UNIT - III

Sentiment Analysis:

A. **Sentiment Analysis Overview**

B. **Sentiment Analysis Applications**

C. **Sentiment Analysis Process**

D. **Sentiment Analysis and Speech Analytics**

A. Sentiment Analysis Overview

- **Sentiment** is a difficult word to define. It is often linked to or confused with other terms like belief, view, opinion, and conviction.
- Sentiment suggests a settled opinion reflective of one's feelings.
- Sentiment analysis is different from simply finding topics in text. Here's how:
- **Feelings vs. Facts**: When we look for sentiment, we're looking for opinions or feelings—like whether something is good or bad. In topic categorization, we just want to know what the text is about, like "sports" or "politics," without needing to know how people feel about it.
- **Emotion Strength**: With sentiment, we often want to measure how positive or negative a text is—not just that it's positive or negative. This is different from topic analysis, where it's enough just to know the subject.
- **Sentiment classification**, on the other hand, usually deals with two classes (positive versus negative), a range of polarity (e.g., star ratings for movies), or even a range in strength of opinion .
- Working with just a few categories seems simple, but it's hard because each category includes many different topics, users, and types of documents.
- **Sentiment analysis** has many names. It's often referred to as opinion mining, subjectivity analysis, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion).
- As a field of research, sentiment analysis is closely related to computational linguistics, natural language processing, and text mining.
- Sentiment analysis is trying to answer the question "What do people feel about a certain topic?" by digging into opinions of many using a variety of automated tools.

- **In a business**, especially in marketing and customer relationship management, sentiment analysis seeks to detect favorable and unfavorable opinions toward specific products and/or services using large numbers of textual data sources (customer feedback in the form of Web postings , tweets, blogs, etc.).
- **Sentiment that appears in text comes in two flavors**: explicit, where the subjective sentence directly expresses an opinion ("It's a wonderful day"), and implicit, where the text implies an opinion ("The handle breaks too easily").
- Most of the earlier work done in sentiment analysis focused on the first kind of sentiment, since it was easier to analyze.
- Current trends are to implement analytical methods to consider both implicit and explicit sentiments.
- Sentiment polarity is a particular feature of text that sentiment analysis primarily focuses on.
- It is usually dichotomized into two—positive and negative—but polarity can also be thought of as a range.
- A document containing several opinionated statements would have a mixed polarity overall, which is different from not having a polarity at all (being objective).

B. Sentiment Analysis Applications

- Traditional sentiment analysis relied on surveys and focus groups, which were expensive and slow, using only small participant samples.
- Now, text analytics automates large-scale data collection and processing, using natural language processing to analyze both facts and opinions from social media and the web efficiently.
- Sentiment analysis is perhaps the most popular application of text analytics, tapping into data sources like tweets, Facebook posts, online communities, discussion boards, Web logs, product reviews, call center logs and recording, product rating sites, chat rooms,etc.
- The following applications of sentiment analysis are meant to illustrate the power and the widespread coverage of this technology

  1. Voice of the customer (VOC)

  2. Voice of the Market (VOM)

  3. Voice of the Employee (VOE)

4. Brand Management

5. Financial Markets

6. Politics

7. Government Intelligence

8. Other Interesting Areas

1. Voice of the customer (VOC)

- Voice of the customer (VOC) is an integral part of an analytic CRM(customer relationship management) and customer experience management systems.
- Sentiment analysis can access a company's product and service reviews (either continuously or periodically) to better understand and better manage the customer complaints and praises.
- VOC, mostly driven by sentiment analysis, is a key element of customer experience management initiatives, where the goal is to create an good relationship with the customer.
- For instance, a **motion picture advertising/marketing** company may detect the negative sentiments toward a movie that is about to open in theatres (based on its trailers), and quickly change the composition of trailers and advertising strategy (on all media outlets) to mitigate the negative impact.
- Similarly, a **software company** may detect the negative buzz regarding the bugs found in their newly released product early enough to release patches and quick fixes to alleviate the situation.
- Often, the **focus of VOC is individual customers**, their service- and support-related needs, wants, and issues.
- VOC draw data from e-mails, surveys, call center notes/recordings, and social media postings.

