163]. For example, the word *inverte* (inverts) is a shifter that negates the polarity of the phrase “*o crescimento de economia*” (growth of the economy), resulting in negative sentiment.

5.2 Shifter Pattern

We introduce two concepts describing the effect of shifters in sentiment analysis: the *head token* and the *dependent token*. Consider the phrase *crescer muito* in Portuguese, which translates as “grow very much.” In this phrase, *crescer* (grow) is identified as the head token, while *muito* (very much) is identified as the dependent token. This information is provided

by a dependency parser as shown in Table 5.1. This relationship can be visualized through a dependency relationship diagram, where the head token is positioned above the dependent token and connected by a link as shown in Figure 5.1.

S/N	tokenid	token	upos	head tokenid	deprel
1	1	crescer	V	0	root
1	2	muito	ADV	1	advmod
2	1	crescimento	N	2	nsubj
2	2	parece	V	0	root
2	3	bom	ADJ	2	xcomp

Table 5.1: Dependencies between tokens provided by the dependency parser

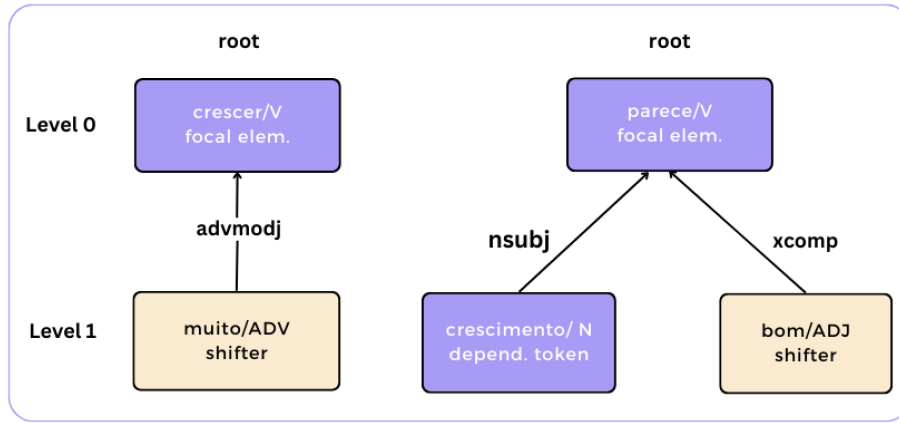


Figure 5.1: Two examples with dependency links between some tokens

A shifter pattern captures how a shifter modifies the polarity of another expression. For instance, the phrase *inverter crescimento* (invert growth) can be represented by a specific pattern that captures how the shifter *inverter* negates or reverses the polarity of *crescimento*. This approach allows us to incorporate contextual information into our sentiment analysis system and can be represented by a pattern below:

$$S^R \leftarrow F_N^+$$

where S^R denotes a reversal/inversion shifter (i.e., *inverter* (invert)) and F_N^+ denotes the focal element that is affected by the shifter (i.e., *crescimento*). The symbol “+” indicates the focal element should have a positive sentiment value. The symbol “N” indicates the syntactic category of the focal element (noun or noun phrase). The arrow “←” used in our example points to the *head token* from the dependent token. In this work, we use a dependency parser for the required language (here, Portuguese) that annotates all tokens in the given sentence

or phrase. Table 5.1 shows the output of the dependency parser for the phrase “*crescer muito* (growing a lot)” and in addition also the phrase “*crescimento parece bom* (the growth seems good)”. Figure 5.1 shows the dependency relationships in the form of a graph, where the head tokens are represented in a box positioned above the dependent token. The dependency relationship between the two is visible through the link connecting them.

For each shifter category (e.g., intensification), we introduce the manually curated shifter patterns in Portuguese by leveraging the syntactic class of the focal element (i.e., V, Adj, N). Table

5.3 Different types of shifter patterns

We manually analyze a corpus of text and identify a set of shifter patterns that can modify the sentiment polarity of words or expressions. We categorize these shifter patterns into three classes: intensification, downtoning, and reversal. In this section, we describe each class of shifter patterns and provide some examples.¹

5.3.1 Intensification

Intensification is an operation that increases the magnitude of the sentiment value associated with a focal element. For example, an intensifier can amplify a positive sentiment value, as in “*crescer muito*” (grow very much), or a negative sentiment value, as in “*piorar muito*” (deteriorate greatly). We formalize this operation as follows:

$$SV(F, S^I) = C_I * SV(F) \quad (5.1)$$

Where F is the focal element, S^I is an intensifier (e.g., *muito*), SV is the sentiment value function, and C_I is a constant greater than 1 (e.g., 2). Table 5.2 presents some examples of shifter patterns that involve intensifiers. A subscript indicates the class of the focal element. For example, F_V (or F_{ADJ} , etc.) means that the focal element should be a verb (or adjective, etc.) for the pattern to apply. The direction of the *head token* is shown by an arrow $\leftarrow$ (or $\rightarrow$). The shifter is shown in column *Token 1* or in *Token 2*, depending on its position relative to the focal element. For example, consider again *crescer muito* (grow much). Since *crescer* precedes *muito*, it is in *Token 1* column and *muito* is in *Token 2*. For readability, we enclose the focal element in parentheses and include it in our tables, although it is not part of the pattern. All patterns require that the focal element be of some specific class.

¹The shifter patterns were designed in collaboration with my supervisor, Prof. P. Brazdil and some linguists from the Faculdade de Letras da UP.

Id	shifter pat.	Token 1	Token 2	 S 	Use
IAV+	$F_V^+ \leftarrow S^I$	(F=crescer (grow))	S=muito (very much)	28	2.7
IAV-	$F_V^- \leftarrow S^I$	(F=prejudicar (harm))	S=muito (very much)	10	1.5
IAJ+	$S^I \rightarrow F_{ADJ}^+$	S=muito (very)	(F=satisfatório (satisfactory))	29	6.8
IAJ-	$S^I \rightarrow F_{ADJ}^-$	S=muito (very)	(F=negativo (negative))	11	4.3
IJN+	$S^I \rightarrow F_N^+$	S=bom (good)	(F=crescimento (growth))	23	9.5
IJN-	$S^I \rightarrow F_N^-$	S=mau (bad)	(F=crise (crisis))	10	3.3
IV'N+	$S^I \leftarrow F_N^+$	S=reforçar (reinforce)	(F=crescimento (growth))	20	2.3
IV'N-	$S^I \leftarrow F_N^-$	S=ampliar (amplify)	(F=crise (crisis))	15	1.8
IVN+	$F_N^+ \leftarrow S^I$	(F=crescimento (growth))	S=reforçado (reinforced)	20	0.8
IVN-	$F_N^- \leftarrow S^I$	(F=crise (crisis))	S=ampliado (amplified)	15	0.3
IN'N+	$S^I \leftarrow F_N^+$	S=aumento (increase of)	(F=crescimento (growth))	11	0.7
IN'N-	$S^I \leftarrow F_N^-$	S=aumento (increase of)	(F=despesa (expenditure))	8	0.8
INN+	$F_N^+ \leftarrow S^I$	(S=crescimento (growth))	F=aumentado (increased)	11	0.8
INN-	$F_N^- \leftarrow S^I$	(S=despesa (expenditure))	F=aumentado (increased)	8	1.0

Table 5.2: Some shifter patterns that include intensification

The column identified as $|S|$ shows the size of the list of words that represent the shifters for each shifter pattern. For example, consider the shifter pattern on the first line (i.e., IAV+) in Table 5.2. We see that the list of manually assembled shifters includes 28 elements (the tokens *muito* and *tanto* form part of this list). The column identified as *Use* shows how many times this pattern was invoked on our test set. As we use six different setups, the numbers shown represent the average values of the test sets across different setups.

We note that, in general, intensifier patterns were invoked quite frequently. For instance, intensifier pattern *IJN+* that covers *bom crescimento* (good growth), among other phrases, was invoked 9.5 times on average on different test sets across different setups.

5.3.2 Attenuation/downtoning

This operation works oppositely to intensification. It decreases the absolute value of the sentiment value of the focal element, as in *crescer pouco* (*grow slightly*). We formulate this operation as follows:

$$SV(F, S^A) = C_A * SV(F) \quad (5.2)$$

Where S^A represents the given attenuator (e.g., *pouco*), and C_A is a constant less than 1 (e.g., 0.5). Table 5.3 shows some shifter patterns with downtoners/attenuators that we have identified.

Id	shifter pat.	Token 1	Token 2	 S 	Use
AAV+	$F_V^+ \leftarrow S^A$	(F=crescer (grow))	S=pouco (not much)	5	0.8
AAV-	$F_V^- \leftarrow S^A$	(F=prejudicar (harm))	S=pouco (not much)	4	0.7
AAJ+	$S^A \rightarrow F_{ADJ}^+$	S=pouco (not much)	(F=satisfatório (satisfactory))	4	1.8
AAJ-	$S^A \rightarrow F_{ADJ}^-$	S=pouco (not much)	(F=negativo (negative))	4	0.8
AJN+	$S^A \rightarrow F_N^+$	S=fraco (weak)	(F=crescimento (growth))	12	1.2
AJN-	$S^A \rightarrow F_N^-$	S=fraca (weak)	(F=crise (crisis))	7	1.3
AV'N+	$S^A \leftarrow F_N^+$	S=diminuir (decrease)	(F=crescimento (growth))	8	0.7
AV'N-	$S^A \leftarrow F_N^-$	S=controlar (control)	(F=despesa (expenditure))	23	1.5
AVN+	$F_N^+ \leftarrow S^A$	(F=crescimento (growth))	S=diminuído (decreased)	8	0.7
AVN-	$F_N^- \leftarrow S^A$	(F=crise (crisis))	S=controlado (controled)	23	1.5
AN'N+	$S^A \leftarrow F_N^+$	S=diminuição (decrease of)	(F=crescimento (growth))	5	0.0
AN'N-	$S^A \leftarrow F_N^-$	S=diminuição (decrease of)	(F=despesa (expenditure))	5	0.2
ANN+	$F_N^+ \leftarrow S^I$	(S=crescimento (growth))	F=diminuído (decreased)	5	0.8
ANN-	$F_N^- \leftarrow S^I$	(S=despesa (expenditure))	F=diminuído (decreased)	5	1.0

Table 5.3: Some shifter patterns that include attenuation/downtoning

5.3.3 Reversal/inversion

The result of this operation depends on whether the focal element has a positive or negative sentiment polarity. For example, in the case of F^+ , reversal inverts the polarity of the sentiment value associated with the focal element. An example of this can be seen in *não é bom* (is not good), where most people would agree that the sentiment value conveyed by this phrase is predominantly negative. This can be defined by the following rule:

$$SV(F^+, S^R) = C_{R+} * SV(F^+) \quad (5.3)$$

Where S^R represents the given inversion shifter (e.g., *não*), and C_{R+} a constant that is negative (e.g., -0.8)². Using an inversion operation applied to a focal element with a negative sentiment value is more complicated. Consider, for instance, the expression *não é mau* (is not bad). Most people would not consider this equivalent to *bom* (good), but only slightly positive, close to the neutral value. So, this can be captured by the following rule:

$$SV(F^-, S^R) = C_{R-} * SV(F^-) \quad (5.4)$$

where and C_{R-} a negative constant near 0 (e.g., -0.2). As shifters also appear in the sentiment lexicon, they may be accompanied by a sentiment value. However, in reversal/inversion operations, we do not want this value to perturb the sentiment value of the whole phrase, and consequently, the value retrieved is multiplied by 0.1 to reduce its impact. Table 5.4 shows some shifter patterns with reversal/inversion that we have identified when analyzing

²We have conducted experiments with several different values and the value -0.8 was selected as it leads to the best results on a validation set

our corpus.

