SENTIMENT ANALYSIS AND EMOTION DETECTION FROM TEXT

A Thesis

Submitted in partial fulfillment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

in

SOFTWARE ENGINEERING

Submitted to

DELHI TECHNOLOGICAL UNIVERSITY, NEWDELHI

Submitted by

Manoj Prajapati

(2k22/SWE/11)

Under the guidance of

Dr. Sonika Dahiya

(Assistant Professor , Department of Software Engineering)

Delhi Technological University, Delhi

DELHI TECHNOLOGICAL UNIVERSITY

SHAHBAD DAULATPUR, DELHI-110042

MAY, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur,Main Bawana Road,Delhi-42

CANDIDATE'S DECLARATION

I hereby want to declare that the thesis entitled "**Sentiment Analysis and Emotion Detection from a Text**" which is being submitted to the Delhi Technological University, in the partial fulfilment of the requirements for the award of degree in Master of Technology in Software Engineering is an authentic work carried out by me. The material contained in the thesis report has not been submitted to any other institution or university for the award of any degree.

Manoj

Manoj Prajapati

(2K22/SWE/11)

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to best of my knowledge.

Sonika 23/Oct/24

Supervisor Signature

Dr. Sonika Dahiya

Scanned with OKEN Scanner

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur,Main Bawana Road,Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

This is to confirm that Manoj Prajapati(2K22/SWE/11) has successfully completed the thesis report titled "Sentiment Analysis and Emotion Detection from a Text" under my supervision as part of the MASTER OF TECHNOLOGY degree in Software Engineering at DELHI TECHNOLOGICAL UNIVERSITY.The thesis embodies the results of original work,and studies are carried out by the student himself,and the contents of the thesis do not form the the basis for the award of any other degree to the candidate or to anybody else from this or any other Univeristy/Institution.

Sonika
03/06/23

Dr. Sonika Dahiya

(Assistant Professor)

Department of Software Engineering

Delhi Technological University

Scanned with OKEN Scanner

ABSTRACT

Detecting emotions and thoughts in text is a fascinating topic in machine learning for natural language processing. Emotions and feelings can often be revealed through specific phrases. Many people around the world use foreign languages, and many documents are written in English. Some individuals do not always use proper punctuation in their texts. Unlike other methods, we study emotions in text with or without punctuation to explore how to design an effective emotional management system using certain beneficial approaches. By developing our methods in a specific way, we can better track and identify feelings more accurately. In this paper, we applied various techniques to identify emotions, including Naïve Bayes classifier, linear SVM, logistic regression, and random forest. Among these, random forest achieved the highest accuracy. The challenge addressed in this paper is recognizing emotions or feelings that are not explicitly shared in posts, blogs, and social media pages using this advanced learning algorithm. The goal of sentiment analysis is to understand people's overall attitudes within a community. The Internet is a place where people from around the world share their opinions on various topics. They discuss everyday issues, complain about products, and give positive or negative feedback. This makes the Internet a valuable source of information for opinion mining and sentiment analysis. In computer science, sentiment analysis has many uses, such as identifying thoughts that can improve customer messaging and understanding opinions in written emails. However, well-written data is often limited to a few expressions, which can reduce the chances of accurate analysis. The authors have collected data from text messages to address these issues. The main aim of the project is to provide a clear written statement that conveys emotion. The proposed model uses two strategies to determine which one works best. The biggest challenge in sentiment analysis is interpreting the text according to the underlying thoughts.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur,Main Bawana Road,Delhi-42

ACKNOWLEDGEMENTS

I take this opportunity to express my deep sense of gratitude and respect towards my guide Dr.Sonika Dahiya, Department of Software Engineering. I am very much indebted for her generosity, expertise and guidance I have received from her while working on this project. Without her support and timely guidance the completion of the project would not be possible.She have guided not only with the subject matter, but also taught the proper style and techniques of documentation and presentation. I would like to express my gratitude to the university for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

Manoj Prajapati

2K22/SWE/11

M.Tech, SWE

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Internet has grown up in such a way where individuals is sharing their experiences and opinions which impacting their life including marketing and communication. Customer assessing only those product and services which have good review and rating on online platform. In some industries, these online evaluations play a vital role in determining the purchasing decision of a consumers. Forrester has made an statement that over 30% user are sharing their rating and reviews on online platform .Consumers also have a greater belief in the fellow customer who have good experiences with the product and services rather than tradition company making claims.