2.Voice of the Market (VOM)

- Voice of the market is about understanding aggregate opinions and trends. It's about knowing what stakeholders—customers, potential customers, influencers, whoever—are saying about your (and your competitors') products and services.

- A well-done VOM analysis helps companies with competitive intelligence and product development and positioning.

3.Voice of the Employee (VOE)

- Traditionally VOE has been limited to employee satisfaction surveys. Text analytics in general (and sentiment analysis in particular) is a huge enabler of assessing the VOE.
- Using rich, opinionated textual data is an effective and efficient way to listen to what employees are saying. As we all know, happy employees empower customer experience efforts and improve customer satisfaction.

4. Brand Management

- Brand management focuses on listening to social media where anyone (past/current/prospective customers, industry experts, other authorities) can post opinions that can damage or boost your reputation.
- There are a number of relatively newly launched start-up companies that offer analytics-driven brand management services for others.
- Brand management is product and company (rather than customer) focused.
- It attempts to shape perceptions rather than to manage experiences using sentiment analysis techniques.

5. Financial Markets

- Predicting the future values of individual (or a group of)stocks has been an interesting and seemingly unsolvable problem.
- What makes a stock (or a group of stocks) move up or down is anything but an exact science.
- Many believe that the stock market is mostly sentiment driven, making it anything but rational (especially for short-term stock movements).
- Therefore, use of sentiment analysis in financial markets has gained significant popularity
- Automated analysis of market sentiments using social media, news, blogs, and discussion groups seems to be a proper way to compute the market movements.
- If done correctly, sentiment analysis can identify short-term stock movements based on the buzz in the market, potentially impacting liquidity and trading.

6. Politics

- As we all know, opinions matter a great deal in politics.
- Because political discussions are dominated by quotes, sarcasm, and complex references to persons, organizations, and ideas, politics is one of the most difficult, and potentially fruitful, areas for sentiment analysis.
- By analyzing the sentiment on election forums, one may predict who is more likely to win or lose. Sentiment analysis can help understand what voters are thinking and can clarify a candidate's position on issues
- Sentiment analysis can help political organizations, campaigns, and news analysts to better understand which issues and positions matter the most to voters.
- The technology was successfully applied by both parties to the 2008 and 2012 American presidential election campaigns.

7. Government Intelligence

- Government intelligence is another application that has been used by intelligence agencies. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications.
- Sentiment analysis can allow the automatic analysis of the opinions that people submit about pending policy or government-regulation proposals.
- Furthermore, monitoring communications for spikes in negative sentiment may be of use to agencies like Homeland Security.

8. Other Interesting Areas

- Sentiments of customers can be used to better design e-commerce sites ,better place advertisements, and manage opinion- or review-oriented search engines.
- E-commerce sites (product suggestions, upsell/cross-sell advertising)
- Better place advertisements (e.g., placing dynamic advertisement of products and services that consider the sentiment on the page the user is browsing)
- Manage opinion- or review-oriented search engines (i.e., an opinion-aggregation Web site, an alternative to sites like Epinions, summarizing user reviews).
- Sentiment analysis can help with e-mail filtration by categorizing and prioritizing incoming e-mails (e.g., it can detect strongly negative or flaming e-mails and forward them to the proper folder)

C. **Sentiment Analysis Process:**

- Because of the complexity of the problem (underlying concepts, expressions in text, context in which the text is expressed, etc.), there is no readily available standardized process to conduct sentiment analysis.
- However, based on the published work in the field of sensitivity analysis so far (both on research methods and range of applications), a multi-step, simple logical process, as given in below figure , seems to be an appropriate methodology for sentiment analysis.
- These logical steps are iterative (i.e., feedback, corrections, and iterations are part of the discovery process) and experimental in nature, and once completed and combined, capable of producing desired insight about the opinions in the text collection.
- Step 1: Sentiment Detection
- Step 2: N-P Polarity Classification
- Step 3: Target Identification
- Step 4: Collection and Aggregation

Step 1: Sentiment Detection

- After the retrieval and preparation of the text documents, the first main task in sensitivity analysis is the **detection of objectivity.**
- Here the goal is to differentiate between a fact and an opinion, which may be viewed as classification of text as objective or subjective. This may also be characterized as calculation of O-S Polarity (Objectivity-Subjectivity Polarity, which may be represented with a numerical value ranging from 0 to 1).
- If the **objectivity value is close to 1**, then **there is no opinion to mine (i.e., it is a fact),** therefore, the process goes back and grabs the next text data to analyze.
- Usually opinion detection is based on the examination of adjectives in text.
- For example, the polarity of "what a wonderful work" can be determined relatively easily by looking at the adjective.

