Chapter 1

Introduction

"What do the customers think about this mobile?" The opinions of customers have consistently been a significant source of information. For example:
Which mobile should I buy ?

The opinion of other users will be important to buy a mobile. Such opinions can be collected from pre-web sources i.e. sources that existed even before the Internet came, such as friends and relatives, associates or customer-feedback, and from sources from internet i.e. blogs, e-commerce sites, review sites and discussion forums. It is critical to find and comprehend viewpoints from user-generated information. Analyzing people's attitudes toward particular entities from text documents is the task of sentiment analysis. Consider the sentence-
"The fried rice is amazing here."
The review above is about *rice*, and it expresses *positive* emotion or sentiment.
A vast amount of reviews may be found on post web sources because of rapid expansion of digital media and information technology advancements. It is impossible to manually interpret the details supplied as perspectives due to the large volume of text contents. As a result, it is required to create an automatic computational framework for assessing opinions concealed behind unstructured language.

On the Internet, people can share their thoughts through text-based platforms. These are social media [33] and online reviews. There are now a huge

number of these online reports being written every day. English is an international language, but more and more Indian users have made internet reviews and opinions available in their own language in recent years. Finally, the end sentiment can not be judged by English alone; this shows how important it is to work on sentiment analysis in Indian languages as well.

One popular Indian language [99] that is spoken by many is Hindi. There are also a lot of Hindi web-pages. There are a lot of websites that offer knowledge in Hindi. People also say what they think and review things in Hindi. Because of this, it is also important to measure how people feel from reports written in Hindi. This study will help with mood analysis for the Hindi language. Here are some examples of reviews written in Hindi:

1. 11.6 इंच के लैपटॉप की सबसे अहम खासियत इसका कॉम्पेक्ट होना है।

   source- https://hindi.gadgets360.com/pc-laptop

   Above review is about *laptop* and sentiment expressed is *positive*.

2. फिल्म एक बढ़िया एंटरटेनर है, जो आपको अंत तक बांधे रखती हैं।

   source- https://www.livehindustan.com/entertainment/

   Above review is about *movie* and sentiment expressed is *positive*.

3. इसके पहले खेले गए पहले एकदिवसीय मुक़ाबले में तो दोनों टीमों के गेंदबाज़ विकेट के मिज़ाज को पढ़ ही नहीं सके।

   source- https://www.bbc.com/hindi/topics/cwr9j8g1kj9t

   Above review is about *game* and sentiment expressed is *negative*.

1.1 Levels of Sentiment Analysis

Sentiment Analysis(SA) is mainly of three types: document-level [15], sentence-level [19] and aspect-level sentiment analysis[96].

1. **Document-Level Sentiment Analysis** Objective of **document-level SA** is to determine a document's total opinion, which typically shows a distinct viewpoint on a subject. A *positive/negative/neutral/conflict* polarity is assigned to a whole document. A document is a collection of sentences. A review in the example below depicts document-level SA.

   Example

   "The camera quality of the phone is excellent. However, the battery life is terrible. I love the sleek design of the phone."

   Document-level classification is most effective when the document is authored by a solitary individual and conveys a subjective viewpoint or attitude about a singular entity.

2. **Sentence-Level Sentiment Analysis** In ***Sentence-level SA*** restricts the analysis to sentences. These sentences could belong to documents or single reviews. The following reviews are examples that depict Sentence-level SA.

   Examples

   "The camera quality of the phone is excellent. - Positive"

   "However, the battery life is terrible. - Negative"

   "I love the sleek design of the phone. - Positive"

   While both document and sentence-level sentiment analysis are useful, both approaches provide an *overall* sentiment orientation.

3. **Aspect-Level Sentiment Analysis** Aspect-level sentiment analysis (ABSA), involves doing a detailed examination of the input and changes its

attention from a full phrase or document to particular entity or particular aspects of an entity. The phrase is partitioned and its opinions are determined depending on its aspects. For instance, within the realm of e-commerce, a target refers to a certain product, while its characteristics such as cost and dimensions, are the specific aspects of that product. Table 1.1 shows the aspect-based sentiment analysis for following sentences:

(a) *"The camera quality of the phone is excellent. - Positive"*

(b) *"However, the battery life is terrible. - Negative"*

(c) *"I love the sleek design of the phone. - Positive"*

Table 1.1: ABSA

Aspect Term	camera quality of phone	battery life	phone
Aspect Category	Accessory	Accessory	Electronics
Opinion Term	excellent	terrible	love
Sentiment Polarity	Positive	Negative	Positive

In above examples *camera quality of phone*, *battery life* and *phone* are aspects and excellent, terrible and love are opinions.

