SENTIMENT ANALYSIS

SAKSHI VERMA

What is sentiment analysis?

- Sentiment analysis is a method used in Natural Language Processing (NLP) that involves extracting emotions from raw texts. It is commonly used on social media postings and customer reviews to automatically determine if specific individuals are positive or negative, as well as why.
- Sentiment analysis enables large-scale and real-time data processing. For example, do you want to evaluate hundreds of tweets, product reviews, or support tickets? One may use sentiment analysis to automatically identify how people are talking about a given issue, acquire insights for data-driven choices, and automate business processes instead of manually going through this data.
- Some examples of how it works:
    1. “*Titanic is a great movie*.” The phrase denotes positive sentiment about the film Titanic.
    2. "*Titanic is not a great movie.*" This phrase treats the movie as not great (negative sentiment)
    3. "*Titanic is a movie.*" When we look at the third one more closely, we'll notice that there isn't a single word in it that can tell us anything about the feeling it conveys. As a result, that is an example of neutral sentiment.

Sentiment Analysis with Python

- Python sentiment analysis is a way of examining a piece of text and determining the hidden sentiment. It is achieved through the use of a combination of machine learning and natural language processing (NLP).

NLP(Natural language Processing) - Natural Language is the medium via which we, as humans, communicate with one another. It might be in the form of speech or text. The automatic manipulation of natural language by software is known as natural language processing (NLP). Natural Language Processing (NLP) is a higher-level term that combines Natural Language Understanding (NLU) and Natural Language Generation (NLG).

How does Sentiment Analysis Work?

Sentiment analysis is a classification method that focuses on identifying an opinionated point of view and its disposition, as well as emphasizing the information that is of special relevance in the process.

With respect to data science, opinion has its own meaning;

- It is a personal empirical experience-based subjective judgment of anything. It is governed in part by factual facts and in part by emotions.
- An opinion may be thought of as a kind of dimension in data on a certain issue. It's a collection of symbols that, when combined, provide a point of view, or perspective, on a certain topic.

Operations on which sentiment analysis can be applied;

- Determining the polarity - i.e., positive or negative
- Identify the opinion holder - on its own and in correlation with the existing audience segments
- Find and extract opinionated data on a specific platform.
- Define the subject matter

Scopes of the sentiment analysis algorithm

- Document-level: fo the entire text
- Sentence-level: obtains the sentiment of a single sentence
- Sub-sentence level: obtains the sentiment of sub-expressions within a sentence

Dealing with opinions is a difficult task. Opinions differ; some of them are categorized as follows;

- **Direct Opinion:** It refers to an opinion expressed directly on an entity or an entity aspect. E.g. "*the responsiveness of the buttons in application X is poor*."
- **Comparative Opinion:** It expresses a relationship between two or more entities' similarities or differences as well as the opinion holder's preference based on some of the entities' shared characteristics. E.g. "*the responsiveness of the button in application X is worse than in application Y.*"
- **Explicit Opinion:** It is a subjective statement that gives a regular or comparative opinion. E.g. "*this chair is rocking.*"
- **Implicit Opinion:** They are implied but not clearly stated. E.g. "*the app started lagging in two days.*".It is important to note that implicit opinions may also have idioms and metaphors, which complicates the sentiment analysis process.

Types of Sentiment Analysis

In the context of business operations, there are different types of sentiment analysis.

- **Fine-grained Sentiment Analysis:** It involves determining the polarity of the opinion. It can be a simple binary positive/negative sentiment differentiation. We can go beyond this into fine-grained sentiment analysis with a larger scale of categories that include; Very positive, positive, neutral, negative and very negative, depending on the case. It is widely used in opinion polls and surveys like 5-star reviews 5 being very positive and 1 being very negative.

- **Emotion Detection:** It is employed to recognize the symptoms of particular emotional states mentioned in the text. Typically, lexicons and machine learning algorithms work together to determine what is what and why.

- **Aspect-based Sentiment Analysis:** Its goal is to convey a perspective on a certain product component. The aspect-based analysis is frequently used in product analytics to monitor how the product is perceived by customers and to identify its strong and weak features.

- **Intent Analysis:** It is all about the action. Its purpose is to determine the type of intention that the message is expressing. It is commonly used in customer support systems to streamline the workflow.

Sentiment Analysis Algorithms :

1. **Rule-Based Approach:**
   This is a practical technique for evaluating text without training or employing machine learning models. The output of this approach is a set of guidelines based on which the text is labeled as positive/negative/neutral. These rules are also known as lexicons. Hence, the Rule-based approach is called the Lexicon-based approach.

   The steps involved are as follows:
   - **Data Preprocessing**
     1. **Cleaning the text:** Involves removal of special characters and numbers from the text.
     2. **Tokenization:** It is the process of breaking the text into smaller pieces called Tokens. It can be performed at sentences called sentence tokenization or word level called word tokenization.
     3. **Enrichment:** POS tagging: POS stands for part of speech. It is a process of converting each token into a tuple having the form(word, tag). POS tagging is essential to preserve the context of the word and is necessary for Lemmatization.
     4. **Stopwords removal:** Stopwords are words that in English convey very little valuable information. We need to remove them as part of text preprocessing.
     5. **Obtaining the stem words:** A stem is a part of a word responsible for its lexical meaning. The popular techniques for getting the root/stem words are Stemming and Lemmatization.
        The main distinction is that stemming simply removes certain characters at the end, and it frequently results in some meaningless root words. Lemmatization produces meaningful root words; however, it requires POS tags of the terms.

   - **Sentiment Analysis (Using any one of them)**
     1. **TextBlob:** It is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. The two measures it uses to analyze the sentiment are:
        - Polarity - It talks about how positive or negative the opinion is. It ranges from -1 to 1 (1 being more positive, 0 neutral and -1 being more negative).
        - Subjectivity - It talks about how subjective the opinion is. Its value ranges from 0 to 1(0 is very objective and 1 is very subjective).