Step 2: N-P Polarity Classification

- The second main task is that of polarity classification.

- Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities, or locate its position on the continuum between these two polarities .
- When viewed as a binary feature, polarity classification is the binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion (e.g., thumbs up or thumbs down).
- **In addition to the identification of N-P polarity, one should also be interested in identifying the strength** of the sentiment (as opposed to just positive, it may be expressed as mildly, moderately, strongly, or very strongly positive).
- Most of this research was done on product or movie reviews where the definitions of "positive" and "negative" are quite clear.
- Other tasks, such as classifying news as "good" or "bad," present some difficulty.
- For instance an article may contain negative news without explicitly using any subjective words or terms. Furthermore, these classes usually appear intermixed when a document expresses both positive and negative sentiments.
- Then the task can be to identify the main (or dominating) sentiment of the document.
- Still, for lengthy texts, the tasks of classification may need to be done at several levels: term , phrase, sentence, and perhaps document level.
- For those, it is common to use the outputs of one level as the inputs for the next higher layer.
- Several methods used to identify the polarity and strengths of the polarity are explained in the next section.

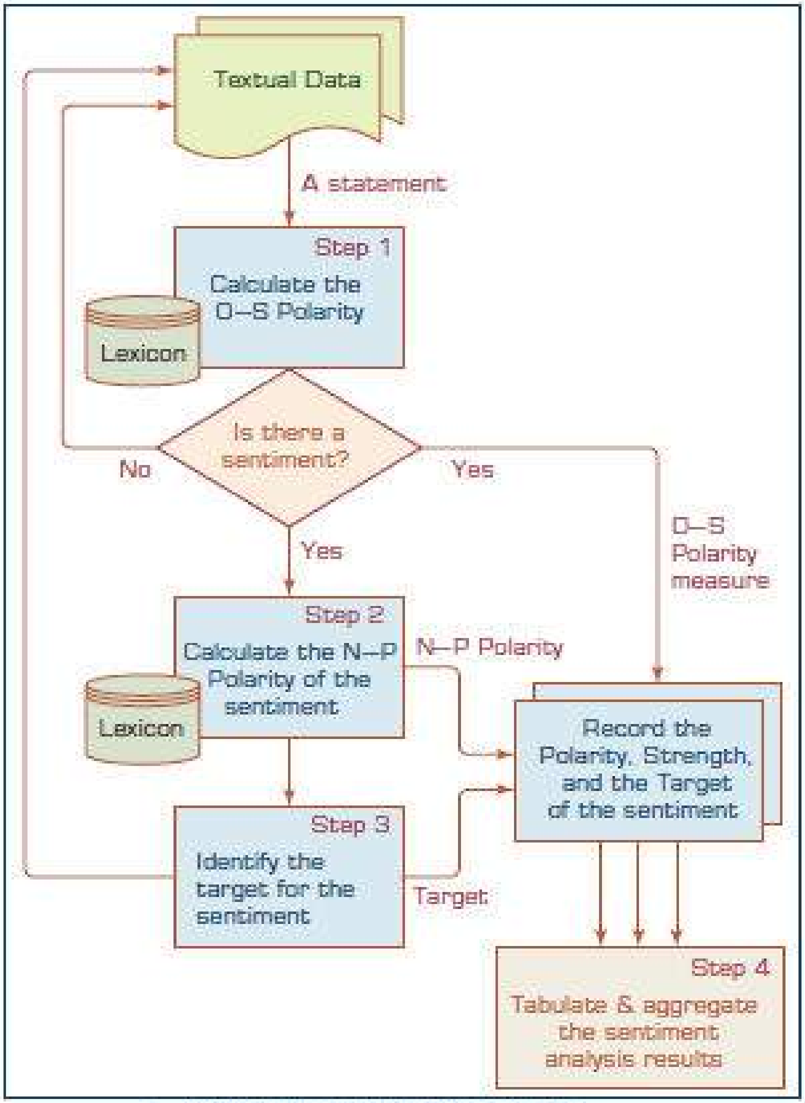
Step 3: Target Identification

- The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, an event, etc.).
- The difficulty of this task depends largely on the domain of the analysis.
- Even though it is usually easy to accurately identify the target for product or movie reviews, because the review is directly connected to the target, it may be quite challenging in other domains.
- For instance, lengthy, general-purpose text such as Web pages, news articles, and blogs do not always have a predefined topic that they are assigned to, and often mention many objects, any of which may be deduced as the target.

- Sometimes there is more than one target in a sentiment sentence, which is the case in comparative texts.
- A subjective comparative sentence orders objects in order of preferences—for example, “This laptop computer is better than my desktop PC”
- These sentences can be identified using comparative adjectives and adverbs (more, less, better, longer), superlative adjectives (most, least, best), and other words (such as same, differ, win, prefer, etc.).
- Once the sentences have been retrieved, the objects can be put in an order that is most representative of their merits, as described in text.

Step 4: Collection and Aggregation

- Once the sentiments of all text data points in the document are identified and calculated, in this step they are aggregated and converted to a single sentiment measure for the whole document.
- This aggregation may be as simple as summing up the polarities and strengths of all texts, or as complex as using semantic aggregation techniques from natural language processing to come up with the ultimate sentiment.



