Chapter 3

NLU in Low-Resource Languages: A case of Sentiment Analysis

Sentiment analysis, a critical facet of NLU, has undergone a transformative journey, evolving from rule-based systems to contemporary methodologies dominated by LLMs. In this chapter we explore the synergy between two distinct yet complementary approaches: LRs, which is a rule-based approach and LLMs, in the context of sentiment analysis for low-resource languages. By amalgamating insights from two experiments, each dedicated to one of these approaches, we unravel the complexities of Sentiment Analysis in diverse linguistic landscapes, with a particular focus on the linguistic richness of India.

The first experiment focuses on the foundational step of categorizing linguistic data into subjective and objective categories. Subjective information, characterized by personal beliefs and opinions, necessitates rigorous filtering to enhance sentiment analysis accuracy. Leveraging lexical cues, the research investigates linguistic devices and subjectivity encoding in diverse Indian languages. This exploration aligns with the overarching RQs, examining linguistic devices' similarity across languages in a sprachbund and evaluating the efficacy of cost-effective, lexicon-heavy, rule-based systems for subjectivity identification.

In exploring morphologically rich Indian languages such as Bengali, Hindi, Kannauji, Khortha, and Odia, the study unveils the impact of linguistic typology on subjectivity encoding. The languages, chosen deliberately to avoid representation bias, form a linguistic mosaic representing the diversity within the Indian subcontinent. As these languages encode information differently, LRs emerge as an efficient means of representation, particularly in morphologically rich languages. The exploration of subjectivity in these languages not only answers RQs but also serves as a foundational step toward understanding linguistic diversity and devising language-specific sentiment analysis strategies.

The second experiment shifts the focus to the application of LLMs, emphasizing the significance of nuanced approach for subjectivity analysis in low-resource languages. The linguistic diversity encapsulated in languages like Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu poses unique challenges for NLP, compounded by a scarcity of labelled datasets. Recognizing this gap, the study underscores the importance of subjectivity analysis, not only for sentiment extraction but also for extracting nuanced opinions and emotions embedded in textual data.

The application of LLMs, with their pre-trained adaptability and transfer learning capabilities, holds promise for effective subjectivity analysis in low-resource languages. However, the study identifies a critical gap in existing research, i.e., the limited exploration of LLMs in the context of languages with constrained linguistic resources. These experiments aim to bridge this void by delving into recent advancements, such as PE and ICL, which have demonstrated their potential in enhancing LLMs' performance in NLP applications.

This chapter embarks on a journey to unravel the dynamics of sentiment analysis in low-resource languages, weaving insights from rule-based subjectivity identification and LLM applications. The overarching objective is to provide a comprehensive understanding of subjectivity encoding, sentiment extraction, and the unique challenges posed by linguistic diversity and resource constraints. To achieve this, the chapter is organized into two distinct subsections that delve into the methodologies, findings, and implications of each experiment, culminating in a holistic synthesis that sheds light on the synergies between rule-based models and LLMs in the realm of sentiment analysis for low-resource languages.

3.1 Sentiment Analysis with LRs

Discourse can be categorized into subjective and objective utterances. Objective statements are factual, verifiable, while subjective ones express the speaker's emotions or opinions. In this experiment we conduct a comparative analysis of linguistic devices demonstrating subjectivity across five Indian languages. It formalizes LRs for subjective constructions and implements a LR-based FST for subjectivity identification. Evaluation on test data from various domains yield good results, with 91% accuracy in the politics

domain and an average accuracy of ~84%, showcasing the effectiveness of the proposed approach in discerning subjective and objective content.

3.1.1 Background

This set of experiments explores use of LR, shedding light on the complexities inherent in languages with rich morphological structures. The research encompasses Bengali, Hindi, Kannauji, Khortha, and Odia. These languages are carefully selected to represent a diverse linguistic landscape, aiming to avoid representation bias and explore the idiosyncrasies embedded in their linguistic typologies.

3.1.2 Related Works

Subjective information, grounded in personal belief, opinion, or preference, poses a significant challenge due to the absence of universally accepted criteria for validating its truth value. Filtering out subjective information becomes imperative in processes such as opinion classification and sentiment analysis.

Minimizing objective information helps: Any oversight in classifying subjective data can lead to spikes in confusion metrics, escalating the cost of further information processing for sentiment analysis. Pang and Lee demonstrated the effectiveness of minimizing objective information before sentiment analysis, highlighting the superiority of a Naive Bayes polarity classifier trained on subjective extracts compared to mixed-content original text (Pang & Lee, 2004).

Lexical cues help in subjectivity identification: Wang and Fu contribute to the discourse by attempting subjectivity identification in Chinese through the utilization of lexical cues. They categorize these cues into five distinct classes: opinion indicator, opinion word, named entity, opinion object, and adverb of degree. Introducing the

concept of sentiment density, they successfully classify sentences into subjective or objective categories based on these subjectivity keywords (Wang & Fu, 2010). This approach emphasizes the role of lexical information in understanding subjectivity within a linguistic context.

Emotion Annotation in Bengali: Das and others have contributed to the exploration of subjectivity in an Indian language by annotating a web-based Bengali corpus based on the emotions attached to words and phrases. This granular classification covers six emotion classes, each with three types of intensities, providing a nuanced understanding of the emotional dimensions within the language (Das & Bandyopadhyay, 2010). This comprehensive emotional annotation serves as a foundation for understanding the nuanced sentiments encoded in Bengali text.

Leveraging Lexicons and WordNet for Subjectivity Identification: Narayan et al. utilize the Hindi Subjective Lexicon and Hindi WordNet to identify the semantic orientation of adjectives and adverbs, enriching the understanding of sentiment nuances within the language (Narayan et al., 2002). Mittal and Agrawal employ negation and discourse relations to identify sentiments in Hindi data, improving the existing Hindi SentiWordNet and achieving an impressive ~80% accuracy in Hindi review classification (Mittal & Agarwal, 2013). These approaches showcase the significance of leveraging linguistic resources and semantic relationships for accurate sentiment analysis.

Generative Lexicons and Language Typology: Some researchers propose a generative lexicon in the Hindi Subjectivity Analysis System. They employ the English OpinionFinder subjectivity lexicon and a small Hindi seed word list to create a subjectivity lexicon. This system demonstrates ~71% agreement with human evaluators and ~80% classification accuracy on parallel Hindi and English datasets (Jha et al., 2016).

The generation of lexicons in this manner provides valuable insights into the efficiency of lexicon-heavy approaches for sentiment analysis.

Language Typology and Morphological Richness: Languages encode and distribute information differently based on their typology. Morphologically rich languages convey substantial information through morphology, while syntactically rich languages encode information in sentence structures. The efficiency of LR in representing information in morphologically rich languages is emphasized, given their free word order attributed to information density at the morphological level (Tsarfaty et al., 2013).

Diverse Linguistic Landscape of India: India, known for its linguistic diversity, is a linguistic area full of morphologically rich languages. This research work zeroes in on the lexical analysis of five morphologically rich Indian languages, carefully avoiding adjacent languages in the language continuum to ensure a diverse representation.

1. **Bengali:** An Indo-Aryan language and the lingua franca of the Bengal region, Bengali boasts 284 million speakers globally (Emeneau, 1956).

2. **Hindi:** An Indo-Aryan language and a lingua franca in India, Hindi claims 592 million speakers in India and ranks as the third most widely spoken language globally (Registrar General India, 2011). Given the vast diversity and variations in linguistic features, some scholars argue that Hindi is better understood as a group of related languages, with modern Hindi serving as a standardized abstraction of this broader family (Kachru, 2006).

3. **Kannauji:** An Indo-Aryan language spoken by 10 million people in Kannauj and adjacent districts in Uttar Pradesh, establishing itself as a stand-alone locale (Chaturvedi, 2015).

4. **Khortha:** A language of the Indo-Aryan family spoken primarily in Jharkhand (Prasāda & Jhā, 1958), with 8 million speakers, making it the second most spoken language after Hindi in the region (Registrar General India, 2011).

5. **Odia:** An Indo-Aryan language spoken in the eastern parts of India, with 35 million speakers across Odisha, West Bengal, Jharkhand, and Chhattisgarh (Registrar General India, 2011). Odia exhibits influence from Dravidian languages, resulting in linguistic features that reflect characteristics of both the Indo-Aryan and Dravidian language families (Patnaik, 2014).

3.1.2.1 RQs

These experiments address two crucial RQs:

RQ1: Do languages in a sprachbund use similar linguistic devices to encode subjectivity? The study explores Subjective Constructions (SC) in five Indian languages with different geographical boundaries. Beyond syntactico-semantic analysis, the research formulates and compares LR to understand the linguistic idiosyncrasies in these languages.

RQ2: Can cost-effective, lexicon-heavy, rule-based systems provide good accuracy in subjectivity identification? The study employs a subset of existing open-source lexical resources in one language to create a rule-based (RB) system for subjectivity identification. Evaluation is performed against domain-agnostic and domain-specific gold-standard test sets.