Id	shifter pat.	Token 1	Token 2	S	Use
RAV+	$S^R \rightarrow F_V^+$	S=não (not)	(F=crescer (grow))	12	0.7
RAV-	$S^R \rightarrow F_V^-$	S=não (not)	(F=prejudicar (harm))	12	6.2
RAJ+	$S^R \rightarrow F_{ADJ}^+$	S=não (not)	(F=satisfatório (satisfactory))	6	1.0
RAJ-	$S^R \rightarrow F_{ADJ}^-$	S=não (not)	(F=mau (bad))	6	1.0
RAN+	$S^R \rightarrow F_N^+$	S= não (no)	(F=crescimento (growth))	7	3.7
RAN-	$S^R \rightarrow F_N^-$	S= não (no)	(F=crise (crisis))	7	3.3
RV'N+	$S^R \leftarrow F_N^+$	S=inverter (reverse)	(F=crescimento (growth))	22	1.7
RV'N-	$S^R \leftarrow F_N^-$	S=inverter (reverse)	(F=crise (crisis))	13	0.7
RVN+	$F_N^+ \leftarrow S^R$	(F=crescimento (growth))	S=invertido (reversed)	22	0.3
RVN-	$F_N^- \leftarrow S^R$	(F=crise (crisis))	S=reduzida (reduced)	13	0.0
RN'N+	$S^R \leftarrow F_N^+$	S=inversão (reversal)	(F=crescimento (growth))	21	0.3
RN'N-	$S^R \leftarrow F_N^-$	S=redução (reduction)	(F=despesa (expenditure))	12	0.2
RNN+	$F_N^+ \leftarrow S^I$	(S=crescimento (growth))	F=reduzido (reduced)	5	0.8
RNN-	$F_N^- \leftarrow S^I$	(S=dívida (debt))	F=reduzido (reduced)	5	1.0
RVV+	$S^R \leftarrow F_V^-$	S=faltar (fail to)	(F=crescer (grow))	1	0.2

Table 5.4: Some shifter patterns that include reversal/inversion

5.4 Combining sentiment lexicons with shifter patterns

This section describes the methodology for using shifter patterns to determine sentiment values in a given text. The process starts with extracting the sentiment values of individual words (unigrams) using a sentiment lexicon and applying relevant shifter patterns to modify the values of focal elements. To illustrate this process with a simple example, *crescer muito*, we outline the following steps:

1. For each sentence, we generate a dependency tree. For instance, the dependency tree for our example “*crescer muito*” is shown in Fig. 5.1
2. For each dependency tree, we identify all the shifter words present within it. In other words, the system searches for words within each dependency tree that would modify the sentiment of other words and label them as shifter words. In this example, the system identifies only one shifter word: *muito*.
3. For each identified shifter word, we select an associated shifter pattern. In our example, the system would identify the definition $IAV+ : F_V^+ \leftarrow S^I$ shown in Table 5.2.
4. We then determine whether the shifter is the *head token* or *dependent token* and consider this information when searching for the focal element. We examine whether the focal element satisfies all requirements stipulated in the definition of the shifter pattern. For

instance, we check whether it belongs to the required class. Considering pattern *IAV+*, it should be a verb (V). Our previous example with “*crecer/V*” satisfies this condition.

5. Finally, for the focal element found, we retrieve its sentiment value from the appropriate sentiment lexicon and apply respective rules to determine its modified value within this phrase. For example, suppose that the sentiment value of the token *crecer* is 0.64. So, the sentiment value of the phrase *crecer muito* is $2 \times 0.642 = 1.284$. This value is used as the modified value of this phrase.

5.4.1 Explanations provided by the system

The explanations provided by our system include all words in the given text, while some of them are accompanied by additional information, such as sentiment values retrieved from the sentiment lexicon, or the values calculated as a result of using shifter patterns. For example, the explanation for the phrase ‘*crecer muito*’ is as follows:

crecer/*0.642/1.285/D *muito*/S/*IAV+*

As shown above, the explanation includes both *crecer* and *muito*, along with additional information such as:

- Sentiment values retrieved from the sentiment lexicon (here “*0.642”);
- Sentiment values that have been derived from the values in the sentiment lexicon (here “1.285”);
- Symbol “S” that accompanies each shifter identified (here it is associated with the token “muito”);
- Identifier of the shifter pattern(s) identified (here “IAV+”);
- Identification with symbol “D” all tokens that are in the scope of the shifter (here the token “*crecer*” is accompanied by “D”, as it is in the scope of the shifter “*muito*”).

5.4.2 Dealing with shifters that affect multiple words in their scope

A given shifter may affect more than one word. For instance, consider the phrase:

não gera receitas fiscais adequadas
(does not generate adequate fiscal income).

The negation shifter *não* (not) affects the phrase *gera receitas fiscais adequadas* (generates adequate fiscal income) in its scope. To identify the scope of the shifter, we use a dependency parser and identify the head token and all tokens that are linked to this shifter by dependence relations. For example, in Figure 5.2 (left), the shifter *não* affects the phrase *gera receitas fiscais adequadas* (generates adequate fiscal income), which is enclosed by an ellipse. We apply the inversion rule (RAV+) to this phrase and generate the following explanation:

The symbol “D” indicates that the tokens are dependent on the shifter’s head token. The inversion rule uses the values of “gera/*0.333/D” and “adequadas/*0.4/D” from the sentiment lexicon to compute the final value of the phrase (-0.627).³

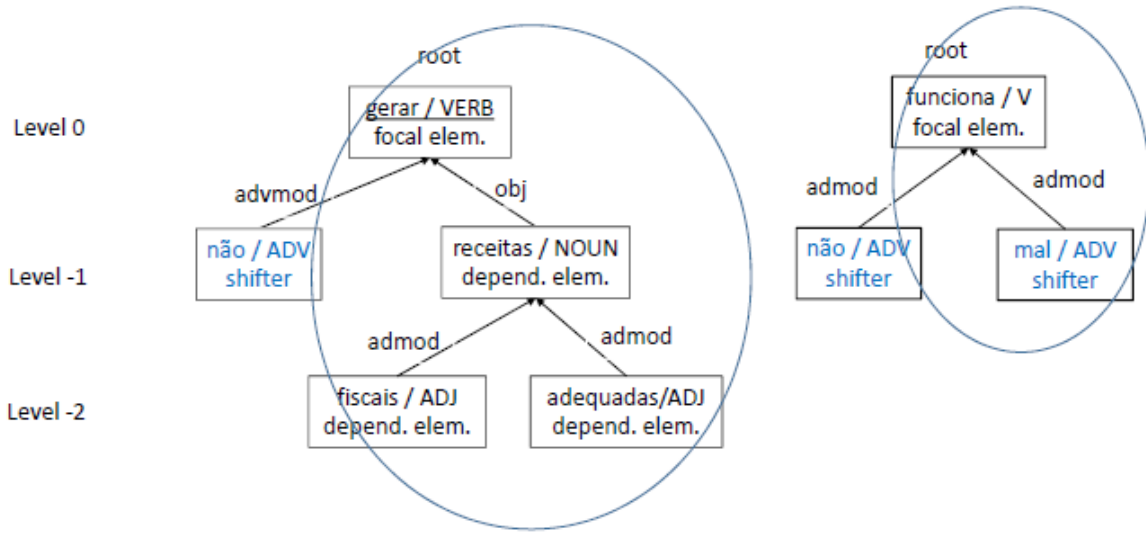


Figure 5.2: Example of tokens linked by dependencies

$$\text{não}/*-0.401/-0.627/\text{S}/\text{RAV+ gera}/*0.333/\text{D receitas}/\text{D fiscais}/\text{D adequadas}/*0.4/\text{D}$$

5.4.3 Processing phrases with multiple shifters

Some sentences may contain multiple shifters, where one is in the scope of another. In such cases, we need to determine the order of applying the shifters to the sentiment values. We adopt a method that conducts the search level by level, starting with the lowest level (here -1), and processes the tokens on each level from right to left. This method ensures that the innermost shifter is applied first, followed by the outermost shifter. For example, in Figure 5.2 (right), the phrase *não funciona mal* (does not work badly) has two shifters *não* and *mal*.

³The final value is obtained as a sum of two parts. The first includes the sentiment values in the focal phrase ($0.4+0.333$), multiplied by constant $C_{R+} = -0.8$, leading to -0.586 . The second one represents the reduced value of the shifter “nã”. This is achieved by $0.1 * -0.401 = -0.040$. The sum of the two parts is -0.627 .

The shifter *mal* is on level -1 and is first processed, followed by the shifter *não* on level 0. Our system generates the following explanation:

não/*-0.401/0.001/S/RAV- funciona/0.625/D mal/-0.286/D/RAV+

Two shifters have been applied to this text (first RAV+ and then RAV-). The final sentiment value of this phrase is slightly positive (0.001), which is correct. Another example is the phrase *inverte o crescimento de economia* (inverts the growth of the economy), which has two shifters *inverte* and *crescimento*. Our system predicts the value -1.151 as the final sentiment value and generates the following explanation:

inverter/*0.4/-1.151/M/RVN+ crescimento/*0.598/D/INN+ economia/*0.445/D

First, shifter pattern INN+ is applied to increase the positive value of the word *economia*. Then, shifter pattern RVN+ is applied to invert the polarity of the phrase *crescimento economia*.

5.5 Experimental Setup and Results

We conducted an experiment leveraging the joint use of shifter rules, shifter patterns, and sentiment lexicons. Subsequently, we compared the accuracy of our experimental results with those obtained by relying solely on sentiment lexicons. Additionally, we carried out two experiments using sentence fractions and sentences.

5.5.1 Lexicons used

In our experimental setup, we used four systems to predict sentiment values on sentence fragments. Two approaches to sentiment analysis were employed in conjunction with sentiment lexicons. The first approach, referred to as SA-EcoLex, used only sentiment lexicons. The second approach, SAP, used both lexicons and shifter patterns.

- SA-EcoLex - A symbolic system that exclusively employs the domain-specific sentiment lexicon, *EcoLex*.
- SAP-EcoLex - A symbolic system that employs symbolic shifter rules, shifter patterns, and the domain-specific sentiment lexicon, *EcoLex*.
- SA-EcoSentiLex - A symbolic system that employs the sentiment lexicon, *EcoSentiLex*, which combines the domain-specific sentiment lexicon, *EcoLex*, with the general-purpose sentiment lexicon, *SentiLex*'.

- SAP-EcoSentiLex - A symbolic system that integrates symbolic shifter rules with the sentiment lexicon, *EcoSentiLex*.

5.5.2 Experiments involving sentence fractions

Setup	1	2	3	4	5	6	Mean
SA-EcoLex	69.5	70.2	68.3	67.0	64.3	66.7	67.7
SAP-EcoLex	69.6	71.2	68.0	69.2	68.1	66.7	68.8
SA-EcoSentiLex	74.2	71.4	70.7	70.5	65.8	68.2	70.1
SAP-EcoSentiLex	75.4	73.1	71.0	71.8	70.8	68.6	71.8

Table 5.5: Weighted F1-score results on sentence fractions

Table 5.5 presents the results of the sentiment analysis systems for different setups using sentence fractions. The mean value across different setups is shown in the last column. SAP-EcoSentiLex achieved the best performance with a mean weighted F1 of 71.8%, indicating that using a combined lexicon that includes both domain-specific and general-purpose lexicons is advantageous. Furthermore, the use of shifter patterns appears to be beneficial, as the SAP-EcoSentiLex system achieves a higher weighted F1 value (+1.7) when compared to SA-EcoSentiLex, which is a similar system but does not use shifter patterns.

5.5.3 Experiments involving sentences

Setup	1	2	3	4	5	6	Mean
SA-Ecolex	53.5	68.1	62.1	66.7	58.9	64.9	62.3
SAP-Ecolex	62.1	61.5	58.1	70.1	71.4	60.5	64.0
SA-EcoSentilex	55.3	64.3	64.7	68.1	58.1	63.3	62.3
SAP-EcoSentiLex	63.7	60.2	62.8	70.8	71.0	57.6	64.3

Table 5.6: Weighted F1-score results on long sentences

Table 5.6 presents the results of the sentiment analysis systems for different setups using sentences. The best result was again obtained with SAP-EcoSentiLex, which achieved a mean weighted F1 of 64.3% in different setups. This value is substantially better (2.0% higher on average) than the performance of a similar SA-EcoSentiLex system that uses only sentiment lexicons. Thus, the positive effect of using shifter patterns is also confirmed in sentences and this result is comparable to that obtained with SAP-Ecolex. However, using the combined sentiment lexicon did not provide an advantage in this setting. Likely, the number of training fractions was insufficient to capture the relationships between the different parts of the sentences. We plan to investigate this issue in future work.