Aspect-Based SA discovers, what people like, what they do not like at a fine-grain level. It is easier to comprehend the sentiment when opinion terms are linked to aspect terms.

The fundamental sentiment analysis problem has two essential elements: target and opinion. The target can refer to any entity or specific component of the entity. Sentiment can be categorized as positive, negative, neutral or conflicting opinions regarding the target. ABSA consists of the following four elements:

(a) An **aspect term** [10] refers to the specific target of an opinion that is expressly mentioned in a given text, such as the word *"camera quality of*

phone" in the sentence *"The camera quality of the phone is excellent."* In cases where the target is implicit represented (for example, when it is referred to as "It is excellent."), the aspect word is indicated as a distinct entity called "null".

(b) **Aspect category** [24] defines a specific classification for an aspect. Each specific subject of interest has its own specified classifications. Accessory and service may be aspect categories within the mobile domain.

(c) **Opinion term** [117] is used to explicitly communicate feeling towards the aspect term. In the sentence *"The camera quality of the phone is excellent."*, the term "excellent" expresses an opinion about the aspect term "camera quality of the phone".

(d) **Sentiment polarity** [31] relates to the way taken by the opinion towards an aspect term, typically categorized as *positive*, *negative*, *neutral* or *conflict*.

In ABSA, aspect of a target is classified into an aspect category and an opinion word into a sentiment polarity.

1.2 Approaches to ABSA

ABSA entails the process of recognizing and extracting feelings related to particular elements or characteristics of entities from textual data. Building an ABSA system for Hindi requires the following tasks.

1. **Data collection and pre-processing**

   Data collection

   - **Source identification:** Identify sources where Hindi text is abun-

dantly available, such as product reviews on e-commerce sites, movie reviews, social media platforms (Twitter, Facebook), news articles, and forums.

- **Web scraping:** Use web scraping tools like BeautifulSoup [28], Scrapy [68], or Selenium [93] to gather text data from these sources.
- **Manual annotation:** Prepare a subset of the data for manual annotation. Involve native Hindi speakers to annotate aspect terms and corresponding sentiments to create a gold standard dataset [45] like IITP-I[2] and IITP-II[3].

Pre-processing

- **Normalization:** Remove punctuation, special characters and normalize [7] elongated words (e.g., "बढिया" to "बढ़िया").
- **Tokenization:** Use a Hindi tokenizer [107] to split sentences into words. Libraries like Indic NLP [9] Library and spaCy [110] Hindi language model can be helpful. One example sentence is tokenized with spaCy model as follows:

  Input sentence: यह एक साधारण वाक्य है।

  This is a simple sentence.

  Output token list: ['यह', 'एक', 'साधारण', 'वाक्य', 'है', '।']
- **Stopword removal:** Remove common Hindi stopwords that do not contribute to sentiment analysis. Following example shows the stopword removal process:

  Input tokens: ['यह', 'एक', 'साधारण', 'वाक्य', 'है', '।']

  Filtered tokens: ['साधारण', 'वाक्य', '।']

- **Stemming/Lemmatization:** Reduce words to their root forms using Hindi stemmers [79] or lemmatizers [88]. Here is an example of stemming and lemmatization for a Hindi sentence:

Input: किताबें पढ़ना अच्छा होता है।

It is good to read books.

Stemmed tokens: ['किताब', 'पढ़', 'अच्छा', 'हो', 'है']

Lemmatized tokens: ['किताब', 'पढ़ना', 'अच्छा', 'होता', 'है']

Stemming reduces words to their base form, sometimes removing suffixes, while lemmatization returns the dictionary form of the words.