This multifaceted exploration into subjectivity identification in morphologically rich Indian languages underscores the importance of linguistic cues, emotional annotation, lexicon generation, and language typology. The research collectively contributes to the understanding of sentiment analysis in diverse linguistic landscapes, offering valuable

insights into the nuances of subjectivity encoding across languages. The chapter seamlessly integrates these diverse methodologies, weaving a comprehensive narrative that advances our understanding of sentiment analysis in the context of low-resource languages with rich morphological structures.

3.1.3 Research Method

In this section, we outline our methodology for subjectivity classification, encompassing the annotation schema, data analysis for five chosen languages, the definition of the sentiment index, and the development of the LR-based FST.

3.1.3.1 Annotation Schema

For glossing linguistic data, we have adopted tags from the Penn tag set (Taylor et al., 2003), thus utilizing a subset of the Penn tag set as our tag-set. While the Penn tag set offers numerous fine-grained tags useful for linguistic analysis, some contribute minimally to computation. To streamline the process, we employ parallel coarse-grained tags in LRs for such fine-grained tags. This approach enhances efficiency by focusing on tags that significantly impact computational outcomes.

Table 3.1
Schema for linguistic annotation

S. No.	Fine-grained tag	Meaning	Coarse-grained tag
1	ADJ	Adjective	ADJ
2	ADV	Adverb	ADV
3	CC	Conjunction	-
4	COP	Copula	V
5	INDF	Indefinite	INDF
6	INTF	Intensifier	INTF
7	N	Noun	N
8	NCOM	Noun common	N
9	NEG	Negation	NEG
10	P	Pre/Post-position	P
11	PRON	Pronoun	PRON
12	PRONDEM	Pronoun demonstrative	PRON
13	PRONPRS	Pronoun personal	PRON
14	QW	Question word	-
15	V	Verb	V
16	VAUX	Verb auxiliary	V
17	VMOD	Verb model	V
18	VN	Verbal noun	V

3.1.3.2 Data Analysis

To construct LRs, we initiated a data collection drive in the selected languages. Language specialists were presented with a questionnaire covering everyday scenarios where individuals express opinions, emotions, or convey factual information. Following this, linguists in each language performed a detailed analysis of both subjective and objective expressions, examining grammatical structures and providing annotations. Particular emphasis was placed on subjective expressions relevant to our subjectivity classification task.

For uniformity, we utilized WX notation (Indian languages in ASCII) to transliterate the utterances (Gupta et al., 2010). The analysis section provides selective examples for different languages. In each example, the first layer presents the WX representation of the analyzed sentence. The second layer includes relevant syntactico-semantic tags corresponding to the words from the annotation schema outlined in Table 3.1. The third layer comprises the English gloss of the sentence, and the fourth layer provides the corresponding English translation.

This multi-layered approach not only facilitates a comprehensive understanding of the data but also ensures a standardized representation across different languages. By aligning the WX representation, syntactico-semantic tags, English gloss, and English translation, we establish a robust foundation for subsequent subjectivity classification.

3.1.3.2.1 Bengali

Bengali employs modifiers, such as adjectives and adverbs, extensively in constructing subjective expressions. The range of constructions varies, encompassing concise one-word sentences with a modifier to more elaborate sentences with multiple phrases. Examples illustrate this diversity:

benAras	xurxAnwa	destination	howe	pAre
PN	ADJ	NCOM	V	VAUX
banaras	terrific	destination	could be	

Banaras could be a terrific destination.

jAygATAa	xAruN
NCOM	ADJ
place	awesome

The place is awesome.

Intensifiers play a crucial role alongside modifiers, amplifying the intensity or degree of opinion and consequently enhancing the overall subjectivity of the sentence. The inclusion of the intensifier *khub* intensifies *bhalo*, transforming it into a robust subjective construction.

khAbArtA	khub	BAlo
NCOM	INTF	ADJ
food	very	good

The food is very good.

Negation is sometimes paired with positive or negative modifiers to augment the subjectivity of sentences. The subsequent examples showcase the incorporation of negation with adjectives. While these sentences maintain subjectivity even without negation, its inclusion fortifies their subjective nature.

hoteltA	ata	pariRkAr	nay
NCOM	ADV	ADJ	NEG
hotel	so	clean	not

The hotel is not so clean.

paribeSTA	ekebArei	BAlo	nay
NCOM	ADV	ADJ	NEG
ambience	at all	good	not

The ambience is not good at all.

Subjective constructions are also formed through the use of adverbs with nouns and the employment of metaphorical proper nouns. In the provided instances, nouns take on a metaphorical role to express an opinion. The overall polarity in such cases hinges on the choice of metaphorical nouns for comparison.

loktA	ekebAre	kAlkeute
NCOM	ADV	NCOM
the man	absolutely	king cobra

The man is absolutely king cobra.

rAjib	ki	ebAr	biBIRaN	halen
PN	QW	ADV	NCOM	V
rajib	QW	this time	bibhishan	become

Has Rajib become Bibhishan this time?

Conjunct verbs serve as another mechanism for creating subjective constructions. When a modifier accompanies a polar conjunct verb, it contributes to the development of a potent subjective construction.

KAbAartA	Axau	BAlo	lAge	ni
NCOM	ADV	ADJ	V	NEG
food	at all	good	feel	not

(I) have not liked the food at all.

Bengali's richness in linguistic devices allows for diverse methods of forming subjective constructions. From the use of intensifiers and negation to the use of adverbs with nouns and metaphorical proper nouns, each approach contributes to subjective expressions in Bengali. The exploration of conjunct verbs with modifiers further emphasizes the versatility and depth inherent in Bengali's subjective constructions.

3.1.3.2.2 Hindi

Subjectivity, akin to Bengali, finds expression in Hindi through the strategic use of adjectives, adverbs, conjunct verbs, and nouns. The subsequent examples delve into the

intricacies of constructing subjective expressions in Hindi. The incorporation of adjectives, both with and without negation, contributes to the formation of subjective constructions in Hindi. In the presented examples, the sentences remain inherently subjective, and the addition or removal of negation does not transform them into objective statements. However, the introduction of negation serves to negate the modifier's scope, intensifying the construction and imparting a robust subjective quality.

table	gaMxI	hE
NCOM	ADJ	COP
the table	dirty	is

The table is dirty.

KAnA	acCA	nahIM	hE
NCOM	ADJ	NEG	COP
the food	good	not	is

The food is not good.

This subjective constructions in Hindi leverage adverbs and negation, with the latter strategically negating the influence of adverbs to create potent subjective expressions. Negation and adverbs adds depth to the subjectivity, shaping nuanced linguistic constructions.

yahAz	XyAna	se	KAnA	nahIM	banAwe	hEM
PRONDEM	ADV	P	NCOM	NEG	V	VAUX
here	carefully		food	not	cook	do

(They) do not cook food here carefully.

mEM	yahAz	kaBI	nahIM	AUzgA
PRONPRS	PRONDEM	ADV	NEG	V
i	here	never		will come

I will never come here.

The utilization of conjunct verbs, particularly those with polar noun or adjective heads, contributes significantly to the formation of subjective expressions. The examples presented showcase instances both with and without negation, highlighting the diverse

ways in which conjunct verbs can shape subjective constructions with varying degrees of intensity.

mEM	isakI	anuSaMsA	karUMgI
PRONPRS	PRONDEM	ADJ	V
i	this	recommend	will

I will recommend this.

mEM	yahAz	Ane	kI	salAha	nahIM	xUzgI
PRONPRS	PRONDEM	VN	P	N	NEG	V
i	here	coming		suggest	not	would

I wouldn't suggest coming here.

Hindi allows for the formation of subjective constructions through the use of polar nouns. However, such constructions typically exhibit a weaker subjective nature unless accompanied by intensifiers, modifiers, or negation. The examples provided elucidate this point, underscoring the role of additional linguistic elements in enhancing the subjectivity of sentences.

use	sabakA	pyAra	milegA
PRONDEM	INDF	N	V
he	from everyone	love	will get

He will get love from everyone.

usako	wArIPa	milanI	cAhie
PRONDEM	N	V	VMOD
she	appreciation	get	should

She should get appreciation.

Hindi, much like Bengali, embraces a multifaceted approach to convey subjectivity. Through adjectives, adverbs, conjunct verbs, and nouns, it conveys subjective

expressions. The nuanced use of negation emerges as a key tool, intensifying subjectivity in specific contexts. The exploration of conjunct verbs sheds light on their role in forming varied subjective constructions. Additionally, the incorporation of polar nouns presents another avenue for subjective expression, emphasizing the linguistic versatility inherent in Hindi's subjective constructions.

3.1.3.2.3 Kannauji

Kannauji, like other languages within the sprachbund, employs linguistic devices for constructing subjective expressions. The examples below illustrate the language's utilization of adjectives paired with nouns to craft subjective constructions:

nIkI	jagaha
ADJ	NCOM
awesome	place

The place is awesome.