5.6 Limitations and Future work

We present limitations that arise when using the lexicon-based approach. These limitations include: (1) the absence of entries in the sentiment lexicon, (2) inconsistent labeling of certain sentence constituents, and (3) erroneous output from the dependency parser. We briefly elaborate on these limitations below:

Missing entries in the sentiment lexicon Our analysis revealed that some errors in sentiment prediction can be attributed to the absence of certain words in our sentiment lexicon. This affected the functioning of the shifter patterns. For example, when an intensifier pattern rule (IAJ+) is applied to the phrase *mais desfavorecido* (more disadvantaged), no impact is observed because ‘disadvantaged’ is not present in our sentiment lexicon and is therefore assigned a sentiment value of 0. Consequently, invoking the IAJ+ pattern and multiplying the sentiment value of 0 by a constant factor (e.g., 2) does not modify the sentiment value and returns a value of 0. One possible solution to this problem would be to generate a sentiment lexicon from a large corpus.

Problems of categorization of some shifter patterns Another challenge we encountered involves deciding whether a certain shifter should be included in attenuation/down-toning patterns or reversal/inversion patterns. For example, consider the phrase *reduzir a crise* (reduce the crisis). One could argue that the term *reduce* merely attenuates the negative sentiment value associated with the term *crisis*, implying that the crisis still persists but in a reduced form. However, one could also argue that the phrase indicates that the problem with the crisis was resolved. If this interpretation is used, then *crisis* forms part of a reversal/inversion pattern. The fact that some errors occurred with sentence fragments like these indicates that care is needed when designing a system. It is important to ensure that the labeling is consistent with the method used to derive sentiment values from a given shifter.

Incoherent labeling of sentences Some errors occur because certain sentences (or sentence fragments) are labeled in a way that is inconsistent with how given shifter patterns are interpreted. For example, consider the sentence *o objetivo é reduzir a dívida* (the aim is to reduce the debt). Some people might consider this sentence to have a positive sentiment value because reducing debt seems like a noble aim. This implies that some future state is considered when attributing sentiment value to this sentence. However, others could argue that the sentence implies that the country still has debt and therefore conveys a negative sentiment value. In this case, it implies that the current state is considered when attributing sentiment value. Thus, a problem arises if one interpretation is used when labeling sentence

fragments and another when applying shifter patterns. This decreases system performance. There are numerous examples similar to the one discussed above; some include modal verbs while others are paraphrases conveying more or less the same meaning:

- *devemos reduzir a dívida* (we should reduce the debt) (+),
- *podemos reduzir a dívida* (we can reduce the debt) (+),
- *é imprescindível reduzir a dívida* (it is necessary to reduce the debt) (+),
- *quem nos dera a dívida menor* (who could arrange us a smaller debt) (-).

In all these examples the dilemma arises regarding what should be the sentiment value. The first three phrases seem to have a positive sentiment value (+), as there seems to be a solution to the problem with the debt. This does not seem to be the case in the last example and so it is annotated with (-).

Dependency parser errors Finally, some errors may arise due to incorrect output from the dependency parser. For example, in the phrase *confunde X com Y* (confuses X for Y), the verb *confunde* was incorrectly lemmatized as *confunder*. As a result, if ‘confundir’ was present in the sentiment lexicon, it would not have been recognized. Furthermore, if a similar phrase to *confunde X com Y* appeared in the training data, an incorrect entry could have been introduced into the induced lexicon. Consequently, if the test data included the same word (albeit with incorrect spelling), our system would still recognize it and thus conceal the error.

Our findings highlight the challenges inherent in the symbolic approach and the need for further investigation and refinement. The presented method could be improved in various ways. Given that the training data was small (data consisting of labeled sentence fractions), one cannot expect the predictive performance to be very high. More training data would help the symbolic approach to construct a more comprehensive sentiment lexicon and thus lead to improved performance. Further improvements can be obtained by enhancing the set of shifter patterns and, in particular, the lists of words acting as shifters that are associated with each pattern. Another interesting research direction is to translate and adapt existing patterns to another language.

Recent approaches for sentiment analysis focus on deep learning. Deep learning models have the ability to learn from vast amounts of data and can construct their own sentiment representations, which are more richer than those of a lexicon-based approach. This line of research is discussed in the next chapter.

5.7 Conclusions

In this chapter, we presented an enhanced lexicon-based sentiment analysis approach that incorporates shifter patterns. These patterns capture contextual information and linguistic knowledge in an interpretable manner. Our experimental results demonstrate that the use of shifter patterns improves the accuracy of the lexicon-based approach. To encourage further research in this area, we have made our shifter patterns publicly available at <https://github.com/shmuhammad2004/shifter-patterns>. Main contribution from this chapter resulted in the following work:

- **Exploiting Shifter Pattern** Semi-Automatic Approaches for Exploiting Shifter Patterns in Domain-Specific Sentiment Analysis [22].

Chapter 6

Deep Learning for Sentiment Analysis

Traditional lexicon-based sentiment analysis methods have limited power in capturing the complexity of human language. In recent years, machine learning and deep learning techniques, particularly transformer-based pre-trained language models (PLMs) such as BERT [38] and RoBERTa [88] have demonstrated superior performance. In this chapter, we leverage pre-trained language models and transfer learning strategies to evaluate their performance on three datasets: EcoSentiPT, NaijaSenti and AfriSenti. Our results show that these methods outperform the enhanced lexicon-based approaches discussed in the previous chapter. We also present results from the first Afrocentric SemEval Shared Task on Sentiment Analysis for African Languages (AfriSenti-SemEval) we organized. ¹. Finally, we present a neuro-symbolic approach that leverages the strengths of both symbolic and deep learning approaches to achieve both performance and explanation.

6.1 Methodology

This section describes the pre-trained models used in our experiments and other transfer-learning strategies.

6.1.1 Pre-Trained Language Models

Pre-trained language models (PLMs) are widely used in natural language processing (NLP) to learn general language representations from large-scale corpora and then fine-tune them for specific downstream tasks [62]. This transfer learning technique can reduce the need to train a new model from scratch for each task and improve performance with limited annotated data

¹I conceived the idea for organizing this SemEval shared task and led its organization in collaboration with HausaNLP and MasaKhaneNLP researchers. See: <https://afrisenti-semeval.github.io/>

[133]. There are various PLMs available, including monolingual and multilingual variants. In this chapter, we use both types of models for sentiment classification on three datasets: EcoSentiPT, a sentiment dataset for finance/economics in European Portuguese; NaijaSenti, a Nigerian sentiment dataset; and AfriSenti, an African sentiment dataset. We compare the performance of six PLMs on these datasets and show that they outperform traditional lexicon-based methods. We briefly describe the six PLMs used in our experiments below.

mBERT is a multilingual variant of BERT pre-trained in 104 languages, including a Nigerian language—Yorùbá. mBERT was pre-trained using masked language modelling (MLM) and the next-sentence prediction task [38]. We fine-tune the **mBERT-base-cased** model with 172M parameters by adding a linear classification layer to the pre-trained transformer model.

XLM-R In contrast to mBERT—which is a multilingual variant of BERT, XLM-R [36] is a multilingual variant of RoBERTa pre-trained in 100 languages, including Hausa as the only Nigerian language. Unlike mBERT, XLM-R only uses Masked Language Modeling (MLM) during pre-training. We used an XLM-R-base with 270M model parameters for fine-tuning on the NaijaSenti corpus.

RemBERT scales up mBERT to a larger model size (559M) and decouples embeddings, which enables a larger output embedding size during pre-training, resulting in stronger pre-training and downstream performance [34]. RemBERT supports three major Nigerian languages but does not support Nigerian Pidgin.

AfriBERTa AfriBERTa is a RoBERTa-style model trained in 11 African languages, including the four Nigerian languages in the NaijaSenti dataset [113]. The model was trained on less than 1GB of data (since most African languages are low-resourced). We use AfriBERTa-large with 126M parameters. Despite its small model size and limited pre-training data, AfriBERTa has been shown to perform competitively on an African NER dataset [65].

mDeBERTaV3 mDeBERTaV3 [57] is a pre-trained model that employs an ELECTRA-style approach [35], where a discriminator is trained to detect replaced tokens instead of predicting masked tokens. Unlike the other four models pre-trained on the Masked Language Modeling (MLM) task, mDeBERTaV3 does not support any Nigerian languages. In our experiments, we use the mDeBERTaV3-base model with 276 million parameters, similar in size to the XLM-R-base.

BERTimbau In contrast to the models mentioned previously, BERTimbau [150] is a monolingual model trained on the largest and most diverse open Portuguese corpus using a BERT-style approach. The pre-training corpus for BERTimbau consists of 2.68 billion tokens from 3.53 million documents, making it one of the largest Portuguese language models currently available.

6.1.2 Fine-tuning Language Model

Fine-tuning is a common transfer learning technique to adapt a pre-trained language model (PLM) to a specific task, domain, or language by updating its parameters on task-specific, domain-specific, or language-specific data [62, 127]. Today, some models are so large that they can achieve good performance without changing any parameters. This mode of operation is commonly called zero-shot learning [23].

In classical transfer learning, a model is first trained using language modelling objective functions such as masked language modelling (MLM) and causal language modelling (CLM) on large-scale unlabelled data, usually obtained from web crawling. This process is called pretraining and its main objective is to teach the models the fundamentals of language structure. After pre-training, the pre-trained models can be adapted or fine-tuned (i.e., its parameters are updated) to specific downstream NLP tasks such as classification [164] as shown in Figure 6.1.

Despite the performance of pre-trained language models, they may still fail to perform well on data substantially different from pre-training data, even after fine-tuning [59]. Thus, a new approach called adaptive pre-training has been shown to achieve good performance compared to fine-tuning alone [53]. Adaptive pre-training is sometimes referred to as adaptive fine-tuning or continual pre-training.

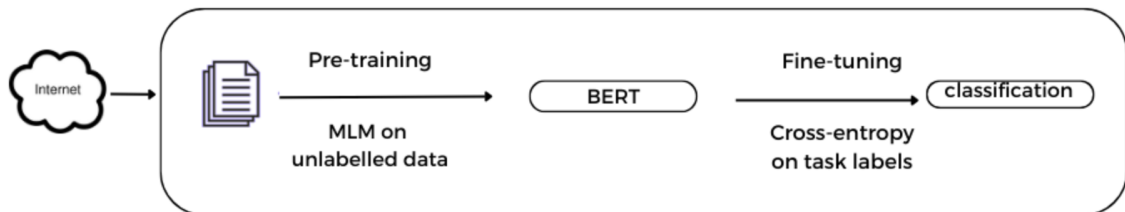


Figure 6.1: Classical transfer learning with pre-training and fine-tuning language model

Adaptive pre-training/fine-tuning is a technique used to improve the performance of pre-trained language models on new data that differs significantly from the pre-training data. This technique involves continuing the pre-training of the initial pre-trained model on more relevant new data with the same objective as in pre-training before fine-tuning it. The

advantage of this technique is that it only requires unlabelled data [141]. There are three types of adaptive pre-training/adaptive fine-tuning [53]:

- **Language Adaptive Pre-training (LAPT):** LAPT is a technique to continue pre-training a language model on a large corpus of text from a different language than the one used for the original pre-training. LAPT can improve the performance of a PLM for various NLP tasks in the target language, such as named entity recognition (NER) [9, 106, 65]
- **Task Adaptive Pre-training (TAPT):** TAPT is a technique to continue the pre-training of a language model on a smaller but more relevant corpus of unlabeled text from the target task, such as the task training data.
- **Domain Adaptive Pre-training (DAPT):** DAPT is a technique to continue the pre-training of a language model on a large corpus of unlabeled text from a specific domain, such as biomedical or news.