Social media, in particular, is playing a major role in making consumer preferences by showing their behaviour and attitudes. Over the years, Internet has made a big difference in purchasing behaviour through social networking. Retailers, who once used to focussed on physical stores for increasing their sales now believe that social medial platforms dramatically increase their visibility. Additionally, brands image can be significantly increased by providing a good behaviour on the social networking. This approach helps companies in collecting valuable information , unlabelled data on demand trend in unethical manner.

Currently, Facebook holds the dominant position in the digital marketing landscape, closely followed by Twitter. Although platforms like blogs, YouTube, and MySpace offer obvious benefits, they are less frequently utilized for marketing purposes.

These dynamics have led to the emergence of a thriving industry, with numerous companies specializing in providing Sentiment Analysis services on social media. Sentiment Analysis and Opinion Mining have become established, albeit still

developing, fields focused on research, development, and innovation. The overarching objective is consistent: to identify "who" is discussing "what," "when," and "in what context."

Social Sentiment Analysis involves using natural language processing (NLP) to examine online social conversations and uncover deeper insights as they relate to a specific topic, brand, or theme. Metrics such as the net sentiment score and brand passion index reveal how users feel about a brand and allow for comparisons with competitors.

The work described aims to provide a comprehensive overview of sentiment analysis by addressing a wide range of techniques and perspectives. It notes that previous surveys often focus primarily on machine learning, transformer learning, and lexicon-based approaches, sometimes overlooking other sentiment analysis techniques. This study distinguishes itself by encompassing the most frequently used methods, thus offering a broader and more inclusive examination of sentiment analysis.

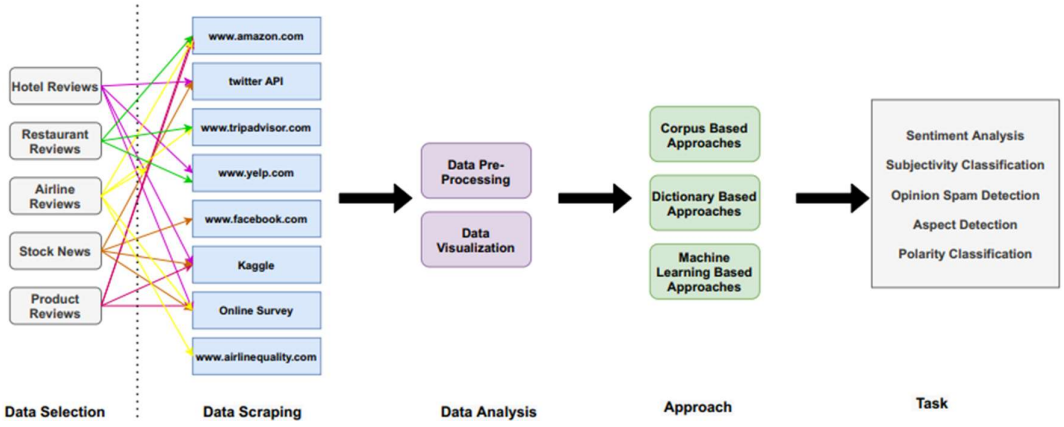


Fig 1.1-General procedure of Sentiment Analysis

This study is not limited to a single task or challenge but rather explores sentiment analysis from multiple angles. It considers various research components related

to sentiment analysis, such as problems, applications, tools, and approaches. By doing so, it offers a thorough investigation that can benefit both scholars and beginners, providing a rich repository of knowledge within a single paper. The paper's significant contributions are outlined as follows:

1. **Comprehensive Literature Review**: This Comprehensive Literature Review incorporate an deep review of existing literature to express the sentiment analysis greatly.It also determines some famous technologies which are used to understand the fundamental component and methodologies of this field.
2. **Methodology Analysis**: It mainly helps in selecting the best available methodologies working best for different application.It mainly includes many different techniques to achieve there appropriateness in diverse contexts.
3. **Classification and Summarization of Approaches**: This one mainly gives the summary of the frequently used technique for sentiment analysis. It helps in understanding the accessible techniques, like, machine learning, lexicon-based analysis, and hybrid analysis. By using these technique the study assist easier comparison.
4. **Benefits and Challenges Summary**: The study generates many challenges and benefits related to sentiment analysis. This helps to keep updated about current and ongoing issue in the field. It also tells what problem can occur while conducting the research further.
5. **Method Comparison and Recommendations**: This study makes comparison in terms of advantages and disadvantages. This is essential for selecting the appropriate method for sentiment analysis. Knowing the strength and weaknesses of each approach ,researcher can make decision which technique is to be used in the research.