Subjective constructions in Kannauji often involve the pairing of adjectives with nouns. This linguistic device, shared with other languages in the sprachbund, lays the foundation for conveying subjectivity. The nuanced choice of adjectives influences the overall tone and sentiment of the expression.

jO	KAnA	bahuwE	svAxiRta	hE
Pronoun	NCOM	Intensifier	ADJ	V
this	food	very	good	is

The food is very good.

To intensify subjectivity, Kannauji incorporates intensifiers, negation, and adverbs into constructions. The use of the intensifier "bahuwE" in a previous example elevates the

degree of subjectivity compared to a construction with a singular polar modifier. The subsequent examples showcase the language's capacity to create stronger subjective constructions by combining adjectives with negation and adverbs. The inclusion of negation, as seen in one example, vividly projects negative polarity. Moreover, the introduction of the adverb "siMgatta" in another example serves to further enhance the degree of subjectivity.

kamarA	sAPZa	nAi	hE
NCOM	ADJ	NEG	V
room	clean	not	is

The room is not clean.

mAhOla	siMgatta	nAi	Tika	hE
NCOM	ADV	NEG	ADJ	V
ambience	at all	not	good	is

The ambience is not good at all.

Kannauji employs simile as a linguistic device for crafting subjective constructions. The choice of similes introduces cultural nuances, influencing the overall polarity of the construction. In the provided example, the term "siyAra," carrying a negative connotation in Kannauji culture, imparts a negative polarity to the construction. Similes, with their explicit projection of opinion, contribute to the creation of robust and subjectively charged expressions.

U	pUrO	siyAra	hE
Pronoun	ADV	NCOM	V
that	totally	fox	Is

That (person) is totally (clever like) a fox.

Kannauji showcases a diverse range of linguistic tools for constructing subjective expressions. Adjectives, intensifiers, negation, adverbs, and similes contributes to the language's ability to convey varying degrees of subjectivity. The examples presented underscore Kannauji's unique approach to articulating opinions and sentiments, enriching its linguistic landscape.

3.1.3.2.4 Khortha

Khortha, akin to the previously discussed languages, employs comparable linguistic devices to convey subjectivity. The following expressions exemplify the formation of subjective constructions through the utilization of adjectives and adverbs:

kamarvA	gaMxA	hao
NCOM	ADJ	COP
room	dirty	is

The room is dirty.

Subjective constructions in Khortha often leverage adjectives and adverbs to encode opinions and sentiments. Through the careful selection and arrangement of these linguistic elements, Khortha speakers craft expressions that carry varying degrees of subjectivity. Adjectives and adverbs contributes to the nuanced articulation of subjective content.

i	AxamI	besa	na	hao
PRONDEM	NCOM	ADV	NEG	COP
this	man	good	not	is

This man is not good.

Khortha utilizes negation as a linguistic device to shape subjective constructions. The subsequent examples illustrate the integration of negation to negate the scope of different linguistic elements, including nouns, pronouns, and adverbs. Through the strategic application of negation, Khortha speakers infuse subjectivity into their expressions, altering the overall meaning and sentiment.

hama	hiMyA	nA	aibo	kaXi
PRONPRS	PRONDEM	NEG	V	ADV
i	here	not	come	ever

I will never come here.

hama	hiMyA	Ave	ke	salAha	nAya	debo
PRONPRS	PRODEM	VN	P	N	NEG	V
i	here	come		suggest	not	would

I wouldn't suggest coming here.

hiMyA	KanavA	XyAna	se	nA	banAvo	ho
PRONDEM	NCOM	ADV	P	NEG	V	VAUX
here	food	carefully		not	cook	

(They) do not cook food here carefully.

Khortha speakers employ conjunct verbs featuring polar heads, such as adjectives or nouns, to convey opinions and subjective viewpoints. The use of a positive polar adjective, "anusaMsA," in the presented sentence, where it forms a conjunct verb in

combination with another verb (ADJ+V), exemplifies this linguistic strategy. The conjunction of a polar adjective with a verb adds a subjective dimension to the expression, allowing Khortha speakers to articulate their sentiments effectively.

hama	ekarA	anusaMsA	karabE
PRONPRS	PRONDEM	ADJ	V
i	this	recommend	will

I will recommend this.

Khortha showcases a rich repertoire of linguistic tools for expressing subjectivity. The seamless integration of adjectives, adverbs, negation, and conjunct verbs with polar heads reflects the language's versatility in capturing a spectrum of subjective nuances. Through these linguistic devices, Khortha speakers navigate subjective expression, contributing to the language's cultural and communicative richness.

3.1.3.2.5 Odia

Odia employs a variety of tools, including modifiers, negations, and polar words, to effectively encode subjectivity. The subsequent examples shed light on how adjectives and adverbs are strategically utilized to construct subjective expressions, showcasing the language's nuanced approach to conveying opinions.

Odia utilizes adjectives and adverbs as linguistic devices to articulate subjective constructions. These constructions, as depicted in the provided examples, demonstrate the language's capacity to express varying degrees of subjectivity. Notably, the second example showcases the formation of two subjective constructions within the same sentence through the combination of two adjectives connected by a conjunction. This illustrates Odia's flexibility in constructing complex subjective expressions.

KAxya	svAxiRta	nUhez
NCOM	ADJ	NEG
food	tasty	not

The food is not tasty.

wAMkara	keSa	lambA	va	camakapurNa
PRONPRS	NCOM	ADJ	CC	ADJ
her	hair	long	and	shiny

Her hair is long and shiny.

mUz	eTAkU	Kevevi	Asibi	nAhIz
PRONPRS	PRONDEM	ADV	V	NEG
i	here	ever	come	not

I will never come here.

The utilization of conjunct verbs serves as another mechanism in Odia for crafting subjective constructions. The subsequent examples illustrate how negation, when applied to conjunct verbs, contributes to the formulation of robust subjective expressions. The deliberate use of negation enhances the intensity of the subjective constructions, emphasizing the language's ability to convey nuanced sentiments.

mUz	ehAkU	sUpAriNa	Karibi
PRONPRS	PRONDEM	ADJ	V
i	this	recommend	will

I will recommend this.

mUz	eTAkU	AsivAkU	parAmarSa	debi	nAhIz
PRONPRS	PRODEM	VN	N	V	NEG
i	here	come	suggest	would	not

I would not suggest coming here.

Odia leverages polar nouns to project opinions and sentiments. The polarity of these constructions hinges on the inherent polar nature of the nouns employed in subjective expressions. The examples provided underscore how the choice of nouns influences the overall polarity of the constructions. Additionally, the incorporation of the indefinite article in the second example serves to heighten the degree of subjectivity, showcasing Odia's capacity for subtle linguistic nuances.

se	praSaMsA	pAibA	Ucita
PRONDEM	N	V	VMOD
She	appreciation	get	should

She should get appreciation.

se	samaswaMkaTArU	prema	pAibe
PRONDEM	INDF	N	V
he	from everyone	love	will get

He will get love from everyone.

Odia demonstrates use of sophisticated linguistic elements to capture subjectivity. By deftly employing modifiers, negations, adjectives, adverbs, conjunct verbs, and polar nouns, speakers of Odia use subjective expression. This linguistic versatility not only enriches the language's communicative potential but also reflects the cultural nuances embedded in Odia discourse.

3.1.3.2.6 Findings

In linguistic expression the degree of subjectivity serves as a crucial determinant, allowing us to categorize constructions into distinct classes: strong and weak. Strong Constructions (SCs) wield a more pronounced influence on overall subjectivity, whereas

Weak Constructions (WCs), while individually less impactful, gain prominence when coexisting with strong counterparts. This dynamic impact underscores the nuanced nature of subjective linguistic constructions.

Modifiers are the architects of Subjectivity: Adjectives and adverbs, operating as modifiers for nouns and verbs, stand out as pivotal carriers of subjective information. These linguistic elements, by virtue of their association with two fundamental syntactic categories, serve as ideal conduits for conveying subjective nuances. Speakers leverage adjectives and adverbs to furnish categorical descriptions, shaping the subjective essence of either the subject or object within an utterance. Furthermore, the strategic amalgamation of modifiers with other linguistic constituents offers a spectrum of subjectivity. Introducing an intensifier alongside an adjective, for instance, amplifies the degree of subjectivity, underscoring the role of modifiers in fine-tuning linguistic expressions. Notably, the conjunction of an adjective with negation initiates a transformative process, altering the polarity and culminating in the formation of robust subjective constructions.

Synergies in Subjectivity: The inherent flexibility of linguistic construction allows for the coexistence of multiple subjective expressions within a single sentence. By orchestrating various linguistic devices, speakers craft a rich articulation of subjectivity. This synergy becomes particularly evident when WCs, individually possessing lesser impact, seamlessly intertwine with their stronger counterparts, collectively contributing to the overarching subjective landscape.