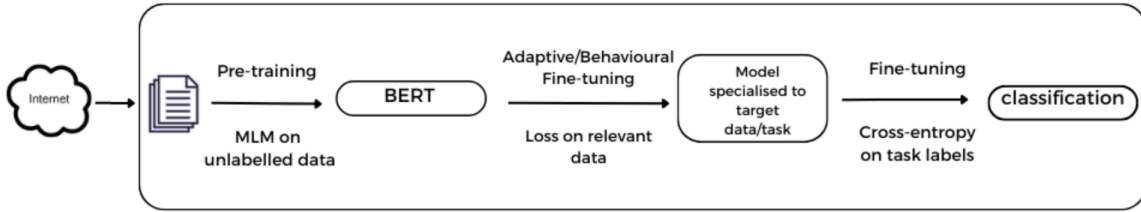


Figure 6.2: Adaptive or continual pre-training and fine-tuning of language models for improved performance on new data.

We report the results of the best system in the AfriSenti-SemEval shared task that explored both task adaptive pre-training (TAPT) and domain adaptive pre-training (DAPT). In our experiment, we only used language adaptive pre-training (LAPT).

6.2 Experiments with EcoSentiPT Dataset

6.2.1 Experimental Setup

We fine-tuned our models using the HuggingFace transformer toolkit [175] with a batch size of 32, a maximum sequence length of 128, 20 epochs and a default learning rate of 5e-5. All experiments were carried out with Nvidia V100 and RTX 2080 GPUs. We used the same setup as with SAP-EcoSentiLex as described in Section 5.5.

6.2.2 Experimental Results

We fine-tuned two pre-trained models, mBert and BERTimbau, for sentiment classification on the EcoSentiPT dataset. We have conducted experiments on sentence fractions and long sentences. The results of the sentence fractions are shown in Table 6.1. The performance of the extended lexicon-based approach SAP-EcoSentiLex is also included for comparison. Both pre-trained language models outperformed the lexicon-based approach, with weighted F1-scores of 91.0% and 75.9%, respectively, compared to 71.8%. BERTimbau achieved a higher score than mBert because it was pre-trained on Portuguese data only, while mBert was a multilingual model with limited transferability. This confirms that domain-specific models improve performance.

	Setup	1	2	3	4	5	6	Mean
1	SAP-EcoSentiLex	75.4	73.1	71.0	71.8	70.8	68.60	71.8
2	mBert	80.4	76.5	72.5	73.9	77.6	74.2	75.9
3	BERTimbau	92.2	88.7	84.3	99.9	93.5	87.5	91.0

Table 6.1: Weighted F1-score results on sentence fractions

The results on long sentences are shown in Table 6.2. We can see that the lexicon-based approach with shifters (SAP-EconSentiLex) achieved 64.3% weighted F1-score on long sentences, which is only 2.5% lower than the deep learning-based approach using mBERT. We did not experiment with BERTTimbau in this setting. This suggests that using shifter patterns captures a great deal of contextual information and thus can compete well with the deep learning approach. However, we note that our experiment involved small training data, which can result in poor performance with the deep learning approach. Moreover, the symbolic approach can provide useful explanations for the results of the sentiment analysis. Therefore, we believe that this direction is worth further investigation.

	Setup	1	2	3	4	5	6	Mean
1	SAP-EcoSentiLex	63.70	60.20	62.80	70.77	71.04	57.58	64.34
2	mBert	76.50	75.80	72.80	47.80	80.10	47.80	66.80

Table 6.2: Weighted F1-score results on long sentences

6.3 Experiments with NaijaSenti Dataset ²

6.3.1 Experimental Setup

Similarly, for the NaijaSenti dataset, we use the same hyperparameter settings and fine-tuning using the HuggingFace transformer tool [175] with a batch size of 32, a maximum sequence length of 128, number of epochs of 20, and default learning rate 5e-5 for all models except for XLM-R and RemBERT where we set the learning rate to 2e-5 to ensure model convergence. All the experiments were performed with Nvidia V100 and RTX 2080 GPUs. We experiment with the following transfer strategies on the NaijaSenti dataset:

- Multilingual Pre-trained Language Models with Fine-Tuning (PLMs),
- Multilingual PLMs with Language Adaptive Pre-training (LAPT) and Fine-Tuning,
- Multi-task Multilingual PLMs,
- Zero-Shot cross-lingual transfer strategy.

In the following section, we briefly explain these approaches.

Multilingual Pre-trained Language Model (PLMs) and Fine-tuning We fine-tuned five multilingual PLMs on the NaijaSenti dataset as shown in Table 6.3. The dataset contains four languages: Hausa (hau), Igbo (ibo), Nigerian-Pidgin (pcm), and Yoruba (yor). The results are quite similar in most cases with about a 1 – 3% performance difference. Performance may depend on the language being seen during pre-training. For example, mBERT has slightly better performance (+0.7%) for *yor* than XLM-R probably because *yor* was seen during pre-training. Similarly, XLM-R performs better for *hau*. RemBERT achieves slightly better performance than mBERT and XLM-R-base, demonstrating that a model with more capacity can improve performance.

Surprisingly, we found that mDeBERTaV3, which has only seen *hau*, provides better results (77.8%) than other models except AfriBERTa, which has been exclusively trained on African languages. mDeBERTaV3 uses replaced token detection (RTD), which has been shown to give superior performance in English [35].

Overall, we found AfriBERTa to be the best baseline model for *hau*, *ibo*, and *yor* because the model is pre-trained on only African languages. The main advantage of AfriBERTa is its smaller model size, which makes it easier to deploy, especially on the African continent,

²A part of this work was developed as a part of a collaborative project which included various researchers from various African countries, where the author of this dissertation led the project.

where most research labs cannot afford powerful GPUs. Unfortunately, AfriBERTa performs poorly on pcm because the model was not pre-trained on English, Nigerian-Pidgin’s lexifier language.

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	81.0 $\pm$ 0.2	81.2$\pm$0.5	70.9 $\pm$ 0.7	80.2 $\pm$ 0.6	78.3$\pm$0.3
mBERT-base	yor	172M	77.8 $\pm$ 0.5	79.8 $\pm$ 0.5	69.0 $\pm$ 0.2	77.6 $\pm$ 0.9	76.9 $\pm$ 0.3
XLm-R-base	hau	270M	78.4 $\pm$ 1.0	79.9 $\pm$ 0.7	73.3$\pm$0.3	76.9 $\pm$ 0.4	77.1 $\pm$ 0.1
mDeBERTaV3-base	hau	276M	79.3 $\pm$ 0.1	80.7 $\pm$ 0.2	72.5 $\pm$ 1.0	78.4 $\pm$ 0.5	77.8 $\pm$ 0.3
RemBERT	hau, ibo, yor	559M	79.0 $\pm$ 0.7	79.9 $\pm$ 0.4	73.3$\pm$1.4	78.0 $\pm$ 0.6	77.5 $\pm$ 0.2

Table 6.3: Weighted F1 evaluation of different Multilingual PLMs. Average and standard deviation over 5 runs. Numbers with “*” are within the standard deviation of the best model.

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multilingual PLMs+LAPT							
mBERT+LAPT (General)	hau / ibo / pcm / yor	172M	80.8 $\pm$ 0.3	80.4 $\pm$ 0.4	70.4 $\pm$ 0.5	80.8 $\pm$ 0.5	78.1 $\pm$ 0.3
mBERT+LAPT (Tweet)	hau / ibo / pcm / yor	172M	79.3 $\pm$ 0.6	77.7 $\pm$ 0.6	70.7 $\pm$ 0.7	76.8 $\pm$ 0.3	76.1 $\pm$ 0.2
XLm-R-base+LAPT (General)	hau / ibo / pcm / yor	270M	81.5$\pm$0.7	80.8 $\pm$ 0.8	70.0 $\pm$ 1.1	80.9$\pm$0.4	78.3$\pm$0.4
XLm-R-base+LAPT (Tweet)	hau / ibo / pcm / yor	270M	79.5 $\pm$ 0.9	77.0 $\pm$ 0.5	71.1 $\pm$ 1.3	76.2 $\pm$ 0.4	75.9 $\pm$ 0.2

Table 6.4: Weighted F1 evaluation of different Models using LAPT. The LAPT is trained on either general corpus or Twitter.

Multilingual PLMs with Language Adaptive Pre-training (LAPT) and Fine-Tuning Many African languages are low-resourced and understudied and thus have limited support from multilingual PLMs. For example, mBERT only supports three African languages (Malagasy, Swahili, and Yorùbá). As shown in Table 6.3, only AfriBERTa supports all languages in our experiment. To adapt the existing PLM to the languages in NaijaSenti, we use language adaptive pre-training (LAPT) [127], which continues pre-training a PLM on unlabeled texts in the target language using the same pre-training objective. We experiment with two collections of monolingual data: (1) Twitter domain, which is often very small (less than 50K tweets for Igbo and Yorùbá, and less than 600K tweets for Hausa and Nigerian-Pidgin), and (2) General domain, which is mostly based on Common Crawl corpora, religious texts, and online news. We also perform vocabulary augmentation using 99 most frequent words inspired by [32, 128] before further pre-training the PLM.

Table 6.4 shows some improvement over mBERT and XLm-R when we apply LAPT on the general domain, on average 2 – 3% on hau, and yor, and < 1% on ibo. For pcm, we only identified an improvement for mBERT (+1.2%). Interestingly, applying LAPT on the Twitter domain did not improve performance. The main reason for this is the small size of the Twitter data. For example, hau was further pre-trained on CC100 [36] corpus with over 318MB and 3 million sentences for the general domain, but for Twitter, we only have

around 512K tweets (32MB), which are often short. In general, we found AfriBERTa to be competitive or better than LAPT for the Nigerian languages except for pcm.

Multi-task Multilingual PLMs In addition to fine-tuning separate models for each language, we trained a joint multi-task sentiment classification model on the four Nigerian languages by aggregating their training sets. The major advantage of this is that having a single model that can classify the sentiment in tweets in all major Nigerian languages facilitates deployment for practical applications. Knowledge from related languages may also be beneficial during transfer. This setting is possible because we are using multilingual PLMs that support multiple languages.

In this setting, we only trained on the best two multilingual PLMs: AfriBERTa and mDeBERTaV3 as shown in Table 6.5. We observe only a slight drop in performance with mDeBERTa (-0.3%) while the AfriBERTa performance is the same. This indicates that we could easily deploy a single sentiment classification model for the four major Nigerian languages, instead of multiple monolingual models.

Model	NG lang. supported	PLM size	hau	ibo	pcm	yor	Avg
Multi-task Multilingual PLMs							
AfriBERTa-large	hau, ibo, pcm, yor	126M	81.2 $\pm$ 0.1	80.6 $\pm$ 0.3	70.9 $\pm$ 0.8	80.5 $\pm$ 0.5	78.3$\pm$0.3
mDeBERTaV3-base	hau	276M	79.4 $\pm$ 0.4	79.6 $\pm$ 0.2	72.7 $\pm$ 0.4	78.4 $\pm$ 0.2	77.5 $\pm$ 0.1

Table 6.5: Weighted F1 evaluation of different Models using Multi-task.

Zero-shot Cross-Lingual Transfer Lastly, we evaluate the zero-shot performance of a sentiment classifier trained on English tweets from SemEval-2017 Task 4 [139] on each of the four Nigerian languages and Table 6.6 shows the results. The English SemEval corpus consists of 11,763 tweets in the training set. pcm has the best zero-shot performance across all models due to its linguistic similarity to English, its lexifier language. AfriBERTa gave the best overall result in zero shot transfer, and it is significantly better than the majority classifier (weighted average) for all languages: hau, ibo, pcm, and yor are better by 41.8%, 20.8%, 4%, and 19.1% respectively.