Overall, this all includes wide range of technique which provides detailed analysis of sentiment analysis. It is like a sea of knowledge where researcher can find the enough knowledge which will be helpful in research work of sentiment analysis.

2.1 Problem Statement:

(i)A task is Sentiment Analysis and Emotion Detection from a Text.

(ii) The shared opinion in a text, a sentence or an entity aspect is positive, negative, or neutral.

(iii)Advanced, “beyond polarity” sentiment classification looks, for instance, at emotional states such as “angry”, “sad”, and “happy”.

Emotion detection is a significant advancement in human-machine interaction, enabling nonliving systems to perceive and respond to human emotions as if they were human themselves. Our proposed model focuses on detecting emotions from text, which inherently lacks tonal or expressive cues, making the task particularly challenging.

Unlike many previous studies that have concentrated on a single dataset, our research leverages three distinct datasets to enhance the robustness and versatility of our model.

Text-based Emotion Recognition model is being implemented across wide range of system.This is proven to be best in business potential in several areas like:-

1. **Customer Reviews**: Businesses can know the purchasing behaviour of consumer related to product and services by analyzing customer reviews.It

can be helpful in improving quality of product and services ,can increase product development and can meet customer's demand.

2. **Service Feedback**: With the help of emotion detection Service providers can get to know the customer experiences and their sentiment and they may try to imrove their customer supper based on this feedback.
3. **Social Media Security**: The model can be helpful in checking the emotional indicators like distress, anger, other significant emotion using social media platform. It can be also useful to protect the users from threats by identifying the potential security threats.
4. **Market Research**: Companies can come to know what conusmers thinks about their product or brand by using this emotion detection,thereby implementing some market strategy based on emotional feedback.
5. **Healthcare**: Emotion detection can be beneficial in healthservices by analyzing patient communication such as anxiety or depression and based on that they can be treated.
6. **Human Resources**: In the workplace, emotion detection is mainly used to know the feedback of the employee so that they can be solved before they raised any issue.

By using an appropriate approach to dataset selection and application,our model proves to be better in detecting emotion from the text with greater accuracy. The versatility of this capability underscores its importance, making emotion detection a key tool in various industries. It can be utilized to improve customer service, enhance social media safety, and innovate

healthcare, demonstrating its broad and impactful applications.

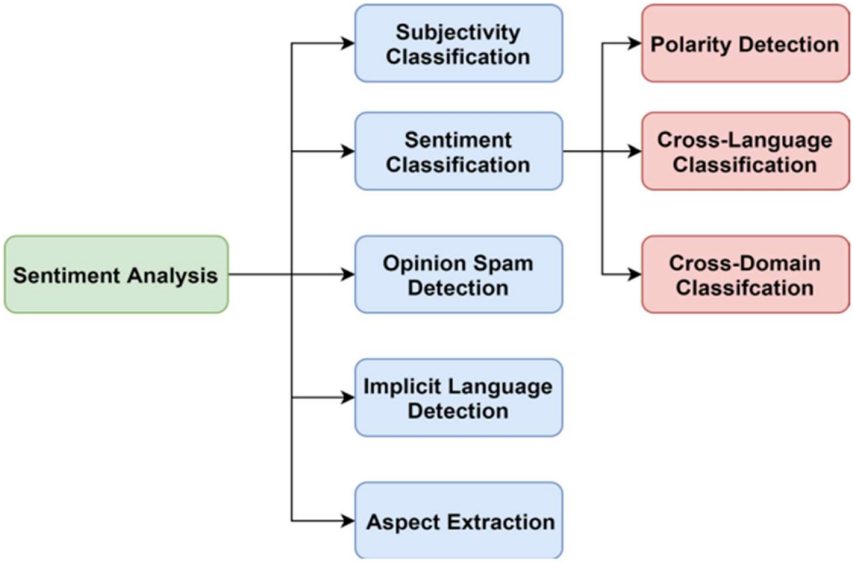


Fig.1.2-Task of Sentiment Analysis

2.2 Task of Sentiment Analysis:-

The tasks of sentiment analysis, as depicted in Figure 2, can be evaluated and explained as follows:

Subjectivity Classification

This is the base of the sentiment analysis i.e first step. It mainly identify subjective hints, emotional phrases, and personal opinions in text. The words that indicate subjectivity are 'hard', 'amazing', and 'cheap'. The main aim of this classification is to differenciate between factual and opinionated.This step make sure that only subjective information is passed as it is essential fo accurate sentiment analysis. By filtering out objective data, this step ensures that only

subjective information is processed further, which is essential for accurate sentiment analysis.