A Multistep Process to Sentiment Analysis.

Example

- Let's take the statement: **"The customer service is excellent, but the delivery was delayed."**

Step 1: Calculate the O–S Polarity

- Determine if the statement is subjective (contains opinion) or objective (factual).
- **Result**: Since the statement expresses opinions (excellent customer service and delayed delivery), it’s classified as **subjective**.

Step 2: Calculate the N–P Polarity

- Determine the polarity (positive or negative) of the sentiment within the statement.

- **Result**:
    - **Customer service**: Positive (e.g., "excellent" has a high positive polarity).
    - **Delivery**: Negative (e.g., "delayed" has a negative polarity).

Step 3: Identify the Target for the Sentiment

- **Objective**: Identify the target(s) of the sentiment in the statement.
- **Result**:
    - Positive sentiment's target is **customer service**.
    - Negative sentiment's target is **delivery**.

Step 4: Record and Aggregate Results

- **Objective**: Store the results in an organized way, such as in an Excel sheet.

Statement ID	Statement Text	O–S Polarity	N–P Polarity	Polarity Strength	Target
1	"The customer service is excellent, but delivery was delayed."	Subjective	Positive/Negative	0.9 / 0.7	Customer Service / Delivery

- After analyzing multiple statements, you can **aggregate** these records by calculating averages, counts, or identifying common targets across multiple entries, making it easier to see overall trends (e.g., if customer service is mostly positive or delivery mostly negative).
- **Methods for Polarity Identification**

    A. Using a Lexicon

    B. Using a Collection of Training Documents

- **Identifying Semantic Orientation of Sentences and Phrases**
- **Identifying Semantic Orientation of Document**
- **Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis**

Methods for Polarity Identification

- Polarity identification deals with identifing the polarity of a text and it can be made at the word, term, sentence, or document level.
- The most granular level for polarity identification is at the word level. (Methods: Using a lexicon , Using a collection of training documents)
- Once the polarity identification is made at the word level, then it can be aggregated to the next higher level, and then the next until the level of aggregation desired from the sentiment analysis is reached. (Identifying Semantic Orientation of Sentences and Phrases, Identifying Semantic Orientation of Document)
- Two dominant techniques used for identification of polarity at the word/term level are

A. Using a lexicon as a reference library (either developed manually or automatically, by an individual for a specific task or developed by an institution for general use)

B. Using a collection of training documents as the source of knowledge about the polarity of terms within a specific domain (i.e., inducing predictive models from opinionated textual documents)

A. Using a Lexicon

- A lexicon is essentially the catalog of words, their synonyms,and their meanings for a given language.
- In addition to lexicons for many other languages, there are several general-purpose lexicons created for English.