Model	hau	ibo	pcm	yor	Avg
AfriBERTa-large	58.4	47.7	50.5	43.1	49.9
mBERT-base	31.0	37.0	50.8	39.5	39.6
XLM-R-base	38.4	37.8	56.3	26.7	39.8
mDeBERTaV3-base	50.1	47.2	57.7	36.4	47.9
RemBERT	54.0	45.4	55.9	30.2	46.4

Table 6.6: Transfer learning experiments. PLMs are trained on English SemEval 2017 and evaluated on Nigerian languages in a zero-shot setting

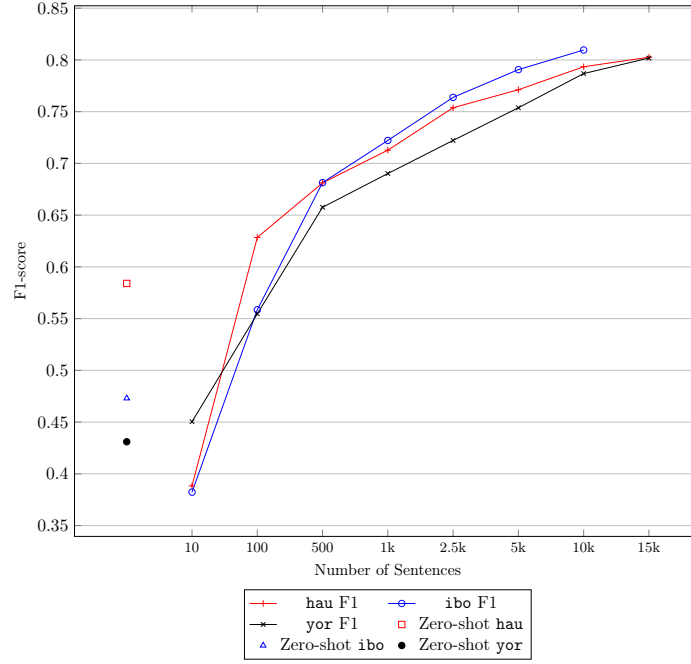


Figure 6.3: Sample efficiency on hau, ibo, and yor using the AfriBERTa model.

Sample Efficiency in Transfer We also investigate how many tweets from each of the Nigerian languages are needed to achieve the zero-shot performance of a model transferred from English and to outperform a majority classifier. Figure 6.3 shows the result of fine-tuning AfriBERTa on hau, ibo, and yor datasets of different sizes (10, 100, 500, 1K, 2.5K, 5K, 10K, and 15K). We observe that with only 10 examples, we obtain an F1 score of 38–40%, which already exceeds the majority vote performance in Table 6.7. Surprisingly, with only 100 sentences, we surpass the zero-shot transfer performance from English, and with at least 1000 sentences, we reach a decent performance of 70% F1. This result demonstrates that we can leverage a multitask sentiment classification model trained on Nigerian languages to quickly adapt to other African languages with as few as 100 or 1000 annotated samples. However, there is still room for improvement, especially in the zero-shot and few-shot cross-lingual transfer settings.

6.4 Experiments with AfriSenti Dataset³

The AfriSenti dataset was previously introduced in Section 3.4. It is a sentiment analysis dataset that covers 14 African languages: Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga and Yoruba [159]. In the following section, we present a baseline experiment using this dataset.

³A part of this work was developed as a part of a collaborative project which included various researchers from various African countries, where the author of this dissertation led the project.

6.4.1 Experimental Setup

We fine-tune two types of multilingual pre-trained language models (PLMs) for 12 AfriSenti languages with training data: massively multilingual PLMs trained on 100 languages from around the world and Africa-centric PLMs trained exclusively on languages spoken in Africa. For the massively multilingual PLMs, we use XLM-R-base and large [36] and mDeBERTaV3 [58]. For the Africa-centric PLMs, we use AfriBERTa-large [113] and AfroXLMR-base large [10], which are XLM-R models adapted to African languages. AfriBERTa was pre-trained from scratch on 11 African languages, including nine of the AfriSenti languages, while AfroXLMR supports 10 of the AfriSenti languages. Additionally, we fine-tune XLM-T [16], an XLM-R model adapted to the multilingual Twitter domain, which supports more than 30 languages but fewer African languages due to the lack of coverage by Twitter’s language API.

6.4.2 Experimental Results

The results of the monolingual baseline models on AfriSenti with different Multilingual PLMs are shown in Table Table 6.7. From the result, AfriBERTa obtained the worst performance on average (61.7), especially for languages it was not pre-trained on, e.g., < 50 for the Arabic dialects. However, it achieved a good result for languages it has been pre-trained on, such as *hau*, *ibo*, *swa*, *yor*. On average, XLM-R-base led to performance comparable to AfriBERTa, worse for most African languages, but better for Arabic dialects and *pt-MZ*.

Lang.	AfriBERTa large	XLM-R base	AfroXLMR base	mDeBERTa base	XLM-T base	XLM-R large	AfroXLMR large
Amharic(amh)	56.9	60.2	54.9	57.6	60.8	61.8	61.6
Algerian Arabic(arq)	47.7	65.9	65.5	65.7	69.5	63.9	68.3
Moroccan Arabic(ary)	44.1	50.9	52.4	55.0	58.3	57.7	56.6
Hausa(hau)	78.7	73.2	77.2	75.7	73.3	75.7	80.7
Igbo(ibo)	78.6	75.6	76.3	77.5	76.1	76.5	79.5
Kinyarwanda(kin)	62.7	56.7	67.2	65.5	59.0	55.7	70.6
Nigerian-Pidgin(pcm)	62.3	63.8	67.6	66.2	66.6	67.2	68.7
Mozambique Portuguess(pt-MZ)	58.3	70.1	66.6	68.6	71.3	71.6	71.6
Swahili(swa)	61.5	57.8	60.8	59.5	58.4	61.4	63.4
Xithonga (tso)	51.6	47.4	45.9	47.4	53.8	43.7	47.3
Twi(twi)	65.2	61.4	62.6	63.8	65.1	59.9	64.3
Yoruba(yor)	72.9	62.7	70.0	68.4	64.2	62.4	74.1
AVG	61.7	61.9	63.9	64.2	64.7	63.1	67.2

Table 6.7: Accuracy scores of monolingual baselines for AfriSenti on the 12 languages with training splits. Results are averaged over 5 runs.

On the other hand, AfroXLMR-base and mDeBERTaV3 achieve similar performance, although AfroXLMR-base performs slightly better for *kin* and *pcm* compared to other models. Overall, considering models with up to 270M parameters, XLM-T achieves the best performance, which highlights the importance of domain-specific pre-training. XLM-T performs

particularly well on Arabic and Portuguese dialects, i.e., **arq**, **ary** and **pt-MZ**, where it outperforms AfriBERTa by 21.8, 14.2, and 13.0 and AfroXLMR-base by 4.0, 5.9, and 4.7 F1 points respectively.

AfroXLMR-large achieves the best overall performance and improves over XLM-T by 2.5 F1 points, which highlights the benefit of scaling for large PLMs. Scaling is of limited use for XLM-R-large, however, as it has not been pre-trained on many of the African languages. Overall, our results demonstrate the importance of both language and domain-specific pre-training as well as the benefits of scale for appropriately pre-trained models.

6.5 AfriSenti-SemEval

We organized SemEval shared task ⁴ using the AfriSenti dataset—Sentiment Analysis for African Languages (AfriSenti-SemEval). AfriSenti-SemEval is the first Africentric SemEval shared task and it is a sentiment classification challenge in 14 African languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga and Yoruba). The task involves classifying data into three categories: positive, negative, and neutral. A total of 44 teams participated in the shared task and 29 submitted system description papers. Task A is monolingual, where the goal is to determine the polarity (positive, negative, or neutral) of tweets in each language separately. This subtask has 12 tracks, one for each language except Oromo and Tigrinya. Task B is multilingual, where the training data consists of all 12 languages in Task A, and the test data is in a different target language. Task C is zero-shot, where the training data can be any or all available languages and the test data is in one of the two unseen languages (Tigrinya or Oromo). We briefly discuss the strategies of the best team in each of the three subtasks.

NLNDE is a team from Bosch Center for Artificial Intelligence. They used language adaptive pre-training (LAPT) and task adaptive pre-training (TAPT) on AfroXLMR-large to achieve the best performance in Task A and Task B. LAPT continued pre-training a pre-trained language model (PLM) on a monolingual corpus of a target African language from the Leipzig Corpus Collection [49], while TAPT continued pre-training on the AfriSenti training corpus of a target language. They showed that LAPT followed by TAPT improved significantly over fine-tuning AfroXLMR-large directly. NLNDE ranked first in 7 out of 12 languages, and, on average, the first in sub-task A and B. Their result outperformed our baseline result on AfriSenti dataset, which only fine-tuned AfroXLMR-large. This showed a significant performance increase by leveraging adaptive-fine tuning.

⁴I initiated and led the SemEval project in collaboration with researchers from various African countries. See :<https://afrisenti-semeval.github.io/>

UCAS-IIE-NLP is a team from the University of Chinese Academy of Sciences. The team used a sentiment lexicon to build a lexicon-based multilingual transformer model based on AfroXLMR-base to facilitate language adaptation and sentiment-aware representation learning. Additionally, they applied a supervised adversarial contrastive learning strategy to improve sentiment-spread representations and enhance model generalization. This demonstrates the usefulness of leveraging sentiment lexicons as side information in building language models. In the next section, we present a neuro-symbolic research approach to sentiment analysis that leverages the advantages of both deep learning and sentiment lexicon.

6.6 A Neuro-Symbolic Approach to Sentiment Analysis

One limitation of deep learning is its inability to explain how sentiment values are derived. In contrast, lexicon-based sentiment analysis systems can offer such explanations but generally have lower performance than the deep learning approach. Consequently, it is worth exploring whether these two approaches can be combined in a symbolic approach to leverage their respective strengths, and consequently achieve both performance and explanation [22, 73, 168]. Recent studies in neuro-symbolic research [60, 20] have shown interest in combining symbolic and deep learning approaches. Symbolic approaches refer to methods that rely on high-level symbolic (human-readable) representations of problems and their solutions, using, for instance, rules or similar constructs.

Neuro-symbolic approaches to sentiment analysis can be divided into two main groups. The first and most widely used approach involves works that aim to improve the performance of the deep learning approach by providing external additional information in a symbolic form, such as sentiment lexicons. The other group involves works whose aim is to provide explanations regarding how the sentiment predictions are obtained while leveraging deep learning approaches to attain better performance. Our approach falls into the latter group.

6.6.1 Methodology

Our neuro-symbolic approach consists of two steps, which follow the methodology described in the previous chapter: (1) lexicon generation using deep learning, and (2) enhanced lexicon-based approach using symbolic rules. The first step uses transformer-based pre-trained language models (BERTibau) to generate a sentiment lexicon that outperforms proportion/statistical-based approaches. The second step uses an enhanced lexicon-based approach incorporating shifter patterns to generate sentiment predictions for new texts. The architecture of our proposed system is illustrated in Figure 6.4.