3.1.3.3 LRs

In the pursuit of unravelling the intricacies of subjectivity within linguistic data, our approach delves into the formulation of LRs. This section navigates the terrain of LRs,

encapsulating the essence of subjective formations through two distinct categories: Recursive rules and Subjectivity representation rules.

At the core of our formulation are Recursive rules, strategically designed with non-terminals on the left-hand side (LHS) and terminals on the right-hand side (RHS). This deliberate arrangement facilitates their reusability, serving as foundational units that can be repeatedly employed in Subjectivity representation rules to depict various facets of subjective constructions. The utilization of = signifies equivalence, || denotes complementary distribution, and {} conveys the optionality of both terminals and non-terminals. Parentheses () are employed to group constituents, emphasizing their collective operation.

Table 3.2 lays out the recursive phrase structure rules, outlining the fundamental components that construct larger linguistic entities. Notably, NP (Noun Phrase) encompasses either a Noun (N) or Pronoun (PRON), with optional adpositions such as preposition or postposition (P). Meanwhile, VP (Verb Phrase) can manifest as a Verb (V) or a combination of VAUX (Auxiliary Verb) and V with interchangeable order. These structured rules sets the stage for the subsequent construction of LRs for subjective expressions.

Table 3.2

Recursive phrase structure rules

NP = (N || PRON) {P}
VP = V {VAUX} || VAUX V

Building upon the recursive structures, Table 3.3 delineates LRs tailored for strong subjectivity constructions. Each rule, denoted as a SC, elucidates the composition of lexical elements contributing to the encoding of subjectivity. From the combination of NP with Adjectives (ADJ), Negations (NEG), Adverbs (ADV), and Verbs (V), these rules capture the essence of strong subjectivity in diverse linguistic expressions.

Table 3.3
LRs for SCs

SC1 = NP ADJ NEG
SC2 = NP ADV {P} NP {NEG} VP
SC3 = NP (N || PRON) ADV {NEG} V
SC4 = NP ADJ V
SC5 = NP NEG V

In contrast, Table 3.4 outlines LRs tailored for weaker subjectivity constructions, labeled as WCs. Here, the rules spotlight scenarios where Negations (NEG) are coupled with Verb Phrases (VP) or where Noun Phrases (NP) engage with Adjectives (ADJ). These rules encapsulate instances where subjectivity is nuanced, providing a comprehensive spectrum that encompasses both strong and WCs.

Table 3.4
LRs for WCs

WC1 = NEG VP
WC2 = NP ADJ

Through these Recursive and Subjectivity representation rules, our methodology not only establishes a structured foundation for encoding subjectivity but also showcases the adaptability and versatility of these rules across varied linguistic contexts. This framework serves as a pivotal guide, unveiling the nuanced layers of subjectivity within linguistic constructions.

3.1.3.4 Subjectivity Index

To capture degree of subjectivity, we introduce the innovative concept of the Subjectivity Index (SI). Recognizing that subjective data encompasses linguistic constructions of

varying strengths, denoted as either strong or weak, the SI is defined through formulas (1) and (2) to provide a nuanced understanding of the subjective content.

$$SI(s) = \frac{\sum_{i=0}^{n} i + \sum_{j=0}^{m} \frac{1}{2} j}{\sum s}, \quad (1)$$

Here n denotes the count of strong subjective constructions, m denotes the count of weak subjective constructions, and Σs denotes the number of words in each sentence. In formula (1) SI is determined by the summation of the count of SCs and half of the count of WCs, divided by the total number of words in the sentence. We assign half weight to weak subjectivity constructions here, as their impact is limited when they appear in isolation. This formula captures the overall subjectivity based on the combination of SCs and WCs. For adjacent strong and weak subjective constructions with a word distance less than five, SI(s) is calculated with Formula (2), providing a refined measure for closely positioned subjectivity expressions. In such cases, both SCs and WCs receive equal weight as WCs combined with strong ones, enhance the overall intensity of subjectivity and contribute equally to its expression.

$$SI(s) = \frac{\sum_{i=0}^{n} i + \sum_{j=0}^{m} j}{\sum s}, \quad (2)$$

These formulas lay the foundation for a robust and nuanced representation of subjectivity, capturing the relation between strong and WCs within a given sentence. Building upon these formulations, we implement a LR-based FST for subjectivity classification in Hindi. The classifier operates through three main steps, as illustrated in Fig. 3.1.

1. **Tokenization and Pruning:** In the initial step, the input data undergoes tokenization and pruning, preparing it for subsequent analysis.

2. **SI Calculation:** The SI for the input data is then calculated based on the previously defined formulas (1) and (2). This step involves a granular analysis of

n-grams within the sentence, with a focus on identifying strong and WCs through LRs.

3. **Subjectivity Classification:** The calculated subjectivity score is compared against a domain ranker, determining the final subjectivity class of the input data. If the global SI surpasses a predefined threshold, the sentence is predicted as subjective; otherwise, it is classified as objective.

The accompanying algorithm outlines the step-by-step process involved in LRs-based subjectivity classification. From tokenization and pruning to the nuanced evaluation of n-grams using LRs, the algorithm provides a comprehensive methodology for subjectivity classification.

Algorithm: LRs-based subjectivity classification

Input: A sentence *s*
Output: Predicted class of *s*: Subjective or Objective

```
Pre-processor: Tokenization, Pruning
SI(s) = 0
for list(n in n-grams) in the sentence s
  for list(lexical-rules) in rule-dict
    if n-gram_i = lexical-rule_i
      if n-gram_i is a SC
        Calculating SI(s) with Formula (1)
      elif n-gram_i is a WC
        if wordDistance (n-gram_i, (n-gram_i-1 or n-gram_i+1)) < 5
          if n-gram_i-1 or n-gram_i+1is a SC
            Calculating SI(s) with Formula (2)
          end if
        end if
      else
        Calculating SI(s) with Formula (1)
      end if
    end if
  end for
  add calculated SI(s) to global SI(s)
  if global SI(s)> ranker threshold
    predict sentence is subjective
  else
    predict sentence is objective
end for
```

Fig. 3.1
The proposed algorithm for subjectivity classification

3.1.3.5 Subjectivity identifier

In our quest for unravelling subjectivity, we introduce the Subjectivity Identifier, a culmination of our subjectivity classification algorithm depicted in Fig. 3.1. This identifier is designed to take textual data as input and discern its subjectivity class, offering a pivotal tool for nuanced linguistic analysis. The high-level architecture of this identifier is illustrated in Fig. 3.2, outlining its comprehensive functionality.

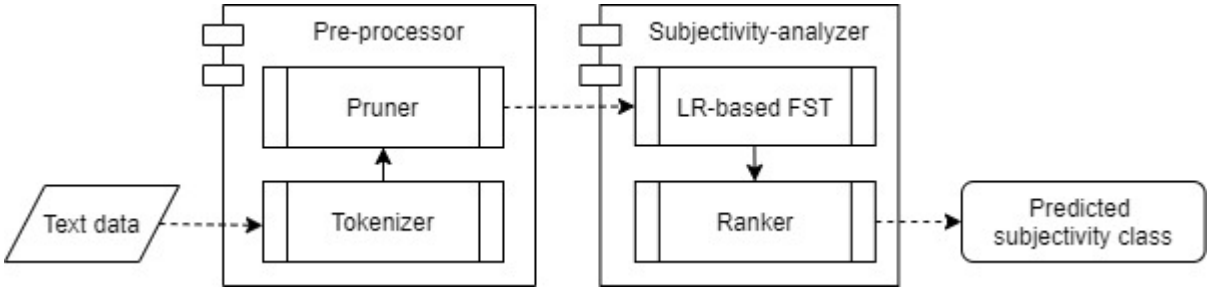

Fig. 3.2
Architecture of the proposed subjectivity identifier

This visual representation provides an overview of the architecture governing the Subjectivity Identifier. Comprising key modules, the identifier ensures a systematic and locale-agnostic approach to subjectivity analysis.

Pre-processor: The initial stage of the Subjectivity Identifier is the pre-processor module, which receives text data input. Notably, this module exhibits locale agnosticism, allowing configurable operations for locale-specific nuances. Comprising micro modules, the pre-processor undertakes structural processing to deliver relevant and unified data to the subsequent subjectivity-analyser module.

Tokenizer: Responsible for cleansing non-semantic characters, standardizing the case of utterances to lowercase, and unifying spelling variations, the tokenizer ensures a consistent input format. Its adaptability extends to supporting locale-specific features.

Pruner: Tasked with cleaning word sequences devoid of value for subsequent modules, the pruner relies on a curated stop word lexicon crafted during resource creation. This step optimizes the data flow for enhanced subjectivity analysis.

Subjectivity-analyser: Following pre-processing, the subjectivity-analyser module receives the refined data and determines its subjectivity class. The adaptability of the micro-modules within the subjectivity-analyser extends to locales exhibiting subjective constructions akin to those observed in Hindi and other languages analysed in Data Analysis section.