     2. **VADER:** It stands for Valence Aware Dictionary and Sentiment Reasoner. Vader sentiment not only tells if the statement is positive or negative along with the intensity of

emotion. The sum of pos, neg and nuwu gives 1.The compound is the metric that ranges from -1 to 1 and is used to draw overall sentiment.

- Positive if compound >= 0.5
- Neutral if -0.5 < compound < 0.5
- Negative if -0.5 >= compound

3. **SentiWordNet:** It uses the WordNet database. It is important to obtain the POS, and lemma of each word. We will then use the lemma, POS to obtain the synonym sets(synsets). We then obtain the positive, negative, and objective scores for all the possible synsets or the very first synset and label the text.
   - if positive score > negative score, the sentiment is positive
   - if positive score < negative score, the sentiment is negative
   - if positive score = negative score, the sentiment is neutral

   Web scrapping Link : Link
   Code link explaining each step involved in the rule-based approach: Link

2. **Automatic Sentiment Analysis:**
   It involves supervised machine learning classification algorithms. An algorithm is trained with many sample passages until it can predict with accuracy the sentiment of the text. Then large pieces of text are fed into the classifier and it predicts the sentiment as negative, neutral or positive.

   Traditional approach:
   This method requires the gathering of a dataset with examples for positive, negative, and neutral classes, then processing this data, and finally training the algorithm based on the examples. These methods are mainly used for determining the polarity of text. Because they are scalable, traditional machine learning techniques like Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) are frequently employed for large-scale sentiment analysis.
   Naive Bayes: The Bayes theorem is used by the Naive Bayes Classifier to forecast membership probabilities for each class, such as the likelihood that a given record or data point belongs to that class. The most likely class is defined as the one having the highest probability. For example, in the sentence ‘I like this product very much, you get a clear sense of the positive sentiment. The classifier calculates each probability value and the class is selected as positive because the positive value outweighs it.

   Code link: Link

   Deep Learning approach:

Sentiment analysis using NLP deep learning is able to learn patterns through multiple layers from unstructured and unlabeled data to perform sentiment analysis. Two techniques of neural networks are common – CNN or Convolutional Neural Networks for processing of images and RNN or Recurrent Neural Networks for NLP tasks.

Code Link: Link

3. **Hybrid Sentiment Analysis:**
   The most advanced, effective, and often applied method for sentiment analysis is a hybrid model. Provided you have well-designed hybrid systems, you can actually get the benefits of both automatic and rule-based systems. Hybrid models can combine the flexibility of customization with the effectiveness of machine learning.

Sentiment Analysis Challenges

1. **Tone:** Brands can face difficulties in finding subjective sentiments and properly analyzing them for their intended tone.
   Soln : smart sentiment API
2. **Polarity:** In-between conjugations of words such as “not so bad” that can mean “average” and hence lie in mid-polarity (-75). Sometimes phrases like these get left out, which dilutes the sentiment score.
   Soln: A topic-based sentiment analysis can give a well-rounded analysis, but with an aspect-based sentiment analysis, one can get an in-depth view of many aspects within a comment.
3. **Sarcasm:** The act of expressing negative sentiment using backhanded compliments can make it difficult for sentiment analysis tools to detect the true context of what the response is actually implying. This can often result in a higher volume of “positive” feedback that is actually negative.
   Soln: A top-tier sentiment analysis API will be able to detect the context of the language used and everything else involved in creating actual sentiment when a person posts something.
4. **Emoji:** Most emotion analysis solutions treat emojis like special characters that are removed from the data during the process of text mining. But doing so means that companies will not receive holistic insights from the data.
   Soln: Data scientists first analyze whether people use emojis more frequently in positive or negative events, and then train the models to learn the correlation between words and different emojis.
5. **Idioms:** Machine learning programs don’t necessarily understand a figure of speech. Hence, when an idiom is used in a comment or a review, the sentence can be misconstrued by the algorithm or even ignored.
   Soln: The neural networks in an emotion mining API are trained to understand and interpret idioms. Idioms are mapped according to nouns that denote emotions like anger, joy, determination, success, etc., and then the models are trained accordingly.
6. **Negations:** Negations, given by words such as not, never, cannot, were not, etc. can confuse the ML model.

Soln: A sentiment analysis platform has to be trained to understand that double negatives outweigh each other and turn a sentence into a positive. This can only be done when there is enough corpus to train the algorithm and it has the maximum number of negation words possible to make the optimum number of permutations and combinations.

7. **Comparative Sentences:** They may not always give an opinion.
   Soln: Sentiment analysis accuracy can be achieved in this case when a sentiment model can compare the extent to which an entity has one property to a greater or lesser extent than another property. And then tie that to negative or positive sentiment.
8. **Employee bias:** Employee feedback is valuable when it comes to shaping company culture, improving sales tactics, and reducing employee turnover. However, due to biases, many businesses find it difficult to separate information. These can come from either the employee or the surveyor, who might not take an ex-comments employee seriously.
   Soln: Text analytics can help read the actual sentiment behind employee feedback and analyze emotional responses to determine bias and eliminate human errors.
9. **Multilingual sentiments:** Each language needs a unique part-of-speech tagger, lemmatizer, and grammatical constructs to understand negations.
   Soln: the sentiment analysis model needs to have a uniquely trained platform and named entity recognition model for each language
10. **Audio Visual data:** Videos are not the same as text data. The challenge is not only that videos need to be transcribed but that they may have captions that need to be analyzed for brand logos.
    Soln: It needs to have a video content analysis model that can break down videos to extract entities and glean insights about customer opinion, product insights, and brand logos.