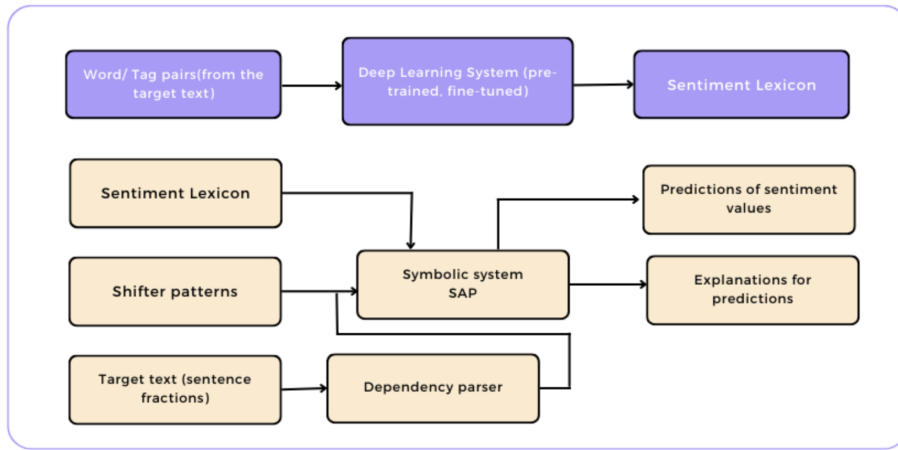


Figure 6.4: Architecture of the proposed neuro-symbolic approach to sentiment analysis

6.6.2 Experiments

This section describes the experimental setup for our neuro-symbolic approach, which consists of the lexicon generation using pre-trained language models and an enhanced lexicon-based approach with shifter patterns. We use the same setup with BERTimbau as in Section 5.5 for the enhanced lexicon-based approach and the same setup as in Section 4.1.3 for the lexicon generation using pre-trained language models. We use the following systems for evaluating our system performance:

- SAP-EcoSentiLex: A symbolic sentiment analysis system that uses symbolic shifter rules in conjunction with the sentiment lexicon *EcoSentiLex*
- SAP-DeepEcoLex: A neuro-symbolic sentiment analysis system that uses symbolic shifter rules in conjunction with the sentiment lexicon *DeepEcoLex*.
- BERTimbau: A deep learning system using BERTimbau.

	Setup	Type	1	2	3	4	5	6	Mean
1	SAP-EcoSentiLex	S	75.4	73.1	71.0	71.0	70.4	68.6	71.6
2	SAP-DeepEcoLex	N-S	83.1	83.4	73.1	71.3	75.8	82.8	78.3
3	BERTimbau	N	92.2	88.7	84.3	99.9	93.5	87.5	91.0

Table 6.8: Weighted F1-score results on sentence fractions

We compare our neuro-symbolic system (N-S), *SAP-DeepEcoLex*, with the symbolic system only (S), *SAP-EcoSentiLex*, and the deep learning system, BERTimbau (N), on sentence fractions and long sentences as shown in Table 6.8 and Table 6.9 respectively. SAP-DeepEcoLex

	Setup	Type	1	2	3	4	5	6	Mean
1	SAP-EcoSentiLex	S	63.7	60.2	62.8	70.8	71.0	57.6	64.3
2	SAP-DeepEcoLex	N-S	46.6	29.0	36.7	22.7	24.2	63.9	37.2
3	BERTimbau	N	76.5	75.8	80.3	47.8	80.1	47.8	66.8

Table 6.9: Weighted F1-score results on long sentences

outperforms EcoSentiLex by $\sim 6\%$ on sentence fraction, showing the benefit of using a lexicon from a pre-trained language model. However, SAP-DeepEcoLex still falls behind BERTimbau, which achieves 91% F1 score. This gap may be due to various shortcomings, including the ones discussed in the previous chapter (Section 5.6). On long sentences, the performance of all the neuro-symbolic systems drops drastically compared to the results on sentence fractions. This may be because the sentences often include rather complex relationships, which were not detected by the systems used here.

6.7 Conclusions and Future Work

This chapter has examined the performance of pre-trained language models (PLMs) for sentiment analysis on three datasets: EcoSentiPT, NaijaSenti, and AfriSenti. We showed that language-specific PLMs, such as BERTimbau [150] for Brazilian Portuguese and AfriBERTa [113] for African languages, achieved better results than massively multilingual PLMs like mBERT. This indicates that PLMs benefit from pre-training on a specific language rather than a large number of languages. However, AfriBERTa performed poorly on Nigerian-Pidgin, which is heavily influenced by English, because AfriBERTa was not pre-trained on English. To address this issue, we applied language adaptive pre-training (LAPT) on NaijaSenti data using XLM-R as the base pre-trained model. LAPT is a technique to fine-tune a PLM on a large corpus of text from a different language than the one used for the original pre-training. We used domain-related Twitter data and a general corpus in this process. LAPT did not improve the performance of XLM-R significantly on Twitter data because the Twitter data was rather small. However, when we used LAPT with a much larger general corpus, we obtained an enhanced performance of XLM-R and made it competitive with AfriBERTa. This demonstrates the effectiveness of adaptive pre-training on language-specific data.

LAPT and domain adaptive pre-training (DAPT) represent continual pre-training techniques on *large* unlabelled corpora. Another effective continual pre-training technique on *small* data is task-adaptive pre-training (TAPT). TAPT is a technique to continue the pre-training of a PLM on a smaller but more relevant corpus of unlabeled text from the target task, such as the task training data. We reported that performing LAPT followed by TAPT with AfroXLMR significantly improved its performance on AfriSenti, beating AfriBERTa and AfroXLMR

without TAPT.

However, despite their excellent performance, deep learning-based approaches have a disadvantage: they do not explain how they compute the sentiment values, whereas lexicon-based approaches do. Moreover, experimental results on EcoSentiPT show that deep learning approaches perform better on short sentence fractions than on long sentences. This may be due to the small size of the training data, which reveals another drawback of deep learning approaches: they require large amounts of data.

Therefore, we investigated whether we can combine deep learning and lexicon-based approaches to exploit their advantages. We proposed a neuro-symbolic approach that uses symbolic rules in conjunction with a lexicon generated by BERTimbau [150]. Although our neuro-symbolic approach does not reach the performance level of deep learning approaches, it has the advantage of providing explanations for sentiment predictions. This is essential for some users who want to understand how the sentiment values are derived, even at the cost of some performance loss. We also observed that our neuro-symbolic approach on sentence fractions outperforms the purely symbolic approach, indicating the benefit of using a lexicon generated by a pre-trained language model. Additionally, we reported the approach of one of the best teams in AfriSenti-SemEval shared task leveraging sentiment lexicon as additional information for pre-training a language model that achieves competitive performance. As a future work, a possible direction is to use deep learning to automatically construct shifter patterns since our current approach only generates a unigram lexicon using deep learning.

In this chapter, we highlight the following main contributions:

- **NaijaSenti:** We created a NaijaSenti dataset and provide a comprehensive evaluation of state-of-the-art methods on the NaijaSenti dataset. We have demonstrated the effectiveness of various transfer learning strategies and provided insights into their performance on low-resource languages. The results were published in a paper entitled *NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis* [161].
- **AfriSenti** We introduced the AfriSenti dataset and provided a comprehensive benchmark experiment on the dataset. Our findings have important implications for future research and pave the way for further advancements in sentiment analysis for African languages. The results were published in a paper entitled *AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages* [159].
- **AfriSenti-SemEval** For the first time, we organized a SemEval shared task focusing on sentiment analysis for African languages. A total of 44 teams participated in the task and 29 teams submitted system description papers. This represents a significant contribution to the field and demonstrates the growing interest in developing NLP methods for African languages. Shared task description paper entitled *SemEval-2023*

Task 12: Sentiment analysis for African Languages [154].

- **Neuro-symbolic Approach for Sentiment Analysis** We propose a new neuro-symbolic approach for sentiment analysis that combines deep learning and lexicon-based methods [160].

Chapter 7

Conclusions and Future Work

"There is no real ending. It's just the place where you stop the story."

Frank Herbert

In this thesis, we described our research in sentiment analysis and proposed several novel approaches. These include the automatic generation of domain-specific sentiment lexicon, the issue of exploiting contextual information when using classical lexicon-based sentiment analysis, and the challenge of dealing with low-resource sentiment analysis corpora for African languages. We also introduced a neuro-symbolic approach to sentiment analysis that provides both explanation and quite good performance. We briefly summarize our contributions, findings, and future research directions in this chapter.

7.1 Contributions and Findings

Sentiment analysis corpora for African languages Chapter 3 introduced AfriSenti and NaijaSenti—the first publicly available large-scale and manually annotated Twitter sentiment dataset for African languages. We proposed methods to enable the collection, filtering, and annotation of such low-resource language data. Additionally, we introduced a manually annotated sentiment lexicon in these languages. We discussed the challenges faced during data collection and annotation, as well as experimental results using state-of-the-art pre-trained language models. Furthermore, we organized AfriSenti-SemEval, the first Afro-centric SemEval shared task on sentiment analysis for African languages. Our work advances sentiment analysis research for under-represented languages and stimulates interest in this area.

Proposal of two automatic approaches for sentiment lexicon generation In Chapter 4 we propose two methods for automatically generating domain-specific sentiment lexicons from labelled corpora. The first uses statistical/machine learning-like techniques to construct the lexicon based on the distribution of ratings of different words in the training data. The second method uses a BERT-based pre-trained and fine-tuned language model to assign sentiment values to words. We also presented a method for combining a domain-specific sentiment lexicon with a general-purpose one. This work contributes to the development of new methods for automatic sentiment-lexicon creation.

We found that sentiment lexicons generated from pre-trained language models generally outperform sentiment lexicons generated from statistical/machine-learning-like approaches. However, the statistical/machine-learning-like approach outperforms the prior work this method has been built on. We also found that the domain-specific and extended sentiment lexicon improves the performance of lexicon-based methods on domain-specific datasets.

Proposal of an enhanced lexicon-based sentiment analysis In Chapter 5 we describe an enhanced lexicon-based sentiment analysis approach that incorporates shifter patterns and rules. These patterns capture some contextual information and linguistic knowledge. We showed that the use of shifter patterns improves the accuracy of the lexicon-based approach. This approach has the advantage that it is easy to generate explanations regarding how the predictions of sentiment values are generated. We found that the proposed extended approach outperforms the vanilla approach.

Benchmarking pre-trained language models for sentiment analysis on our dataset In chapter 6, we examined the performance of pre-trained language models (PLMs) for sentiment analysis on three datasets: EcoSentiPT, NaijaSenti, and AfriSenti. Our experimental study showed that language-specific PLMs (such as AfriBERTa and BERTIMbau) achieved better results than massively multilingual PLMs such as mBERT. This indicates that PLMs benefit from pre-training on a specific language rather than a large number of languages. This work contributes a novel approach for neuro-symbolic sentiment analysis and provides extensive benchmark experiments on AfriSenti and NaijaSenti datasets, which can serve as future references for sentiment analysis in African languages.

We also showed that when Language Adaptive Pre-training (LAPT) is used with a large corpus that is not in the same domain, it presents better performance than simply fine-tuning PLMs. This demonstrates the effectiveness of adaptive pre-training. In addition, we found that performing LAPT followed by Task Adaptive Pre-training (TAPT) significantly improved performance. We also found that deep learning approaches perform better on short sentence fragments than on long sentences on the EcoSentiPT dataset. This may be due to the small size of the training data, which reveals another drawback of deep learning

approaches: they require large amounts of data.

Proposal of neuro-symbolic approach to sentiment analysis In chapter 6, we describe a neuro-symbolic approach to sentiment analysis that combines deep learning-based and lexicon-based methods. We used transformer-based pre-trained language models (BERT-Ibau) to generate a sentiment lexicon, which we used to generate predictions using our enhanced lexicon-based method that incorporates shifter patterns.

7.2 Future work

In this section, we outline some possible future directions.

Extend AfriSenti Datasets A future work may extend the AfriSenti dataset to more low-resource African languages and explore other aspects of sentiment analysis beyond polarity classification, such as aspect-based sentiment analysis or emotion detection, which are understudied in low-resource African languages. In addition, it is possible to investigate datasets for sentiment analysis from various domains other than social media.

Language Adaptive Pre-Training (LAPT) with Large Domain Data We found that LAFT with domain-specific data performs worse than LAPT with out-of-domain data. A possible future work is to investigate LAFT with sufficient data from the same domain.

Twitter-Specific Language Model Afro-centric pre-trained language models are becoming increasingly available. However, they are pre-trained on general corpora. A future direction is to create a domain-specific Afro-centric semantic pre-trained language model using only Twitter data. Such a pre-trained language model has been shown to outperform general pre-trained language models [110].