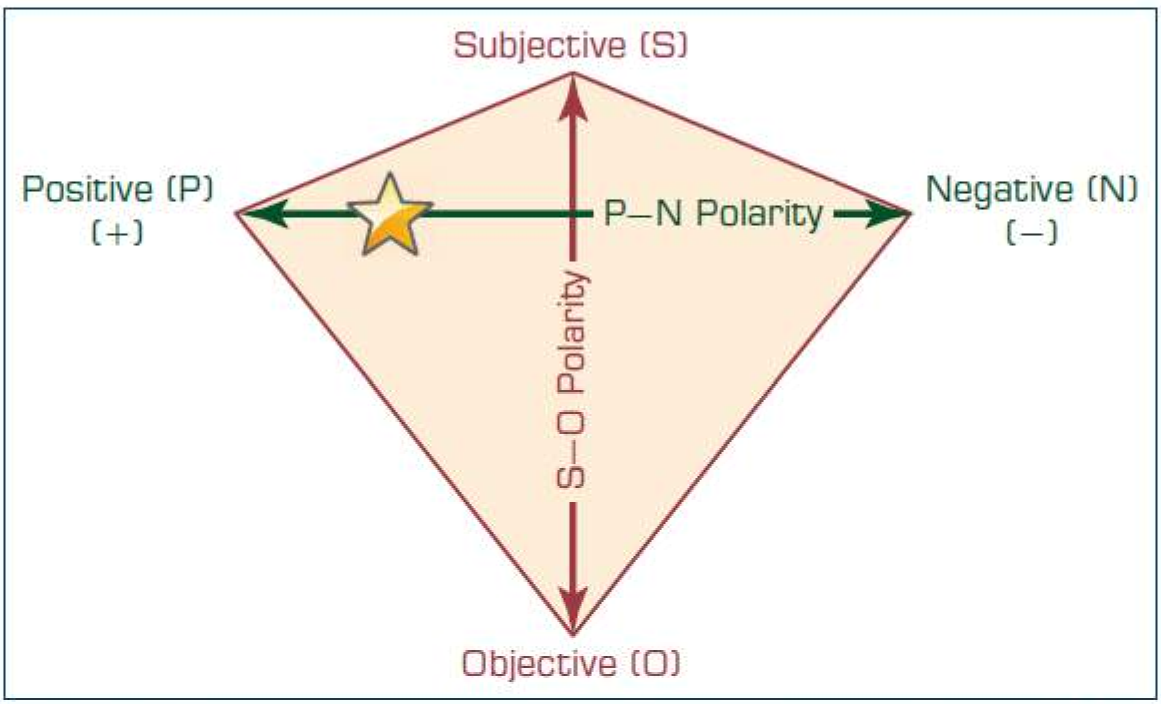
1. The most popular general-purpose lexicon is WordNet, created at Princeton University, which has been extended and used by many researchers and practitioners for sentiment analysis purposes.

- As described on the WordNet Web site (wordnet. princeton.edu), it is a large lexical database of English, including nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (i.e.,synsets), each expressing a distinct concept.
- Synsets are interlinked by means of conceptual-semantic and lexical relations.

2. An interesting extension of WordNet was created by Sebastiani (2006) where they added polarity (Positive-Negative) and objectivity (Subjective-Objective) labels for each term in the lexicon.

- To label each term, they classify the synset (a group of synonyms) to which this term belongs using a set of ternary classifiers (a measure that attaches to each object exactly one out of three labels), each of them capable of deciding whether a synset is Positive, or Negative, or Objective.

- The resulting scores range from 0.0 to 1.0, giving a graded evaluation of opinion-related properties of the terms.
- These can be summed up visually as in Below Figure .
- The edges of the triangle represent one of the three classifications (positive, negative, and objective).
- A term can be located in this space as a point, representing the extent to which it belongs to each of the classifications.



A Graphical Representation of the P–N Polarity and S–O Polarity Relationship.

3. A similar extension methodology is used to create SentiWordNet, a publicly available lexicon specifically developed for opinion mining (sentiment analysis) purposes.

- SentiWordNet assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity.

4. Another extension to WordNet is WordNet-Affect, developed by Strapparava and Valitutti (Strapparava and Valitutti, 2004).