Sentiment Classification

Sentiment classification mainly helps in determining the sentiment of each piece of text(positive,negative,neutral). This task is also known by the name of opinion analysis because it mainly focus on the sentiment of text expressed in the opinion.. Cross-domain and cross-language analysis are mainly used to extract common features from the sentiment in different domain.Many problem arise when ambiguity occurs i.e same word having different meaning.

Opinion Spam Detection

The e-commerce and review platform is rising day by day and with the increase of this ,detection of fake reviews is important. The main aim of this Opinion Spam detection is to detect the fake review which may impact the brand's image or product. They mainly catch the fake review by examining the content of reviews, IP address and geolocation. Machine Learning algorithm help in analyzing the content to prove it is genuine although access to metadate can sometimes be limited.

Implicit Language Detection

It mainly includes sarcasm ,irony and humor which makes a tough task for sentiment analysis to detect. These kind of speech can change the sentiment of the sentence There is need of sign for detecting implicit language like emotions, laughter, and punctuations.

Aspect Extraction

It mainly breaks down the text to investigate the specific part of a subject.They mainly follow three steps i.e extracting aspects, classifying their polarity, and aggregating the results.Using the frequency-based,syntax based and machine learning methods,aspects are predefined .In product reviews the word which are frequently used represent key aspectsIt differenciates from general sentiment analysis by giving more detailed insights.

In summary, these tasks are mainly responsible for improving the effectiveness of sentiment analysis by identifying fake reviews,finding out specific aspects and understanding implicit language like sarcasm. Each task is important in improving the accuracy of the sentiment analysis in different context.

2.3 EMOTION ANALYSIS FROM TEXT:-

Text is considered to be the the primary method for people to share their opinion about other individuals, events, or things. Extracting emotions from text is challenging due to the nature of the data. Emotion detection is easier when specific words clearly indicate a particular emotion, but often emotions are expressed subtly. Additionally, a single piece of text can contain multiple emotions. Texts may include ambiguous emotions and words, words with multiple meanings, or different words conveying the same emotion. Texts can also contain sarcasm, slang, multilingual content, spelling errors, acronyms, and grammatically incorrect sentences, which are common online. These complexities make automatic emotion detection very difficult. As a result, research on extracting emotions from text is a popular topic, with researchers worldwide focused on developing modifications, improvements, and new approaches to address these challenges.

CHAPTER 2

LITERATURE REVIEW

In this section I will write a literature review on some research papers based on sentiment analysis and emotion detection from text.

In this paper discussion involves two different types of writing styles: formal and informal. Similarly corpus for experimentation can be divided into: formal text corpus and informal text corpus. Brief surveys about both styles are presented here.

2.1 Formal text Corpus

Formal writing style is commonly used in various types of texts, including literary arts (such as poems, novels, essays, and plays), official documents, and legal documents. Among these, literary arts represent one of the more complex examples of formal writing.

Barros L. et al.[1] conducted an experiment to automatically categorize poems based on their emotional content, using Quevedo's Spanish poetry and referencing Bleuca's Categorization for comparison. Their work had two main objectives: first, to determine if the original manual classification of poems could be distinguished by the sentiment they conveyed, and second, to explore if automatic learning techniques could improve classification results. They used the Weka toolset to build a Decision Tree for classification, achieving an initial accuracy of 56.22%, which increased to 75.13% with the application of a resample filter. The experiment demonstrated that sentiment analysis is effective for classifying poems and that there is a significant relationship between the detected emotions and Bleuca's categories.

He Z.S. [3] used a Support Vector Machine (SVM) based method to distinguish between bold-and-unconstrained and graceful-and-restrained styles of poetry. Firstly He converted the poem into Vector Space Model(VSM) and then he selected the important words.Using the SVM, poetry style was classified and then features numbers and feature items were observed.With the help of 10-fold cross-validation,the method achieved accuracy of 88.6%.

Kumar V. and Minz S. [4] used three algorithms K-nearest neighbor (KNN), Naïve Bayesian (NB), and Support Vector Machine (SVM) and find the best model suitable for classifying the poetry style.They took the help of information gain for feature selection.Out of three algorithms SVM achieved the highest accuracy of 93.25%.