Dialect Adaptive-Pre-training We showed that LAPT works well in improving language models. A future research direction is to explore Dialect Adaptive-Pre-training. For example, one can use a large unlabeled corpus in European Portuguese to perform continual pre-training of BERTimbau—which was pre-trained on Brazilian Portuguese and evaluate its benefits.

Automatic Generation of Emotion Lexicon We have proposed two methods for automatically generating sentiment lexicons using pre-trained language models. In future work,

our approach for using a pre-trained language model to generate sentiment lexicons could be adapted to automatically generate emotion lexicons.

Automatic Generation of Shifter Patterns In this thesis, we manually create a shifter pattern. Future work is to investigate the automatic generation of shifter patterns and rules using pre-trained language models.

Extend neuro-symbolic to other languages Another interesting line would be to supply the DL not only with the target text but also with various forms of knowledge in a symbolic form, such as POS tags attached to individual words, dependency relationships, knowledge graphs, structures permitting to identify synonyms. Another future direction is to extend our neuro-symbolic approach to other domains and languages, especially low-resource ones.

Overall, our work contributes to the advancement of sentiment analysis and contributes to the advancement of sentiment analysis in low-resource African languages by providing new data, methods, and insights for sentiment analysis.

References

- [1] Saheed Salahudeen Abdullahi, Sun Yiming, Shamsuddeen Hassan Muhammad, et al. Deep sequence models for text classification tasks. In *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6. IEEE, 2021.
- [2] Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, and **Muhammad, Shamsuddeen**. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France, June 2022. European Language Resources Association.
- [3] Amina Imam Abubakar, Abubakar Roko, Aminu Muhammad Bui, and Ibrahim Saidu. An enhanced feature acquisition for sentiment analysis of english and hausa tweets. *International Journal of Advanced Computer Science and Applications*, 12(9), 2021.
- [4] David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv:2205.02022*, 2022.
- [5] Basant Agarwal and Namita Mittal. Machine learning approach for sentiment analysis. In *Prominent feature extraction for sentiment analysis*, pages 21–45. Springer, 2016.
- [6] Sagar Ahire. A Survey of Sentiment Lexicons. 2000.
- [7] Ibrahim Said Ahmad, Azuraliza Abu Bakar, Mohd Ridzwan Yaakub, and Shamsuddeen Hassan Muhammad. A survey on machine learning techniques in movie revenue prediction. *SN Computer Science*, 1:1–14, 2020.
- [8] Nora Al-Twairesh, Hend Al-Khalifa, AbdulMalik Al-Salman, and Yousef Al-Ohali. Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72, 2017.

- [9] Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France, May 2020. European Language Resources Association.
- [10] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.
- [11] Sattam Almatarneh and Pablo Gamallo. Automatic construction of domain-specific sentiment lexicons for polarity classification. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15*, pages 175–182. Springer, 2018.
- [12] Sattam Almatarneh and Pablo Gamallo. Comparing supervised machine learning strategies and linguistic features to search for very negative opinions. *Information (Switzerland)*, 10(1), 2019.
- [13] Orestes Appel, Francisco Chiclana, Jenny Carter, and Hamido Fujita. A hybrid approach to the sentiment analysis problem at the sentence level. *Knowledge-Based Systems*, 108:110–124, 2016.
- [14] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet. *Analysis*, pages 1–12, 2010.
- [15] JoAnne Banks-Wallace. Talk that talk: Storytelling and analysis rooted in african american oral tradition. *Qualitative Health Research*, 12(3):410–426, 2002. PMID: 11918105.
- [16] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, 2022.
- [17] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. Bad Is Stronger Than Good. *Review of General Psychology*, 2001.
- [18] Bob Beddor. Subjective Disagreement. *Noûs*, 53(4):819–851, December 2019.
- [19] Tadesse Destaw Belay, Abinew Ali Ayele, Getie Gelaye, Seid Muhie Yimam, and Chris Biemann. Impacts of homophone normalization on semantic models for amharic.

- In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106, 2021.
- [20] Tarek R. Besold, Artur d’Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C. Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. *Neural-Symbolic Learning and Reasoning: A Survey and Interpretation*, pages 1–51. IOS press, 2022.
 - [21] Pavel Brazdil, Purificação Silvano, Fátima Silva, **Muhammad, Shamsuddeen**, Fátima Oliveira, João Cordeiro, and António Leal. Extending general sentiment lexicon to specific domains in (semi-)automatic manner. In Sara Carvalho and Renato R. Souza, editors, *SALLD-1: Proceedings of the Workshop on Sentiment Analysis & Linguistic Linked Data*. CEUR, Vol. 3064, 2021.
 - [22] Pavel Brazdil and **Shamsuddeen, Muhammad** et.al. Semi-automatic approaches for exploiting shifter patterns in domain-specific sentiment analysis. *Mathematics*, 10(18), 2022.
 - [23] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
 - [24] Henrico Bertini Brum and Maria das Graças Volpe Nunes. Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*, 2017.
 - [25] Aminu Bui. OpenAIR at RGU Contextual Lexicon-based Sentiment Analysis for Social Media. 2016.
 - [26] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio (Editors) Feraco. *A Practical Guide to Sentiment Analysis*. Springer, 2017.
 - [27] Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. Senticnet 7: a commonsense-based neurosymbolic ai framework for explainable sentiment analysis. *Proceedings of LREC 2022*, 2022.
 - [28] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the National Conference on Artificial Intelligence*, 2014.

- [29] Jan Carter-Black. Teaching cultural competence: An innovative strategy grounded in the universality of storytelling as depicted in african and african american storytelling traditions. *Journal of Social Work Education*, 43(1):31–50, 2007.
- [30] Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.
- [31] Rosario Catelli, Serena Pelosi, and Massimo Esposito. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3), 2022.
- [32] Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online, November 2020. Association for Computational Linguistics.
- [33] Paula Chesley, Bruce Vincent, Li Xu, and Rohini K. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Spring Symposium - Technical Report*, 2006.
- [34] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, 2021.
- [35] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.
- [36] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.
- [37] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*, 2021.
- [38] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [39] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdooomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2015.
 - [40] V Dogra, S Verma, Kavita, P Chatterjee, J Shafi, J Choi, and MF Ijaz. A complete process of text classification system using state-of-the-art NLP models. *Computational Intelligence and Neuroscience*, 2022.
 - [41] David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. 2020. Url: <http://www.ethnologue.com>.
 - [42] Rosa Falotico and Piero Quatto. Fleiss’ kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470, 2015.
 - [43] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.
 - [44] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online, June 2021. Association for Computational Linguistics.
 - [45] Ana Catarina Forte and Pavel B Brazdil. Determining the level of clients’ dissatisfaction from their commentaries. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 74–85. Springer, 2016.
 - [46] Quintin Gee. Review of script displays of african languages by current software. *New Review of Hypermedia and Multimedia*, 11:247 – 255, 2005.
 - [47] Deepanway Ghosal, Devamanyu Hazarika, Abhinaba Roy, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3198–3210, Online, July 2020. Association for Computational Linguistics.
 - [48] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

- [49] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).
- [50] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [51] H Grote and F Schmidt. Mad-x-an upgrade from mad8. In *Proceedings of the 2003 Particle Accelerator Conference*, volume 5, pages 3497–3499. IEEE, 2003.
- [52] Itisha Gupta and Nisheeth Joshi. Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. *Journal of intelligent systems*, 29(1):1611–1625, 2020.
- [53] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.
- [54] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 595–605. NIH Public Access, 2016.
- [55] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.
- [56] Hongyu Han, Jianpei Zhang, Jing Yang, Yiran Shen, and Yongshi Zhang. Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools and Applications*, 77(16):21265–21280, 2018.
- [57] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021.
- [58] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.
- [59] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedziec, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics.
- [60] Pascal Hitzler and Md Kamruzzaman Sarker. *Neuro-symbolic Artificial intelligence: State of the Art*. IOS press, 2022.

- [61] Mirka Honkanen and Julia Müller. Interjections and emojis in nigerian online communication. *World Englishes*, 2021.
- [62] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.
- [63] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.
- [64] 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*, pages 168–177, 2004.
- [65] Jade ; Neubig Graham ; D’souza Daniel ; Kreutzer Julia ; Lignos Constantine ; Palen-Michel Chester ; Buzaaba Happy ; Rijhwani Shruti ; Ruder Sebastian ; Mayhew Stephen ; Abebe Azime Israel ; **Muhammad, Shamsuddeen** ; Ifeoluwa Adelani, David ; Abbott et al. Masakhaner: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.
- [66] Abimbola Rhoda Iyanda and Omolayo Abegunde. Predicting sentiment in yorùbá written texts: A comparison of machine learning models. In *Proceedings of SAI Intelligent Systems Conference*, pages 416–431. Springer, 2019.
- [67] Valentin Jijkoun, Maarten De Rijke, and Wouter Weerkamp. Generating focused topic-specific sentiment lexicons. In *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010.
- [68] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.
- [69] Dame Jovanoski, Veno Pachovski, and Preslav Nakov. Sentiment analysis in twitter for macedonian. *arXiv preprint arXiv:2109.13725*, 2021.
- [70] Giuseppe Jurman, Samantha Riccadonna, and Cesare Furlanello. A comparison of mcc and cen error measures in multi-class prediction. *PLOS ONE*, 7(8):1–8, 08 2012.
- [71] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *COLING/ACL 2006 - EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 355–363, 2006.

- [72] Chamandeep Kaur, Mawahib Sharafeldin Adam Boush, Samar Mansoor Hassen, Wafaa Abushmlah Hakami, Mohammed Hassan Osman Abdalraheem, Najla Mohammed Galam, Nedaa Abdulaziz Hadi, Nedaa Abdulaziz Hadi, and Atheer Omar S Benjeed. Incorporating sentimental analysis into development of a hybrid classification model. *Int. J. Health Sci. (IJHS)*, 6(S1):1709–1720, 2022.
- [73] J. Kocoń, J. Baran, M. Gruza, A. Janz, M. Kajstura, P. Kazienko, W. Korczyński, P. Miłkowski, M. Piasecki, and J. Szolomicka. Neuro-symbolic models for sentiment analysis. In *Computational Science – ICCS 2022: 22nd International Conference*, pages 667—681, 2022.
- [74] Taiwo Kolajo, Olawande Daramola, and Ayodele Adebisi. Sentiment analysis on naija-tweets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 338–343, 2019.
- [75] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PLoS ONE*, 10(12):e0144296, 2015.
- [76] Tim Kreutz and Walter Daelemans. Enhancing General Sentiment Lexicons for Domain-Specific Use. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1056–1064, 2018.
- [77] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmongkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Irero Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, **Muhammad, Shamsuddeen Hassan**, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.
- [78] Akshi Kumar and Teeja Mary Sebastian. Sentiment Analysis: A Perspective on its Past, Present and Future. *International Journal of Intelligent Systems and Applications*, 4(10):1–14, 2012.
- [79] Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad, and Richard Johansson. An arabic tweets sentiment analysis dataset (atsad) using distant supervision and self training. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 1–8, 2020.
- [80] Kevin Christophe Labille, Susan Gauch, Sultan Alfarhood, Kevin Labille, and Susan Gauch. Creating Domain-Specific Sentiment Lexicons via Text Mining Creating

- Domain-Specific Sentiment Lexicons via Text Mining. In *Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, number September, 2017.
- [81] Hoai Bac Le and Huy Nguyen. Twitter sentiment analysis using machine learning techniques. In *International Conference on Computer Science, Applied Mathematics and Applications*, 2020.
- [82] Colin Leong, Herumb Shandilya, Bonaventure FP Dossou, Atnafu Lambebo Tonja, Joel Mathew, Abdul-Hakeem Omotayo, Oreen Yousuf, Zainab Akinjobi, Chris Chinenye Emezue, Shamsudeen Muhammad, et al. Adapting to the low-resource double-bind: Investigating low-compute methods on low-resource african languages. *arXiv preprint arXiv:2303.16985*, 2023.
- [83] Bing Liu. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, 2(1):627–666, 2010.
- [84] Bing Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press, 2020.
- [85] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining Text Data*, volume 9781461432, chapter 1, pages 415–463. Springer, Boston, MA, 2012.
- [86] H. Liu and P. Singh. ConceptNet - a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.
- [87] 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.
- [88] 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*, abs/1907.11692, 2019.
- [89] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. Automatic construction of a context-aware sentiment lexicon: An optimization approach. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pages 347–356, 2011.
- [90] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuuttila. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.
- [91] Granville J Matheson. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ*, 7:e6918, 2019.
- [92] Million Meshesha and C. V. Jawahar. Indigenous scripts of african languages. *Indilinga: African Journal of Indigenous Knowledge Systems*, 6:132–142, 2008.