LR-based FST: This module incorporates Hindi-specific generative rules formulated in the LRs section, delineating linguistic devices used to convey subjective ideas. Leveraging these rules, the LR-based FST classifies data into subjective and objective categories.

Ranker: The ranker processes scores provided by the FST, employing internal logic that factors in the type of linguistic constructions, their collocations, and the domain of the given text data. This meticulous ranking enhances the precision of subjectivity classification.

In essence, the Subjectivity Identifier emerges as a sophisticated tool, adept at navigating the intricacies of linguistic subjectivity. Its modular design, combined with locale adaptability, positions it as a versatile asset for researchers and linguists seeking a deeper understanding of subjective elements within textual data.

3.1.4 Experiments and results

To scrutinize the efficiency of our subjectivity classification approach, we conduct a series of experiments on diverse datasets from different domains. This section delves into the experimental setup, presents the statistical details of the experimental corpus, and

dissects the results while conducting a defect analysis to pinpoint areas for enhancement in our proposed solution.

3.1.4.1 Experimental Setup

For our experimentation, we curate data from online news portals and blogs, tapping into the richness of online linguistic content. The pivotal task of annotating the test set and reviewing the annotated data is entrusted to five language specialists. To gauge the reliability of the annotations, we perform an annotation comparison test, yielding a Kappa score of 0.95 across all domains which shows a good consistency and agreement among annotators.

Table 3.5
Statistics of experimental corpus

Domain	Sentence count	Word count	Subjective content %	Objective content%
Entertainment	1000	12000	80	20
Lifestyle	2000	16000	75	25
Politics	2000	27000	70	30
Sports	1000	13000	80	20
Technology	1000	15000	70	30

Recognizing the imbalances in the datasets due to uneven distribution of subjective and objective content, we undertake a strategic approach. Randomly down-sampling sentences from the major class and up-sampling sentences from the minor class ensure an equitable representation of subjective and objective content, laying the groundwork for robust performance evaluation.

3.1.4.2 Results

We allocate 30% of the corpus for refining and enhancing the coverage of our LR-based FST. The remaining 70% serves as the test set. Employing Precision, Recall, Accuracy, and F1 metrics, our approach yields promising results with F1 scores ranging from 0.79 to 0.92 across various domains, culminating in an overall F1 score of 0.84. The political domain emerges victorious with the highest F1 score, while the technology domain grapples with the lowest F1 score.

In Table 3.6, the Confusion Matrix offers a comprehensive view of the subjectivity classification results across different domains. The matrix captures the number of TPs, True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) for each domain.

Table 3.6
Confusion matrix

Domain	True-positive	True-negative	False-positive	False-negative
Entertainment	219	207	31	43
Lifestyle	207	212	43	38
Politics	234	221	16	29
Sports	197	210	53	40
Technology	192	204	58	46
SUM	1049	1054	201	196

Table 3.7 further refines the evaluation by providing Precision, Recall, Accuracy, and F1 metrics for each domain. Precision, representing the ratio of Positives to the sum of TPs and FPs, varies across domains, with Politics exhibiting the highest precision at 0.9360. Recall, indicating the ratio of TPs to the sum of TPs and FNs, is noteworthy, with Sports registering 0.8312. Overall Accuracy, derived from the sum of TPs and TNs over the total positives and negatives, hovers around 0.8412. The F1 metric, a harmonic mean of Precision and Recall, consolidates the performance metrics, with Politics achieving the

highest F1 score at 0.9123. This matrix underscores the robustness of the subjectivity classification system, offering insights into its nuanced performance across diverse domains.

Table 3.7
Accuracy matrix

Domain	Precision TP / (TP + FP)	Recall TP / (TP + FN)	Accuracy (TP + TN) / (P + N)	F1 2TP / (2TP + FP + FN)
Entertainment	0.8760	0.8359	0.8520	0.8555
Lifestyle	0.8280	0.8449	0.8380	0.8364
Politics	0.9360	0.8897	0.9100	0.9123
Sports	0.7880	0.8312	0.8140	0.8090
Technology	0.7680	0.8067	0.7920	0.7869
SUM	0.8392	0.8426	0.8412	0.8409

As we scrutinize these results, we are cognizant of the evolving landscape of linguistic subjectivity. The future trajectory of our work involves expanding the coverage of lexical resources, aiming to elevate the precision of our system particularly in domains that currently present suboptimal performance. In doing so, we endeavour to create a subjectivity classification approach that not only adapts to diverse linguistic landscapes but also continuously evolves to meet the challenges presented by dynamic linguistic expressions.

3.1.4.3 Defect Analysis

Upon scrutinizing sentences categorized as FPs and FNs, we identified three overarching classes of issues: Ambiguous cases, Coverage gap, and Divergence. Ambiguous cases often represent edge scenarios where a sentence can be interpreted as either subjective or objective. The reliance of LRs on WordNet for POS tags and synonymous words in FST

parses poses a challenge in accurately determining sentiment index for lexemes not covered in WordNet. This limitation contributes to the relatively diminished performance observed in the Technology domain test set.

Moreover, our analysis revealed instances where the SI of identical linguistic constructions exhibited variation based on domain and context. These divergent cases led to misclassifications, emphasizing the need for context-aware adjustments in subjectivity identification.

3.1.5 Discussion and Conclusion

Languages employ diverse linguistic devices to encapsulate emotions and opinions, a phenomenon evident in our data analysis presented in Section 2.2. Our findings indicate that languages within a sprachbund exhibit a shared repertoire of linguistic devices for encoding subjective information. Additionally, we showcased how generative rules effectively represent subjective properties within linguistic constructions.

Given the analogous use of linguistic devices in sprachbund languages, the generative rules devised for subjective constructions can be repurposed for subjectivity identification tasks in other languages within the same sprachbund. This cross-applicability underscores the potential for resource-efficient approaches in multilingual subjectivity analysis.

Introducing the SI as a metric and implementing an LR-based FST for subjectivity classification, our system demonstrates commendable overall accuracy. Nevertheless, there remains room for enhancement, particularly in domains where performance can be bolstered through the integration of updated lexical resources into LR-based FST. As part of our future work, we aim to extend lexical resource coverage, addressing the identified defect categories and refining the precision of the system across various domains. This

iterative refinement process will contribute to the ongoing development and optimization of our subjectivity identification approach.

3.2 Sentiment Analysis through PE and ICL

This set of experiments focuses on Sentiment Analysis in low-resource languages, including Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu, employing ICL and PE. Utilizing GPT-4 and BARD LLMS, these experiments provide crucial insights for NLP applications. Effective language-specific prompts enhance adaptability across various NLP tasks, emphasizing the importance of diverse training datasets. The profound impact of ICL underscores the need for nuanced expression understanding. Language-specific accuracy variations highlight the significance of tailored approaches, especially in less common languages, urging considerations for linguistic minority contexts. Addressing challenges like data scarcity, cultural biases, and linguistic nuances, the study proposes advanced solutions, including data augmentation, adaptive fine-tuning, culture-aware embeddings, and context-aware prompts. The research also introduces new evaluation metrics for low-resource languages and emphasizes ethical considerations, promoting fairness and inclusivity in AI deployment within diverse linguistic landscapes. Overall, the findings contribute to the refinement of subjectivity analysis through LLMS, advocating for language-specific strategies and ethical AI practices.

3.2.1 Background

Subjectivity analysis, a vital component of NLP, involves unravelling sentiments, opinions, and emotions embedded in textual data. In the digital communication realm, comprehending the subjective nature of language is crucial for applications ranging from

sentiment analysis in customer reviews to identifying bias in news articles (Dwivedi & Ghosh, 2022). However, subjectivity analysis research has primarily focused on major languages with ample linguistic resources. This set of experiments addresses a significant gap by concentrating on subjectivity analysis in low-resource languages, specifically within the Indian linguistic context.

The study of subjectivity in low-resource languages holds immense importance. Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu represent the diverse linguistic tapestry of the Indian subcontinent. The linguistic diversity across these languages, coupled with the scarcity of labeled datasets, presents unique challenges for NLP. Subjectivity analysis in these languages emerges as a critical research area, offering insights into cross-cultural communication, sentiment expression, and cultural nuances within textual content.

Subjectivity analysis is pivotal for extracting meaningful insights from textual data. In sentiment analysis, discerning whether a text conveys a positive, negative, or neutral sentiment is crucial for applications like customer feedback analysis, brand perception monitoring, and product reviews. Moreover, subjectivity analysis extends beyond sentiment, encompassing the identification of opinions, emotions, and nuanced expressions that contribute to a text's overall subjective tone (Cortis Brian, 2021; Wankhade et al., 2022).

In low-resource languages, subjectivity analysis gains prominence due to the unique linguistic challenges they pose (Dwivedi & Ghosh, 2022). Understanding sentiments and opinions within a language-specific context is vital for building effective NLP applications that cater to the diverse linguistic expressions in these languages. Furthermore, subjectivity analysis forms the foundation for more granular language

understanding tasks, including context-aware machine translation, question-answering systems, and dialogue systems.