Sailunaz et.al[12] conducted an study using a hybrid method combining lexicon-based and Naive Bayes techniques. Initially, from small labelled dataset confidence paramenter is known using lexicon based method.Then Naïve Bayes classifer is trained on testing set to achieve a accuracy of 77.16%.

Pang et.al[2] conducted a study and introduced a hybrid model which mainly focussed on user reviews and comment on youtube.focusing on user reviews and comments on YouTube. This model uses a flexible grid and word patterns to find hidden problems and understand nostalgic comments from Oscar-nominee film trailers. The study demonstrates the framework's effectiveness in identifying sentiment and latent valence problems within consumer feedback.

Tahani Almanie et.al[16] conducted an study and introduced a framework which helped in monitoring emotional tweets from Saudi Arabian cities. The system uses over 4000 emotional words and emojis to track tweets in various Saudi dialects, displaying the results on an interactive map. This map reflects the

predominant emotions in each city, providing real-time insights into public sentiment.

2.2 Informal Text Corpus

Alswaidan et. al[13] present a framework for in-depth real-time sentiment analysis on Twitter. This method provides valuable insights into public perception regarding specific cases, companies, or individuals, enabling timely interventions to improve or enhance the situation at hand.

Balakrishnan et al.[14] analyzed customer sentiments towards digital payment applications using a hybrid approach. This approach combines supervised methods, such as Support Vector Machines, Random Forests, and Naive Bayes, with unsupervised methods like Latent Dirichlet Allocation. By doing so, they assess feelings and identify emerging topics from customer reviews, offering a comprehensive understanding of user sentiment.

Alswaidan et.al[13] critically review state-of-the-art text emotion recognition methods. Their review compares various approaches, detailing their features, benefits, and drawbacks. The review underscores the importance of mixed techniques and highlights benchmark datasets and the impact of basic NLP tasks such as speech and parsing.

Bhumi Shah et.al[15] explore techniques for detecting sarcasm in text. Their study examines recent advances in artificial intelligence and deep learning, noting that unigram methods are employed in about 50% of English documents, with an average accuracy of 70.9%. They identify Twitter, Amazon, and other social media platforms as primary sources of sarcastic content.

These studies collectively showcase the diverse applications and methodologies in sentiment analysis, emphasizing the significance of hybrid approaches, real-time analysis, and the detection of nuanced sentiments like sarcasm and nostalgia.

In their work, Poria et al. [6] introduced an innovative approach to concept-level sentiment analysis that integrates linguistics, commonsense computing, and machine learning. This method improves the accuracy of tasks such as polarity detection by enabling sentiments to flow from one concept to another based on the dependency relations within the input sentence. This approach enhances the understanding of the contextual role of each concept within the sentence.

Tan and Zhang [5] conducted an empirical study on sentiment categorization of Chinese documents. They evaluated four feature selection methods—Mutual Information (MI), Information Gain (IG), Chi-Square (CHI), and Document Frequency (DF)—along with five learning methods: Centroid Classifier, K-Nearest Neighbour, Winnow Classifier, Naive Bayes, and Support Vector Machine (SVM). Their experiments on a Chinese sentiment corpus consisting of 1021 documents revealed that IG was the most effective method for selecting sentimental terms, while SVM demonstrated the highest performance for sentiment classification.

Emotion	Text
Sadness	Hiding a fake smile
Joy	Heading to campus in the rain to take grad photos
Fear	I was riding with a friend in his car, and we were traveling at 120 km/h on the snow-covered motorway. I felt like getting out.
Sadness	Oh, that's too bad. Should I call a doctor?
Sadness	Bro team lost... .... Least he played with swag... #freshcut
Neutral	My mother-in-law used to do the same thing to us. If we weren't disciplining the kids enough, then we were disciplining them too much. She also complained about the food we gave them, the schools we sent them to, and just about everything else.

Surprise	I almost forgot my hair was red until I looked in the mirror
Disgust	It attributed the poor performance to the departure of over 330,000 dissatisfied East Germans who moved to the West due to their discontent with Communism.

Table2.1:A SAMPLE LIST OF EMOTIONS AND TEXT OBTAINED FROM EMOTION_DATASET

Afzaal et al.[7] introduced an innovative method for aspect-based sentiment classification, which accurately identified features and achieved top classification accuracy. This method was implemented as a mobile application to help tourists find the best hotels in town, and was tested with real-world data sets, showing effectiveness in both recognition and classification.