2. **Aspect Term Extraction (ATE)**

   Rule-based methods

   - **Pattern matching:** Create rules to identify patterns where aspect terms are likely to appear, such as "X is good" or "I like Y".
   - **Part of Speech(POS) tagging:** Utilize POS tagging [66] to detect nouns and noun phrases, which frequently correspond to aspect terms. Tools like spaCy or the Stanford NLP [95] POS-tagger can be adapted for Hindi. POS tagging for one example sentence using spaCy is illustrated below:

     Input: राम स्कूल जा रहा है।

     Ram is going to school.

     Output:

     राम: PROPN

     स्कूल: NOUN

     जा: VERB

     रहा: AUX

     है: AUX

     ।: PUNCT

     राम (Ram) is identified as a Proper Noun (PROPN). स्कूल (school) is identified as a Noun (NOUN). जा (going) is identified as a Verb (VERB). रहा (is) is identified as an Auxiliary Verb (AUX). है(is) is identified as an Auxiliary Verb (AUX). । (.) is identified as Punctua-

tion (PUNCT).

Machine Learning methods

- **Sequence labeling:** Use sequence labeling models like Conditional Random Field(CRF) [105] or BiLSTM-CRF [61] to identify aspect terms in sentences. These models consider the context of words, making them effective for extraction tasks. Sequence labeling for one example sentence is illustrated below:

  Input: राम स्कूल जा रहा है।

  Output:

  राम: **B-PER** (Beginning of Person)

  स्कूल: **B-LOC** (Beginning of Location)

  जा: **O** (Indicates words that do not belong to any specific entity)

  रहा: **O** (Indicates words that do not belong to any specific entity)

  है: **O** (Indicates words that do not belong to any specific entity)

- **Name Entity Recognition(NER) models:** Fine-tune NER models [115] are used to recognize aspect terms. Pre-trained models can be adapted for this task using transfer learning [116]. NER labels for one example sentence are illustrated below:

  Input: राम स्कूल जा रहा है।

  Output:

  राम: **PERSON**

  स्कूल: **LOCATION**

  NER model typically categorize entities into predefined classes, such as PERSON, ORGANIZATION, LOCATION etc.

  Rule-based methods rely on specific domain knowledge, Machine Learning methods learn by identifying patterns in the data.

3. **Sentiment Classification**

Lexicon-based methods

- **Sentiment lexicons:** Utilize Hindi sentiment lexicons [59] such as SentiWordNet [54] for Hindi to allocate sentiment ratings to aspect phrases. These lexicons are made up of words with sentiment polarity annotations. An sentiment lexicon based analysis for example sentence is illustrated below:

  Input: राम स्कूल जा रहा है।

  Breakdown of sentence

  following are the aspect phrases:

  (a) राम - referring to the person.

  (b) स्कूल - referring to the location or context of the action (going).

  Sentiment lexicon lookup

  Using a Hindi sentiment lexicon, following are sentiment ratings for the key words in the sentence:

  (a) राम: Neutral or possibly positive, depending on context. (e.g., sentiment score: 0.1)

  (b) स्कूल: Generally positive, as it usually connotes a place of learning and growth. (e.g., sentiment score: +0.5)

  (c) जा रहा है।: Neutral in terms of sentiment, as it describes an action without an emotional connotation. (e.g., sentiment score: 0)

  Sentiment allocation

  (a) राम: 0.1 (neutral/positive)

  (b) स्कूल:+0.5 (positive)

  (c) जा रहा है।: 0 (neutral)

  Overall sentiment score calculation

  Overall Sentiment: (0.1 + 0.5 + 0) / 3 = +0.2 (which indicates a slightly positive sentiment overall).

Machine Learning Methods

- **Traditional models:** Train classifiers such as Support Vector Machine(SVM) [76], Naive Bayes(NB) [92] or Random Field(RF) [34] on features like n-grams [112], TF-IDF [106] and sentiment lexicon scores [30].
- **Feature engineering:** Retrieve features like word n-grams, character n-grams, TF-IDF scores, and POS tags.
- **Cross-validation:** Employ k-fold cross-validation to ensure the resilience of models and mitigate the risk of over-fitting [40].