The application of LLMs to subjectivity analysis in low-resource languages holds promise for several reasons. Firstly, their pre-trained nature enables them to capture fine-grained linguistic patterns, making them adaptable to diverse linguistic contexts. Secondly, their transfer learning capabilities facilitate effective knowledge transfer from high-resource languages to low-resource languages, mitigating challenges posed by limited labelled data.

3.2.2 Related Works

The landscape of subjectivity analysis has undergone a transformative journey, evolving from rule-based systems to the contemporary dominance of advanced ML and LLMs. This literature review highlights the progression of methodologies, the impact of transformer-based models, and the recent emphasis on ICL and PE, particularly in the context of LLMs (Dong et al., 2022; P. Liu et al., 2023). Despite a wealth of research on subjectivity analysis, a notable gap exists in the context of utilizing LLMs in low-resource language settings, forming the focus of this study (W. X. Zhao et al., 2023).

Early approaches to subjectivity analysis heavily relied on rule-based systems, lexicons, and handcrafted features (Ahmad Azuraliza; Yaakub Mohd Ridzwan, 2019; Chang Hsin-Ying; Chen Long-Sheng; Chang Chia-Wei, 2020; Dwivedi & Ghosh, 2022; Kaity Vimala, 2020; Kang Seong Joon; Han Dongil, 2012; Taboada Julian; Tofiloski Milan; Voll Kimberly; Stede Manfred, 2011). However, the limited capacity of these approaches to capture complex contextual nuances prompted a paradigm shift with the advent of deep learning (Acheampong Henry; Chen Wenyu, 2021; Wadawadagi Veerappa B., 2020; Xia Yitai; Pan Xiaoting; Zhang Zuopeng Justin; An Wuyue, 2019; Yadav Dinesh Kumar,

2019). Breakthroughs included the introduction of Bidirectional Encoder Representations from Transformers (BERT) and the Generative Pre-trained Transformer (GPT) series (Vaswani et al., 2017). BERT, introduced by Devlin et al. (Devlin et al., 2019), marked a significant milestone in NLP by capturing granular linguistic dependencies through its bidirectional attention mechanism.

Advancements in the transformer architecture gave rise to models like GPT and BARD (Brown et al., 2020; Xu et al., 2023). GPT-4, the fourth iteration, presented a massive autoregressive LM with 1.8 Trillion parameters, showcasing remarkable NLU capabilities (Alec et al., 2019). However, a gap emerged in applying these models to languages with limited linguistic resources, particularly evident in low-resource languages.

Recent studies aimed to refine LLMs for specific tasks through fine-tuning and transfer learning, with RoBERTa by Liu et al. standing out (X. Liu et al., 2019). It addressed BERT's limitations and represented a significant step in enhancing models for subjectivity analysis.

While these strides paved the way for more effective subjectivity analysis, a crucial aspect remained relatively unexplored – the application of these models in low-resource languages. The linguistic diversity and data scarcity in languages like Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu present unique challenges requiring tailored solutions.

Recent research emphasizes the impact of PE and ICL on LLM performance in NLP applications (Dong et al., 2022). The proposal of XLNet by Yang et al., a generalized autoregressive pretraining method, surpassed BERT in various benchmarks and underscored the importance of capturing bidirectional context for enhanced language understanding (Yang et al., 2019).

PE involves the careful design of input prompts to guide the model's attention toward specific aspects of subjectivity. By tailoring prompts to the linguistic characteristics of each language, researchers aim to enhance the models' ability to capture subjective content effectively.

ICL, another critical aspect, involves exposing LLMs to diverse examples that encompass a wide range of linguistic expressions, idiomatic phrases, and culturally specific sentiments. This exposure to diverse contextual information during prompting aims to augment the models' understanding of nuanced expressions in low-resource languages.

While LLMs have shown remarkable performance in high-resource languages, their adaptability and effectiveness in languages with limited linguistic resources remain an open area for exploration. The study bridges this gap by shedding light on the unique challenges posed by low-resource languages and proposing tailored strategies to enhance subjectivity analysis in these linguistic contexts (Dwivedi et al., 2024).

3.2.3 Research Method

The research methodology employed in this study is tailored to analyze and address the distinctive challenges associated with subjectivity analysis in low-resource languages. Integrating innovative techniques such as PE and ICL, the study leverages the advanced capabilities of GPT-4 and BARD LLMs. The objective is to identify gaps in these models and enhance their comprehension of subjectivity by adapting approaches to the linguistic nuances of each language while overcoming challenges related to data scarcity. This section discusses the key components of the research methodology.

3.2.3.1 PE

PE serves as a pivotal component of the research methodology, with the goal of guiding models toward a nuanced understanding of subjectivity in each selected language. For this research, customized language-specific prompts are meticulously crafted to cater to the unique linguistic characteristics and cultural nuances of the experimental languages. Well-crafted prompts play a crucial role in directing the models' attention to relevant aspects of subjectivity. An example baseline English prompt for sentiment analysis is provided:

> *[Context]*
>
> *Analysing subjectivity in sentences.*
>
> *[Instructions]*
>
> *Evaluate the degree of subjectivity within the provided sentence.*
>
> *Assign a score between 0 (completely objective) and 1 (completely subjective) based on the perceived subjectivity.*
>
> *Provide an explanation for the assigned score.*
>
> *[Constraints]*
>
> *Limit the evaluation to the sentence's content, avoiding external information or context.*
>
> *Focus on linguistic features within the sentence, such as tone, sentiment, and opinion expressions.*
>
> *Ensure the assessment is language-specific, considering nuances and cultural influences.*

This baseline prompt template is adapted for each experimental language to ensure it captures language-specific expressions and cultural subtleties contributing to subjectivity. By tailoring prompts to the linguistic nuances of each language, the aim is to enhance the models' sensitivity to culturally specific sentiments and expressions, ultimately improving accuracy in subjectivity analysis.

3.2.3.2 ICL

In addition to PE, ICL is employed as a crucial technique to expose the models to diverse linguistic expressions, idiomatic phrases, and culture-specific sentiments during the prompting phase. The ICL process involves providing the LLMs with curated contrastive examples for each class (Subjective and Objective for our experiments) during prompting. This process enables the models to adapt to the intricacies of each language and task, capturing the unique linguistic nuances that influence subjectivity. Reference examples for two to ten-shot learning for each language are carefully curated to encompass the wide range of subjectivity, capturing the nuances of sentiment and opinion prevalent in diverse textual content.

The significance of ICL lies in its ability to augment the models' understanding of nuanced expressions that may not be explicitly covered during model training. This exposure to a diverse set of examples facilitates the use of existing models while making them more adept at capturing subjective content in low-resource languages. ICL enhances adaptability by providing models with a broader understanding of language-specific subjectivity, thereby contributing to more accurate and context-aware subjectivity analysis in diverse linguistic landscapes.

3.2.3.3 Language Selection

Selection of Bengali, Gujarati, Hindi, Kannada, Maithili, Marathi, Tamil, Telugu, and Urdu for this experiment is grounded in their linguistic diversity, offering a distinctive opportunity to explore subjectivity analysis in a multi-faceted linguistic landscape.

Firstly, these languages encompass diverse scripts, ranging from variations of the Brahmic script in Bengali, Gujarati, Hindi, Kannada, Marathi, Tamil, and Telugu to the Perso-Arabic script in Urdu. This script diversity introduces unique challenges, enabling

a comprehensive assessment of the model's adaptability. Additionally, these languages exhibit varying morphological complexities, from the agglutinative nature of Kannada to the fusional characteristics in Hindi. The morphological intricacies impact the expression of subjectivity, rendering this set ideal for investigating the model's ability to discern subjective nuances across different morphological structures.

Syntactic variations, manifested in distinct sentence structures across these languages, enhance the experiment's robustness. Divergent word orders, case systems, and syntactic constructions pose challenges for subjectivity analysis. Lastly, the semantic diversity arising from cultural and linguistic nuances enriches the experiment's depth. Variances in sentiment expression, idiomatic usage, and cultural connotations provide a comprehensive evaluation ground, ensuring the models' sensitivity to semantic variations. In summary, the strategic selection of these languages for experimentation capitalizes on their script, morphological, syntactic, and semantic diversity, rendering them an ideal set of languages for a comprehensive exploration of subjectivity analysis in a low-resource language setting.

3.2.3.4 Data Collection and Preparation

For evaluation, a balanced dataset comprising 1000 sentences, evenly split between 500 subjective and 500 objective sentences, is curated and manually crafted by Hindi linguists. Each sentence is accompanied by its English translation to ensure cross-lingual relevance. To validate the accuracy of cross-lingual translations, these sentences are then translated into the remaining Indian languages and back-translated to English. Three native speakers individually validate each sentence, classifying them as either 'subjective' or 'objective.' This rigorous validation process ensures the precision of each sentence's class.