- They label WordNet synsets using affective labels representing different affective categories like emotion,cognitive state, attitude, feeling, and so on.
- WordNet has also been directly used in sentiment analysis.
- For example, Kim and Hovy (Kim and Hovy, 2004) and Hu and Liu (Hu and Liu, 2005) generate lexicons of positive and negative terms by starting with a small list of "seed" terms of known polarities (e.g., love, like, nice, etc.) and then using the antonymy and synonymy properties of terms to group them into either of the polarity categories

B. Using a Collection of Training Documents

- It is possible to perform sentiment classification using statistical analysis and machine learning tools that take advantage of the vast resources of labeled (manually by annotators or using a star/point system) documents available.
- Product review Web sites like Amazon, C-NET, ebay, RottenTomatoes, and the Internet Movie Database (IMDB) have all been extensively used as sources of annotated data.
- The star (or tomato, as it were) system provides an explicit label of the overall polarity of the review, and it is often taken as a gold standard in algorithm evaluation.
- A variety of manually labeled textual data is available through evaluation efforts such as the Text Retrieval Conference (TREC) and Cross Language Evaluation Forum (CLEF).
- Individual researchers and research groups have also produced many interesting data sets
- Once an already labeled textual data set is obtained, a variety of predictive modeling and other machine-learning algorithms can be used to train sentiment classifiers.
- Some of the most popular algorithms used for this task include artificial neural networks, support vector machines, k-nearest neighbor, Naive Bayes, decision trees, and expectation maximization-based clustering.

Identifying Semantic Orientation of Sentences and Phrases

- Once the semantic orientation of individual words has been determined, it is often desirable to extend this to the phrase or sentence the word appears in. The simplest way to accomplish such aggregation is to use some type of averaging for the polarities of words in the phrases or sentences.
- Though rarely applied, such aggregation can be as complex as using one or more machine-learning techniques to create a predictive relationship between the words (and their polarity values) and phrases or sentences.

Identifying Semantic Orientation of Document

- Even though the vast majority of the work in this area is done in determining semantic orientation of words and phrases/sentences, some tasks like summarization and information retrieval may require semantic labeling of the whole document (REF).
- Similar to the case in aggregating sentiment polarity from word level to phrase or sentence level, aggregation to document level is also accomplished by some type of averaging.
- Sentiment orientation of the document may not make sense for very large documents, therefore, it is often used on small to medium-sized documents posted on the Internet.

Technology Insights Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis

- Congressional Floor-Debate Transcripts: Published by Thomas et al. (Thomas and B. Pang, 2006); contains political speeches that are labeled to indicate whether the speaker supported or opposed the legislation discussed.
- Economining: Published by Stern School at New York University,it consists of feedback postings for merchants at Amazon.com.
- Cornell Movie-Review Data Sets: Introduced by Pang and Lee (Pang and Lee, 2008); contains 1,000 positive and 1,000 negative automatically derived document-level labels, and 5,331 positive and 5,331 negative sentences/snippets.
- Stanford—Large Movie Review Data Set: A set of 25,000 highly polar movie reviews for training, and 25,000 for testing.
- MPQA(Multi-Perspective Question Answering) Corpus: Corpus and Opinion Recognition System corpus contains 535 manually annotated news articles from a variety of news sources containing labels for opinions and private states (beliefs, emotions, speculations, etc.).
- Multiple-Aspect Restaurant Reviews: Introduced by Snyder and Barzilay , it contains 4,488 reviews with an explicit 1-to-5 rating for five different aspects: food, ambiance, service, value, and overall experience.

D. Sentiment Analysis and Speech Analytics

- **Speech analytics is a growing field of science that allows** users to analyze and extract information from both live and recorded conversations.
- It is being used effectively to gather intelligence for security purposes, to enhance the presentation and utility of rich media applications, and perhaps most significantly, to deliver meaningful and quantitative business intelligence through the analysis of the millions of recorded calls that occur in customer contact centers around the world.
- Sentiment analysis, as it applies to speech analytics, focuses specifically on assessing the emotional states expressed in a conversation and on measuring the presence and strength of positive and negative feelings that are exhibited by the participants.
- One common use of sentiment analysis within contact centers is to provide insight into a customer's feelings about an organization, its products, services, and customer service processes, as well as an individual agent's behavior.
- Sentiment analysis data can be used across an organization to aid in customer relationship management, agent training, and in identifying and resolving troubling issues as they emerge.