- [93] Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon KABONGO KABENAMUALU, Elizabeth Salesky, Irero Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete AGBOLO, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. Biblets: a large, high-fidelity, multilingual, and uniquely african speech corpus. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2383–2387. ISCA, 2022.
- [94] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 1995.
- [95] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*, 2021.
- [96] Saif Mohammad. A Practical Guide to Sentiment Annotation: Challenges and Solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, San Diego, California, 2016. Association for Computational Linguistics.
- [97] Saif M. Mohammad. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement (Second Edition)*. Elsevier, 2021.
- [98] Saif M. Mohammad, Svetlana Kiritchenko, and Xiao-Dan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *International Workshop on Semantic Evaluation*, 2013.
- [99] S.M. Mohammad, M. Salameh, and S. Kiritchenko. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130, 2016.
- [100] Antonio Moreno-Ortiz, Javier Fernández-Cruz, and Chantal Pérez Chantal Hernández. Design and evaluation of sentiecon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5065–5072, 2020.
- [101] Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. An algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1202–1210, 2020.
- [102] Igor Mozetič, Miha Grčar, and Jasmina Smailović. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036, 2016.

- [103] Andrius Mudinas, Dell Zhang, and Mark Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12*, pages 1–8, 2012.
- [104] Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey. Domain-based lexicon enhancement for sentiment analysis. In *SMA@ BCS-SGAI*, pages 7–18, 2013.
- [105] Putra Fissabil Muhammad, Retno Kusumaningrum, and Adi Wibowo. Sentiment analysis using word2vec and long short-term memory (lstm) for indonesian hotel reviews. *Procedia Computer Science*, 179:728–735, 2021.
- [106] Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online, June 2021. Association for Computational Linguistics.
- [107] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter, 2019.
- [108] Nikita Nangia and Samuel R Bowman. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. *arXiv preprint arXiv:1905.10425*, 2019.
- [109] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwale Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Irero Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics.
- [110] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets, 2020.

- [111] Emmanuel Sunday Nwofe. Pro-biafran activists and the call for a referendum: A sentiment analysis of ‘biafraexit’ on twitter after uk’s vote to leave the european union. *Journal of Ethnic and Cultural Studies*, 4(1):65, 2017.
- [112] Emeka Ogbuju and Moses Onyesolu. Development of a general purpose sentiment lexicon for Igbo language. In *Proceedings of the 2019 Workshop on Widening NLP*, page 1, Florence, Italy, August 2019. Association for Computational Linguistics.
- [113] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [114] Tolulope Olagunju, Oladapo Oyebode, and Rita Orji. Exploring key issues affecting african mobile ecommerce applications using sentiment and thematic analysis. *IEEE Access*, 8:114475–114486, 2020.
- [115] Sunday Adewale Olaleye, Ismaila Temitayo Sanusi, and Jari Salo. Sentiment analysis of social commerce: a harbinger of online reputation management. *International Journal of Electronic Business*, 14(2):85–102, 2018.
- [116] Ikechukwu Onyenwe, Samuel Nwagbo, Njideka Mbeledogu, and Ebele Onyedinda. The impact of political party/candidate on the election results from a sentiment analysis perspective using# anambradecides2017 tweets. *Social Network Analysis and Mining*, 10(1):1–17, 2020.
- [117] Sylvester Olubolu Orimaye, Saadat M Alhashmi, and Siew Eu-gene. Sentiment analysis amidst ambiguities in youtube comments on yoruba language (nollywood) movies. In *Proceedings of the 21st International Conference on World Wide Web*, pages 583–584, 2012.
- [118] Xi Ouyang, Pan Zhou, Cheng Hua Li, and Lijun Liu. Sentiment analysis using convolutional neural network. In *Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 2359–2364. IEEE, 2015.
- [119] Oladapo Oyebode and Rita Orji. Social media and sentiment analysis: The nigeria presidential election 2019. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0140–0146. IEEE, 2019.

- [120] Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. Semantic enrichment of nigerian pidgin english for contextual sentiment classification. *arXiv preprint arXiv:2003.12450*, 2020.
- [121] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.
- [122] Bo Pang. Automatic analysis of document sentiment. *Klinische Monatsbl*, 211(August):126, 2006.
- [123] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
- [124] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.
- [125] S. Park, W. Lee, and I.-C. Moon. Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 56:38–44, 2015.
- [126] Tal Perry. LightTag: Text Annotation Platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
- [127] Jonas Pfeiffer, Ivan Vuli, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. In *Proceedings of EMNLP 2020*, 2020.
- [128] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [129] Christopher Potts. On the negativity of negation. *Semantics and Linguistic Theory*, 20(20):636, 2015.
- [130] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online, August 2021. Association for Computational Linguistics.
- [131] Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 2009.

- [132] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding domain sentiment lexicon through double propagation. In *IJCAI International Joint Conference on Artificial Intelligence*, 2009.
- [133] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, 2020.
- [134] Ochilbek Rakhmanov. A comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments. *Procedia Computer Science*, 178:194–204, 2020.
- [135] Justus J Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa. *Online submission*, 2005.
- [136] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter S. Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):1–12, 11 2016. Copyright - EPJ Data Science is a copyright of Springer, 2016; Last updated - 2017-02-06.
- [137] Eshrag Refaee and Verena Rieser. An arabic twitter corpus for subjectivity and sentiment analysis. In *LREC*, pages 2268–2273, 2014.
- [138] Eshrag Refaee and Verena Rieser. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 2268–2273, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).
- [139] 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.
- [140] Paul Rozin and Edward B Royzman. Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, 5(4):296–320, 2001.
- [141] Sebastian Ruder. Recent Advances in Language Model Fine-tuning. <http://ruder.io/recent-advances-lm-fine-tuning>, 2021.
- [142] Kevin P Scannell. Statistical unicodification of african languages. *Language resources and evaluation*, 45(3):375–386, 2011.
- [143] Teven Le Scao et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

- [144] Jonas Paul Schöne, Brian Parkinson, and Amit Goldenberg. Negativity spreads more than positivity on twitter after both positive and negative political situations. *Affective Science*, 2(4):379–390, 2021.
- [145] Marc Schulder, Michael Wiegand, and Josef Ruppenhofer. Automatic generation of lexica for sentiment polarity shifters. *Natural Language Engineering*, 27(2):153–179, 2021.
- [146] Abdullah Aziz Sharfuddin, Md Nafis Tihami, and Md Saiful Islam. A deep recurrent neural network with BiLSTM model for sentiment classification. In *Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–4. IEEE, 2018.
- [147] Fátima Silva, Purificação Silvano, António Leal, Fátima Oliveira, Pavel Brazdil, João Cordeiro, and Débora Oliveira. Análise de sentimento em artigos de opinião. *Linguística: Revista de estudos linguísticos da Universidade do Porto*, 13:79–114, 2018.
- [148] Marshall S. Smith, Daniel M. Ogilvia, Philip J. Stone, Dexter C. Dunphy, and John J. Hartman. The General Inquirer: A Computer Approach to Content Analysis., 1967.
- [149] Fredrik Sommar and Milosz Wielondek. Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification. 2015.
- [150] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer, 2020.
- [151] V S Subrahmanian and Diego Reforgiato. AVA Adjective-Verb-Adverb Combinations for Sentiment Analysis. *IEEE Intelligent Systems*, 23(August):43–50, 2008.
- [152] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307, 2011.
- [153] Cátia Tavares, Ricardo Ribeiro, and Fernando Batista. Sentiment analysis of portuguese economic news. In *Proceedings of the 10th Symposium on Languages, Applications and Technologies (SLATE 2021), Article 17*, volume 94, pages 17:1–17:13. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.
- [154] **Muhammad, Shamsuddeen** et al. SemEval 2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada, July 2023.
- [155] **Muhammad, Shamsuddeen Hassan**. An overview of sentiment analysis approaches. In *MAP-i Seminar Proceedings*, pages 65–70, 2019.

- [156] **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio Jorge. Incremental approach for automatic generation of domain-specific sentiment lexicon. In *Proceedings of Advances in Information Retrieval, LNCS*, volume 12036, pages 619–623. Springer, 2020.
- [157] **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio Jorge. Combining symbolic and deep learning approaches for sentiment analysis. In Pascal Hitzler, Md Kamruzzaman Sarker, and Audrey Eberhart, editors, *Frontiers in Artificial Intelligence and Applications*, volume 369 of *Frontiers in Artificial Intelligence and Applications*, pages 506–521. IOS Press, August 2023.
- [158] **Muhammad, Shamsuddeen Hassan**, Pavel Brazdil, and Alípio M. Jorge. Leveraging symbolic and deep learning techniques for explainable sentiment analysis. In *22nd EPIA Conference on Artificial Intelligence*. EPIA, 2023.
- [159] **Muhammad, Shamsuddeen Hassan** et al. AfriSenti: a Twitter sentiment analysis benchmark for African languages. *arXiv preprint arXiv:2302.08956*, 2023.
- [160] **Muhammad, Shamsudeen Hassan**, Pavel Brazdil, and Alípio Jorge. Combining symbolic and deep learning approaches for sentiment analysis. In Pascal Hitzler, Md Kamruzzaman Sarker, and Aaron Eberhart, editors, *Compendium of Neuro-Symbolic Artificial Intelligence*.
- [161] **Shamsuddeen Hassan Muhammad**, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa’id, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. Naijasenti: A Nigerian twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference*, pages 590–602, Marseille, France, June 2022. European Language Resources Association.
- [162] S Thara and Prabakaran Poornachandran. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388, 2018.
- [163] Radoslava Trnavac, Debopam Das, and Maite Taboada. Discourse relations and evaluation. *Corpora*, 11(2):169–190, 2016.
- [164] Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers*. " O’Reilly Media, Inc.", 2022.
- [165] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

- [166] Uduak Umoh, Imo Eyoh, Etebong Isong, Anietie Ekong, and Salvation Peter. Using interval type-2 fuzzy logic to analyze igbo emotion words. *Journal of Fuzzy Extension and Applications*, 1(3):217–240, 2020.
- [167] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, 2013.
- [168] Wouter van Atteveldt, Mariken ACG van der Velden, and Mark Boukes. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140, 2021.
- [169] C.J. van Rijsbergen. *Information retrieval (2nd ed.)*. Butterworth-Heinemann, 1979.
- [170] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [171] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.
- [172] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.
- [173] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- [174] Alfian Farizki Wicaksono, Clara Vania, Bayu Distiawan, and Mirna Adriani. Automatically building a corpus for sentiment analysis on indonesian tweets. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 185–194, 2014.
- [175] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of*

- the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
- [176] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.
 - [177] Mengzhou Xia, Guoqing Zheng, Subhabrata Mukherjee, Milad Shokouhi, Graham Neubig, and Ahmed Hassan Awadallah. Metaxl: Meta representation transformation for low-resource cross-lingual learning. *arXiv preprint arXiv:2104.07908*, 2021.
 - [178] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.
 - [179] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2):183–207, 2018.
 - [180] Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, 2020.
 - [181] Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.
 - [182] Chiara Zucco, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. Explainable sentiment analysis with applications in medicine. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1740–1747. IEEE, 2018.
 - [183] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs, 2011.