The resulting dataset, featuring similar sentences across different Indian languages and English, establishes a representative benchmark for evaluating the cross-lingual performance of any subjectivity identification model. Subjective utterances within this dataset span a spectrum of subjectivity, including positive and negative sentiments, opinions, and emotionally charged expressions. This annotated dataset, crucial for both evaluation and ICL, empowers models to learn from examples of subjectivity and objectivity in each language, even in low-resource language scenarios where labeled data is scarce.

Despite the challenge posed by limited labeled data in such languages, the careful curation and creation of sentences from diverse sources such as social media, news articles, and online forums significantly contribute to mitigating this challenge. The resulting dataset not only facilitates the evaluation of subjectivity analysis models but also serves as a valuable resource for training and enhancing the adaptability of these models in low-resource language settings.

3.2.3.5 Model Selection

The decision to employ GPT-4 and BARD LLMs as the primary tools for this investigation is rooted in their cutting-edge capabilities in language understanding. Harnessing these advanced models not only addresses the void in subjectivity analysis for low-resource languages but also contributes to a broader comprehension of languages within diverse cultural contexts.

GPT-4 and BARD, epitomizing LLMs, have showcased unparalleled performance in comprehending and generating human-like text as part of their emergent abilities. GPT-4, with its augmented model size and extensive training data, manifests superior contextual understanding and has exhibited remarkable results across various language

tasks. BARD, designed with bidirectional autoregressive decoding, complements GPT-4 in our experimental setup by offering an alternative approach to language modeling and understanding. Both models capitalize on the transformer architecture, now a standard for numerous NLP tasks.

3.2.3.6 Experimental Setup

The experimental setup entails the utilization of GPT-4 and BARD for subjectivity analysis in both English and selected Indian languages. Subjectivity analysis is conducted through PE for one experiment and ICL via selective examples for another. Given their substantial model size and exposure to multilingual training data, GPT-4 and BARD are expected to demonstrate improved contextual understanding when combined with PE and ICL.

The models remain unfine-tuned, with hyperparameter configuration omitted for simplicity. Instead, subjectivity analysis is executed by configuring the models through system prompts. The emphasis is on leveraging widely available PE and ICL techniques without delving into the complexities of fine-tuning and hyperparameter optimization. This streamlined approach seeks to evaluate the models' performance in subjectivity analysis across diverse linguistic landscapes without introducing unnecessary intricacies.

3.2.4 Results

The results section provides an in-depth analysis of subjectivity identification experiments conducted using GPT-4 and BARD models across nine low-resource Indian languages. The primary aim is to assess and compare the models' performance in handling subjectivity, with a focus on adaptability to diverse linguistic contexts. Quantitative

metrics, including precision, recall, and F1 scores, are employed to offer a detailed understanding of the models' ability to distinguish between subjective and objective data. A comparative analysis of performance metrics reveals nuanced differences between GPT-4 and BARD models for each language. These metrics demonstrate the models' effectiveness in discerning subjective content, and the variations across languages underscore the necessity of considering multiple metrics for a comprehensive evaluation of model performance.

In Table 3.8, we present accuracy scores for PE experiments, comparing baseline English prompts with native language prompts. The results show improvement in accuracy ranging from 100 to 700 basis points across languages. Notably, GPT-4's model for Telugu exhibits the most substantial improvement across metrics, while Bengali, Maithili, and Urdu demonstrate comparatively lower accuracy improvement.

Table 3.8
English and Native Language Prompt accuracy Scores

Language	Model	Baseline English Prompt			Native Language Prompt		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Bengali	GPT-4	0.81	0.80	0.80	0.82	0.83	0.82
	BARD	0.78	0.80	0.79	0.79	0.82	0.80
Gujarati	GPT-4	0.75	0.70	0.72	0.76	0.74	0.75
	BARD	0.76	0.79	0.77	0.78	0.82	0.80
Hindi	GPT-4	0.82	0.80	0.81	0.84	0.82	0.83
	BARD	0.79	0.83	0.81	0.83	0.85	0.84
Kannada	GPT-4	0.76	0.72	0.74	0.78	0.74	0.76
	BARD	0.71	0.73	0.72	0.72	0.76	0.74
Maithili	GPT-4	0.62	0.58	0.60	0.64	0.59	0.61
	BARD	0.60	0.55	0.57	0.63	0.57	0.60
Marathi	GPT-4	0.75	0.78	0.76	0.77	0.79	0.78
	BARD	0.81	0.80	0.80	0.82	0.83	0.82
Tamil	GPT-4	0.73	0.78	0.75	0.75	0.81	0.78
	BARD	0.81	0.84	0.82	0.83	0.85	0.84
Telugu	GPT-4	0.79	0.72	0.75	0.80	0.83	0.81
	BARD	0.81	0.83	0.82	0.83	0.84	0.83
Urdu	GPT-4	0.78	0.76	0.77	0.79	0.78	0.78
	BARD	0.77	0.79	0.78	0.78	0.81	0.79

Table 3.9 presents accuracy scores for different ICL experiments, ranging from 0-shot to 10-shot. The 0-shot experiment serves as a baseline, indicating the initial performance of the models. Generally, we observe an improvement across metrics with an increase in the number of shots for both GPT-4 and BARD models across languages.

Table 3.9
Comparison of 0 to 10-shot accuracy results for ICL experiments

		0-Shot			2-Shot			5-Shot			10-Shot		
Language	Model	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
	GPT-4	0.81	0.8	0.8	0.82	0.82	0.82	0.83	0.84	0.83	0.8	0.78	0.79
Bengali	BARD	0.78	0.8	0.79	0.79	0.81	0.8	0.81	0.82	0.81	0.77	0.76	0.76
	GPT-4	0.75	0.7	0.72	0.77	0.78	0.78	0.74	0.76	0.75	0.71	0.7	0.7
Gujarati	BARD	0.76	0.79	0.77	0.78	0.8	0.79	0.75	0.78	0.76	0.68	0.67	0.67
	GPT-4	0.82	0.8	0.81	0.83	0.84	0.83	0.85	0.86	0.85	0.8	0.78	0.79
Hindi	BARD	0.79	0.83	0.81	0.81	0.82	0.81	0.84	0.85	0.84	0.78	0.77	0.77
	GPT-4	0.76	0.72	0.74	0.78	0.76	0.77	0.75	0.74	0.74	0.69	0.67	0.68
Kannada	BARD	0.71	0.73	0.72	0.74	0.75	0.75	0.72	0.71	0.71	0.66	0.65	0.65
	GPT-4	0.62	0.58	0.6	0.64	0.61	0.62	0.6	0.57	0.58	0.57	0.54	0.55
Maithili	BARD	0.6	0.55	0.57	0.63	0.6	0.61	0.59	0.56	0.57	0.54	0.52	0.53
	GPT-4	0.75	0.78	0.76	0.77	0.8	0.78	0.79	0.82	0.8	0.77	0.75	0.76
Marathi	BARD	0.81	0.8	0.8	0.79	0.81	0.8	0.8	0.83	0.82	0.76	0.74	0.75
	GPT-4	0.73	0.78	0.75	0.75	0.8	0.77	0.78	0.83	0.8	0.76	0.79	0.77
Tamil	BARD	0.81	0.84	0.82	0.78	0.82	0.8	0.8	0.84	0.82	0.73	0.76	0.74
	GPT-4	0.79	0.72	0.75	0.8	0.82	0.81	0.82	0.84	0.83	0.8	0.78	0.79
Telugu	BARD	0.81	0.83	0.82	0.82	0.85	0.83	0.83	0.85	0.84	0.78	0.76	0.77
	GPT-4	0.78	0.76	0.77	0.79	0.78	0.78	0.78	0.77	0.77	0.76	0.74	0.75
Urdu	BARD	0.77	0.79	0.78	0.78	0.8	0.79	0.76	0.75	0.75	0.75	0.73	0.74

In Bengali, GPT-4's F1 score increases from 0.80 in the 0-shot to 0.83 in the 5-shot experiment, demonstrating the positive impact of ICL. However, for the 10-shot experiment, there is a slight degradation in performance to 0.79. Similar patterns emerge

across languages, but the performance trends vary, with some degradation observed in Tamil even for the 2-shot experiment.

3.2.5 Discussion

This section delves into the outcomes derived from experimental results, the encountered challenges, mitigation strategies, and potential avenues for future research.

3.2.5.1 Findings

The performance comparison of models across languages highlights their adaptability to diverse linguistic contexts, accentuating the importance of nuanced performance metrics. While GPT-4 and BARD demonstrate robust performance in widely spoken languages like Hindi and Bengali, a potential performance degradation is evident in less common languages like Maithili and Urdu. This decline is reflected in lower precision, recall, and F1 scores, attributed to several factors:

1. **Limited Training Data:** Languages with scant resources often suffer from a dearth of training data, hindering models from accurately capturing unique linguistic nuances in less common languages.
2. **Cultural Nuances:** Subjectivity, influenced by cultural nuances, may not be adequately represented in the training data of less common languages, impacting the models' understanding of subjectivity.
3. **Model Bias:** Pre-training on large datasets, dominated by major languages, introduces biases affecting model performance on less common languages. Incomplete mitigation of biases by ICL with limited data is observed.