Sentiment prediction with SVM classifier for example sentence "राम स्कूल जा रहा है।" is as follows. The sentence would first be vectorized using techniques like TF-IDF or Count Vectorization. Afterward, the trained SVM model would classify the sentiment, which, in this neutral context, might predict "*neutral*" sentiment depending on the training data.

Deep learning models

- **Long-short term memory(LSTM)/BiLSTM:** Use LSTM [44] or BiLSTM [125] networks to capture the context and sequential nature of text.
- **Convolutional Neural Networks(CNN):** Use CNN [72] to capture local patterns in text, which can be useful for sentiment classification.
- **Attention mechanisms:** Integrate attention processes [81] to specifically target significant words and elements within a sentence.

Sentiment prediction using LSTM model for example sentence "राम स्कूल जा रहा है।" is as follows. The sentence would first be tokenized and converted into sequences. After passing through the LSTM layers, the model would likely predict a "*neutral*" sentiment.

Transformer-based models

- **Pre-trained models:** Fine-tune transformer models [47] like mBERT, IndicBERT or MuRIL for aspect-specific sentiment classification. These models have been trained in advance using extensive multilingual datasets [85] and are capable of effectively processing Hindi text.
- **Fine-tuning:** Fine-tune these models on the annotated dataset for aspect-specific sentiment analysis.

To predict the sentiment for example sentence "राम स्कूल जा रहा है।" using a fine-tuned model like mBERT, the sentence would be tokenized and processed through mBERT's transformer layers. The fine-tuned mBERT model would likely classify the sentiment as "*neutral*", assuming the model has been fine-tuned for sentiment analysis on similar Hindi-language data.

4. **Aspect-Sentiment Pair Extraction**

   - **Pipeline approach:** Integrate Aspect Term Extraction(ATE) with sentiment categorization into a single pipeline. Extracting aspect terms from sentences is the initial stage, and then proceed to classify the sentiment for each extracted aspect.
   - **Joint models:** Develop joint models that perform ATE and Aspect Sentiment Classification(ASC) simultaneously. These models can be implemented using multi-task learning frameworks.

5. **Evaluation**

   - **Metrics:** Utilize metrics [80] like precision, recall, F1-score and accuracy to assess the performance. It is essential to conduct distinct assessments for aspect term extraction and sentiment classification.
   - **Confusion matrix:** Examine confusion matrices to gain insight into the specific types of mistakes made by the model.
   - **Error analysis:** Perform detailed error analysis to identify and address common failure points.

6. **Post-processing and integration**

    - **Aggregation:** Aggregate aspect-sentiment pairs to provide an overall sentiment summary for each entity. For example, aggregate sentiments about different features of a product.
    - **Visualization:** Use visualization tools [23] like matplotlib, seaborn or Plotly to present the results. Visualizations can include aspect-based sentiment heatmaps, word clouds and sentiment trend graphs.

1.3 Evaluation Metrics

Confusion matrix is utilized to compute evaluation measures. A confusion matrix is a tabular representation that provides a concise summary of how well a machine learning model performs on a certain set of test data. It is a method of visually representing the number of correct and incorrect occurrences based on the predictions made by the model. It is commonly employed to assess the effectiveness of classification models, which seek to forecast a category label for each input instance. The machine learning model's output shows the count of instances generated on the test data.

A confusion matrix has two dimensions: actual and predicted, and it is divided into four quadrants:

- **True Positive (TP):** The situations in which the model accurately predicts the positive class.
- **True Negative (TN):** The situations in which the model accurately predicts the negative class.
- **False Positive (FP):** The situations in which the model forecasts the positive class inaccurately.

- **False Negative (FN):** the situations in which the model predicts the negative class inaccurately.