How Is It Done?

- The core of automated sentiment analysis centers around creating a model to describe how certain features and content in the audio relate to the sentiments being felt and expressed by the participants in the conversation.
- Two primary methods have been deployed to predict sentiment within audio: acoustic/phonetic and linguistic modeling.

The Acoustic Approach

- The acoustic approach to sentiment analysis relies on extracting and measuring a specific set of features (e.g., tone of voice, pitch or volume, intensity and rate of speech) of the audio.
- These features can in some circumstances provide basic indicators of sentiment.
- For example, the speech of a surprised speaker tends to become somewhat faster, louder, and higher in pitch.
- Sadness and depression are presented as slower, softer, and lower in pitch.

- An angry caller may speak much faster, much louder, and will increase the pitch of stressed vowels.
- There is a wide variety of audio features that can be measured.

The most common ones are as follows:

• Intensity: energy, sound pressure level

• Pitch: variation of fundamental frequency

• Jitter: variation in amplitude of vocal fold movements

• Shimmer: variation in frequency of vocal fold movements

• HNR: harmonics-to-noise ratio

- Speaking rate: number of phonemes, vowels, syllables,

or words per unit of time

- When developing an acoustic analysis tool, the system must be built on a model that defines the sentiments being measured.
- The model is based on a database of the audio features and how their presence may indicate each of the sentiments (as simple as positive, negative, neutral, or refined, such as fear, anger, sadness, hurt, surprise, relief, etc.) that are being measured.
- To create this database, each single-emotion example is preselected from an original set of recordings, manually reviewed, and annotated to identify which sentiment it represents.
- The final acoustic analysis tools are then trained (using data mining techniques) and a predictive model is tested and validated using a different set of the same annotated recordings.
- As sophisticated as it sounds, the acoustic approach has its deficiencies.
- First, because acoustic analysis relies on identifying the audio characteristics of a call, the quality of the audio can significantly impact the ability to identify these features.
- Second, speakers often express blended emotions, such as both empathy and annoyance (as in “I do understand, madam, but I have no miracle solution”), which are extremely difficult to classify based solely on their acoustic features.

- Third, acoustic analysis is often incapable of recognizing and adjusting for the variety of ways that different callers may express the same sentiment. Finally, its time-demanding and laborious process make it impractical for use with live audio streams.

The Linguistic Approach

- Conversely, the linguistic approach focuses on the explicit indications of sentiment and context of the spoken content within the audio.
- linguistic models acknowledge that , when in a charged state, the speaker has a higher probability of using specific words, exclamations, or phrases in a particular order.
- The features that are most often analyzed in a linguistic model include:
- • Lexical: words, phrases, and other linguistic patterns
- • Disfluencies: filled pauses, hesitation, restarts, and nonverbals such as laughter or breathing
- • Higher semantics: taxonomy/ontology, dialogue history,and pragmatics

1. The simplest method, in the linguistic approach, is to catch within the audio a limited number of specific keywords (a specific lexicon) that has domain-specific sentiment significance.

- This approach is perhaps the least popular due to its limited applicability and less-than-desired prediction accuracy.

2. Alternatively, as with the acoustic approach, a model is built based on understanding which linguistic elements are predictors of particular sentiments, and this model is then run against a series of recordings to determine the sentiments that are contained therein.

- The challenge with this approach is in collecting the linguistic information contained in any corpus of audio. This has traditionally been done using a **large vocabulary continuous speech recognition (LVCSR)** system, often referred to as speech-to-text.
- However, LVCSR systems are prone to creating significant error in the textual indexes they create.
- In addition, the level of computational effort they require—that is, the amount of computer processing power needed to analyze large amounts of audio content—has made them very expensive to deploy for mass audio analysis.

3. Yet, another approach to linguistic analysis is that of phonetic indexing and search.

- Among the significant advantages associated with this approach to linguistic modeling is the method’s ability to maintain a high degree of accuracy no matter what the quality of the audio source, and its incorporation of conversational context through the use of structured queries during analysis