The efficacy of language-specific prompts emerges as a key finding, highlighting their crucial role in enhancing adaptability across diverse linguistic landscapes. This suggests broader applications beyond subjectivity analysis, emphasizing the potential of localized and customized prompts in various NLP tasks. Furthermore, the impact of ICL underscores the necessity of diverse examples to expose models to varied linguistic expressions and cultural sentiments, emphasizing the significance of incorporating nuanced language features during inference to improve model performance.

The challenges faced during experiments include the observed performance degradation in less common languages, highlighting the impact of limited training data, cultural nuances, and model bias. These challenges underscore the need for targeted strategies to address issues related to linguistic minority contexts, such as data scarcity, cultural biases, and training challenges. Additionally, the necessity of developing nuanced approaches for each language is emphasized, recognizing the diverse linguistic landscapes that GPT-4 and BARD models encounter.

Mitigating the challenges involves developing strategies tailored to each language. Addressing limited training data requires efforts to enhance data availability and diversity in less common languages. Incorporating cultural nuances in training data and refining ICL methods can improve models' sensitivity to diverse linguistic expressions. Tackling model bias entails exploring ways to minimize biases introduced during pre-training on large datasets, ensuring fair and equitable performance across languages.

3.2.5.2 Challenges Encountered and Addressed

The research journey encountered formidable challenges, each demanding nuanced solutions for a robust outcome. The scarcity of labelled data in less common languages posed a significant hurdle, challenging the models' capacity to generalize effectively. To

address this, diverse data from sources like social media, news articles, and online forums were curated and handcrafted. While this approach helped mitigate scarcity for evaluation and ICL to some extent, future research should explore innovative data augmentation techniques to enhance datasets in low-resource languages.

Cultural nuances and biases, inherent in the pre-training phase on large datasets, emerged as another significant challenge, particularly impacting model performance in less common languages during evaluation. The response to this challenge involves a multifaceted approach. Incorporating cultural examples, ICL, and prompting strategies tailored to less common languages helped mitigate this challenge. This approach seeks not only to address biases but also to enhance the models' sensitivity to diverse cultural expressions. By acknowledging and navigating these challenges, the research contributes to the ongoing refinement of methodologies in multilingual NLP use cases, paving the way for more equitable and effective LMs.

3.2.5.3 Significance of Multilingual Subjectivity Analysis

This study's deliberate focus on multilingual subjectivity analysis holds paramount significance in shaping the broader discourse on LM applicability within diverse linguistic landscapes. Beyond its implications for sentiment analysis, the insights garnered carry weight in critical domains such as content moderation, opinion mining, and personalized content recommendation. This emphasis on multilingual subjectivity analysis not only refines our understanding of LMs' adaptability but also underscores their potential impact across a spectrum of real-world applications.

3.2.5.4 Potential Avenues for Future Research

In envisioning the future trajectory of this research, several key directions emerge to fortify the robustness and inclusivity of subjectivity analysis and other NLP applications in low-resource languages. A pivotal direction is the investigation and crafting of advanced data augmentation techniques tailored to less common languages. This approach has the potential to significantly enhance model generalization, with techniques such as cross-lingual data synthesis and transfer learning warranting exploration to enrich training datasets. These endeavours offer a promising pathway toward more comprehensive LM adaptation.

Moreover, a crucial focus is on developing fine-tuning strategies explicitly designed for the challenges posed by limited linguistic resources. Adaptive fine-tuning approaches, dynamically adjusting to the linguistic characteristics of each language, present a promising avenue for refining model performance. Integrating cultural embeddings and context-aware prompts into the training process emerges as another critical direction. This integration aims to heighten models' sensitivity to cultural nuances, fostering more accurate subjectivity analysis in low-resource languages.

Additionally, there is a pressing need for the development of new evaluation metrics specifically tailored to the unique challenges of low-resource languages. Current metrics may fall short in capturing the intricacies of linguistic diversity and scarcity of labeled data in these languages. Creating evaluation frameworks that align with the specific linguistic nuances of each language will provide more accurate assessments of model performance in diverse linguistic landscapes.

Lastly, ethical and responsible AI considerations must remain integral to the discourse. Ensuring fairness, transparency, and inclusivity in NLP research is paramount, with a particular emphasis on addressing biases in both training data and model predictions. This

comprehensive approach seeks to prevent perpetuating inequities, particularly in minority language communities, marking a crucial stride towards ethical and equitable LM development. By embracing these multidimensional directions, the research can contribute significantly to advancing the field of NLP in low-resource languages while upholding ethical standards and promoting inclusivity.

3.2.6 Conclusion

The culmination of this research presents a nuanced understanding of subjectivity analysis in low-resource Indian languages, viewed through the lens of LLMs. By delving into PE, ICL, and the adaptability of models across linguistic diversity, this study offers valuable insights for NLP and establishes critical considerations for future research.

Revelations from the research findings illuminate key aspects shaping the landscape of subjectivity analysis in low-resource languages. The efficacy of language-specific prompts, demonstrated by the models' accurate interpretation of subjective expressions, underscores the significance of tailored approaches. ICL emerges as a pivotal factor, emphasizing the imperative need for diverse training datasets that expose models to the varied linguistic expressions prevalent in different languages. Language-specific performance variations underscore the uniqueness of each language and the necessity of adapting models to capture distinct linguistic nuances. Concurrently, the observed degraded performance in less common languages sheds light on challenges related to data scarcity, cultural nuances, and biases introduced during pre-training.

3.2.6.1 Contributions to NLP Research

The contributions of this research transcend the domain of subjectivity analysis, impacting broader NLP research. Insights gained from experiments with GPT-4 and

BARD models provide a blueprint for developing LMs that are more inclusive and adaptable to diverse linguistic contexts. Although the research primarily focuses on Indian languages, the methodologies and insights can be extended to other low-resource languages globally. The principles of PE and ICL, being language-agnostic, offer a foundation for adapting LMs to diverse linguistic environments.

3.2.6.2 Implications for Multilingual NLP Applications

Understanding subjectivity across languages is pivotal for deploying NLP applications successfully in multicultural and multilingual settings. The adaptability of GPT-4 and BARD to different linguistic nuances signifies progress in achieving effective communication in a globalized digital landscape. From sentiment analysis in customer reviews to opinion mining in social media, the subjectivity analysis capabilities of these models have far-reaching implications.

Challenges encountered during the research, such as data scarcity and biases, highlight the ethical dimensions of developing and deploying LLMs. The responsible development of AI technologies demands careful consideration of potential biases and the impact of these models on linguistic minority communities. Ethical considerations, transparency, and fairness should be integral to the development and deployment of LMs, particularly in low-resource language settings. This holistic approach ensures that AI technologies not only advance linguistic understanding but also uphold ethical standards and foster inclusivity.

3.2.6.3 Future Directions and Research Avenues

This research paves the way for several promising avenues for future exploration. Tailored data augmentation techniques designed for less common languages hold the

potential to significantly improve the generalization capabilities of LMs. The development of fine-tuning strategies explicitly addressing the challenges posed by limited linguistic resources is a crucial area for further investigation. Furthermore, the incorporation of cultural embeddings and context-aware prompts into the training processes constitutes a critical approach for enhancing the models' sensitivity to cultural nuances, an aspect examined and advocated in subsequent chapters, thereby facilitating more precise and contextually grounded subjectivity analysis.

3.2.6.4 Limitations and Considerations for Improvement

Recognizing the study's limitations is vital for fostering continuous improvement in NLP research. Challenges associated with less common languages, including limited availability of labelled data and cultural biases, signal areas where future research can make substantial contributions. Exploring innovative strategies to overcome these challenges is essential for the development of more robust and unbiased LMs.

This research significantly contributes to the ongoing evolution of LMs, particularly in the era of large pre-trained models. As LMs advance, understanding their adaptability to linguistic diversity becomes increasingly pertinent. The study offers a snapshot of the current state of subjectivity analysis in low-resource languages, acting as a catalyst for future advancements in the field.

The societal impact of NLP advancements, especially in low-resource language settings, is profound. These advancements can empower communities with limited linguistic representation, fostering inclusivity in digital communication. The applications of improved subjectivity analysis extend to various domains, positively impacting education, healthcare, governance, and community engagement. The research

underscores the transformative potential of LMs in enhancing the lives of individuals and communities.

In conclusion, this research not only advances our understanding of subjectivity analysis in low-resource languages but also lays a foundation for the responsible development of LMs. The relation of PE, ICL, and the models' performance across diverse linguistic contexts throws light on the complexities of language understanding. As the field progresses, continuous collaboration between researchers, developers, and communities is essential to ensure that LMs evolve ethically, inclusively, and responsibly. This holistic approach is crucial for the sustainable and equitable development of language technologies, aligning with the evolving needs of diverse linguistic communities worldwide.