Table 1.2 is typical representation of a confusion matrix:

Table 1.2: Confusion matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	TN	FP
Actual Positive	FN	TP

1. **Precision(P):** is a way to measure how well a model can make good predictions. In other words, precision is ratio of correct positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

2. **Recall(R):** checks how well a classification model can find all the important examples in a collection. Recall is ratio of true positive predictions to true positive and false negative predictions.

$$Recall = \frac{TP}{TP + FN}$$

3. **F1-score:** The F1-score is the weighted average of Precision and Recall. Hence, when computing this score, both incorrect positive and incorrect negative results are taken into account.

$$F1 - score = \frac{2 * R * P}{R + P}$$

4. **Accuracy(A):** is a metric utilized to evaluate the efficacy of the model. The ratio of total correct instances to the all instances is referred to as the accuracy.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

1.4 Applications of ABSA in Hindi

While much of the work in ABSA has been focused on English, there's growing interest and application in other languages, including Hindi. Here are some key applications of ABSA in Hindi:

1. *Social media monitoring:* Businesses monitor social media platforms to understand consumer sentiment about their brand and products, helping them manage their reputation and make informed marketing decisions. E-commerce platforms like Amazon and Flipkart, which operate in India, analyze customer reviews written in Hindi to extract sentiments and understand customer satisfaction and issues.

   Service-oriented businesses like restaurants, hotels, and telecom companies use sentiment analysis to analyze feedback from Hindi-speaking customers to improve their services.

2. *News and media analysis:* News organizations and analysts use sentiment analysis to understand the public mood regarding current events and issues reported in Hindi newspapers, websites, and TV channels.

   By analyzing the sentiment of articles and reader comments, media companies can identify emerging trends and topics of interest in the Hindi-speaking population.

3. *Political SA:* Companies and political organizations use ABSA to gauge public opinion about products, services or political events by analyzing tweets, Facebook posts, and comments written in Hindi. Political parties analyze sentiments in social media posts, speeches, and public comments in Hindi to gauge public opinion and adjust their strategies during elections.
   Government agencies use sentiment analysis to understand public opinion on new policies and initiatives by analyzing responses in Hindi.

4. *Healthcare:* Hospitals and healthcare providers analyze feedback from patients in Hindi to improve healthcare services and patient experience.
   Sentiment analysis of social media and forum posts in Hindi can help identify mental health issues and trends among the Hindi-speaking population.

5. *Entertainment industry:* Film producers and marketers analyze reviews and social media posts in Hindi to understand audience reactions to movies, TV shows, and web series. ABSA helps track public sentiment about songs, albums, and celebrities among Hindi-speaking fans.

1.5 Challenges in ABSA of Hindi language

The main challenges of ABSA for Hindi language are:

1. *Unorganized information:* The information on the internet is largely unorganized; the same entities, people, places, objects, and events are covered in a variety of ways. The internet is a repository for information from a variety of sources, including books, journals, online documents, health records, company logs, internal files of businesses, and multimedia platforms that include text, photos, audio, and video content.

2. *Word sense ambiguity:* It is common knowledge that a word's meaning can vary depending on the sentence in which it is used. There are words that may

seem similar but have various meanings. Take, for example, the word "कुल" (kul)(family ancestry/total) which can have two meanings, ""वंश"(vansh) may mean the family ancestry and "सब"(sab) may mean the total of some quantity.

3. *Spelling variations:* Words with the same sense and meaning might have multiple spellings in the Hindi language. The following sentences illustrate how the Hindi word "relation" can be spelled differently.

   English: Ram and Shyam do not have good relation.

   Sentence 1 (Hindi :) राम और श्याम के बीच अच्छे संबंध नहीं हैं।

   Sentence 2 (Hindi :) राम और श्याम के बीच अच्छे सम्बन्ध नहीं हैं।

4. *Paired words :* The Hindi language uses a several paired terms. These paired words may consist of two distinct, significant words together.

   As an illustration,

   मेल-भाव,भेद-भाव ।

   Both words have different meaning, first word मेल-भाव means to *reconciliation* and second word भेद-भाव means to *discrimination*. The meaning of word भाव is *rate*.

5. *Complex interdependencies:* Following possible contextual interdependendencies are in Hindi language:

   (a) **Contextual relationships:** Words or phrases can have meanings that depend on their context. For instance, the sentiment of a sentence can change significantly based on the presence of specific words or their order. For example, "बहुत अच्छा" (very good) has a positive sentiment, while "अच्छा नहीं" (not good) has a negative sentiment.

   (b) **Word order and syntax:** Hindi is a subject-object-verb (SOV) language, meaning the word order can affect the meaning and sentiment.

For example, **"मुझे यह किताब पसंद है"** (I like this book) and **"यह किताब मुझे पसंद है।"** (This book is liked by me) convey the same sentiment but have different word orders.

(c) **Negation and modifiers:** Negation words like **"नहीं"** (not) or modifiers like **"थोड़ा"** (a little) can change the sentiment of a sentence. For example, **"यह खाना अच्छा है।"** (The food is good) vs. **"यह खाना अच्छा नहीं है।"** (The food is not good). Negation word **नहीं** changes the sentiment of sentence.

"यह खाना थोड़ा तीखा है।" (This food is little spicy) Here, **"थोड़ा"** (a little) modifies the adjective **"तीखा"** (spicy). It specifies the degree of spiciness, indicating that the food is mildly spicy rather than extremely spicy.

(d) **Long-range dependencies:** Relationships between words or phrases that are far apart in a sentence can impact the sentiment. For instance, **"मैंने उसे हमेशा खुश देखा, लेकिन आज उदास लग रही है।"** (I have always seen her happy, but today she looks sad) requires understanding long-range dependencies to correctly determine the sentiment. The sentiment of above sentiment is negative.

6. *Limited resources:* The lack of a thoroughly annotated corpus also poses a significant challenge for performing ABSA in the Hindi language.

1.6 Problem Statement

People express their opinions towards multiple entities in one document. High-level sentiment analysis i.e. document-level sentiment analysis does not discover sentiment polarities at fine level. Aspect category wise sentiment polarity aggregation helps to understand the sentiment analysis problem better. It presents sentiments about each aspect category in the document.

In spite of large number of Hindi language speakers, only a little research is done in ABSA for Hindi. Many issues in this direction are currently unexplored due to the lack of language processing tools. Language processing systems have transitioned to machine learning algorithms, which offer a high level of accuracy and reliability. However, due to a lack of resources, machine learning and neural network-based approaches are not yet widely applied in Hindi.

This thesis aims to examine machine learning based sentiment analysis and propose a machine learning based framework to aggregate aspect category based sentiment in Hindi document. Following objectives are identified for this are:

1. To carry out a literature review of existing methods for Aspect-Based Sentiment Analysis, data sets and sentiment analysis systems.
2. Post-web reviews collection and sentiment labelled Hindi dataset development.
3. A frame-work for aspect-term category wise sentiment polarity aggregation for a given Hindi document.

1.7 Thesis outline

The organization of the thesis is as follows:

Chapter 1 provides a concise introduction to levels of sentiment analysis. It provides insights to ABSA, applications and challenges associated to ABSA for low-resource language Hindi. The introduction discusses evaluation measures used for performance measures of the proposed approaches. Furthermore, it addresses the problem statement and organization of the thesis.

Chapter 2 discusses the background of sentiment analysis, Aspect Term Extraction(ATE), Aspect Category Detection(ACD), Aspect Sentiment Classification(ASC) and Aspect Category-Sentiment Polarity pair. Furthermore, this chap-

ter provides an in-depth exploration of developed Hindi dataset(TU-HSA) and well-accepted Hindi datasets. These datasets are used in machine learning based experiments in this work.

Chapter 3 provides a comprehensive report on a preliminary investigation of the SA approach using the TU-HSA dataset. It offers valuable insights into the usefulness of SA within TU-HSA dataset.

Chapter 4 investigates the deep-learning approach for ATE task for ABSA for Hindi language. The chapter offers comprehensive explanation of the methodology employed. This chapter presents the advantages of proposed model for low-resource settings i.e. Hindi.

Chapter 5 presents an experiment for Aspect Category Detection(ACD) using well-accepted dataset. This chapter showcases the practical application and effectiveness of ACD for Hindi language providing valuable insights into its real-world utility.

Chapter 6 presents a machine-learning based multitasking framework to aggregate category-wise sentiment polarities for given Hindi document. This chapter provides insights to significance and evaluation metrics based performance analysis of proposed multi-tasking framework over well-accepted dataset.

Chapter 7 functions as a thorough conclusion that succinctly outlines the main discoveries and contributions of the thesis. Furthermore, it delineates prospective paths for future investigation and advancement in the domain of sentiment analysis for the Hindi language.