Ahmad et al.[9] focused on analyzing tweets about two halal products: halal cosmetics and halal tourism. They used Twitter's search function to extract data, which was then filtered using a new model. Deep learning models were applied to evaluate the tweets, employing RNN, CNN, and LSTM to enhance accuracy and build prediction methods. The results showed that the combination of LSTM and CNN achieved the highest accuracy.

Safari et al.[10] introduced a Naïve Bayes(NB) approach which helped in classifying sentiment in product reviews They did it by improving the parameter

evaluation method in Naïve Bayes to support continuous learning The output evaluated that their model got high accuracy on Amazon product.

Writing style		Author	Methodology			Dataset		Performance
			Approach	Feature/Algo	Toolset			
INFORMAL	Micro blog	Cho S.H. and Kang B.H. [20]	SVM	TF-IDF	KLT Korean Language Technology	Twitter, face book, me2day	Korean language	F-Score: 0.85for positive class
		Samsudin N. et.al [26]	KNN	FS-INS	weka	Online messages	Bahasa Rojak(Malayasian)	Accuracy : 69.09 %
	Chat	Lu C.Y.[7]	Emotion detection engine based on web text mining	Semantic role labeling	Implemented using ASP.Net			Accuracy:75%
	Email	Silva A.V.[1]	NB, Decision Trees, AdaBoost, SVM, Random Forest			Enron corpus	English	Accuracy:62% (SVM)
	Review sites	Pang B.[3]	SVM,NB,ME	Unigram, bigram	SVMlight	Movie review corpus	English	Accuracy: 78.7 (unigram), 77.1 (bigram)

Table 2.2: Differential Analysis of Informal Text in Classification

Wankhade et al.[8] conducted a study and introduced a model which help addressing the shortcoming of review of Arabic hotel reviews. They developed and trained SVM and Deep RNN models using word, lexical, morphological, semantic, and syntactic features. The proposed models were evaluated using a dataset of Arabic hotel reviews, and the results indicated that SVM outperformed the RNN model.

In 2020, Ullah et al.[11] investigated the impact of various events from 2012 to 2016 on stock markets. They used a Twitter dataset to analyze the sentiment of these events. The dataset included millions of tweets, which were used to determine the sentiment associated with each event.

CHAPTER 3

METHODOLOGY

There are several methods for detecting emotions and sentiments in text. Emotion and sentiment recognition is part of Affective Computing, and the computational techniques used in this field have been categorized in various ways by researchers. Generally, these methods can be grouped into four categories: 'Keyword-based Method,' 'Lexicon-based Method,' 'Machine Learning Method,' and 'Hybrid Method.' Existing studies have used features such as unigrams (single words), n-grams (multiple words), emoticons, hashtags, punctuation, and negations for emotion detection. The following subsections describe the different approaches to textual emotion detection**.**

3.1 Keyword-based Method

The keyword-based approach to emotion detection in text is one of the simplest and most intuitive methods. The core idea is to identify patterns that match predefined emotion keywords. The first step in this approach is to pinpoint words in a sentence that convey emotion. In this case every word is tagged with a Part-Of-Speech(POS) taggerto extract Noun,Verb, Adjective which are mainly responsible for detecting emotion in linguistic and emotion based research

When these words are identified then they the predefined list of emotion representing words are compared against these words.Those emotion which matches the given keyword is known as emotion of the sentence.Suppose there are number of emotion correspond to a word then various strategies can be employed to select the best emotion out of it.One simple approach is finding the probability score of each word for each emotion and the emotion which have

highest probability score is selected .Other approach can be just selecting the first matching emotion as words primary emotion.

Making of keyword dictionary varies among researcher and it mainly depends on specific emotion studied.Even researcher try to create their own keyword dictionary which will be helpful in their research.

The method can be summarized as follows:

1. **Word Extraction**: Use Part-Of-Speech to tag words to identify Noun,Verb, Adverb and Adjective.
2. **Keyword Matching**: Since there is predefined list of emotion keywords Compare so make comparison with these words.
3. **Emotion Assignment**: Since there can be multiple emotion correspond to a word so use any approach to find the best emotion.
4. **Dictionary Construction**: Expansion of keyword dictionary can be done using WordNet for a comprehensive set of emotion words.

By following these steps, the keyword-based method gives a systematic way to detect emotions in text.

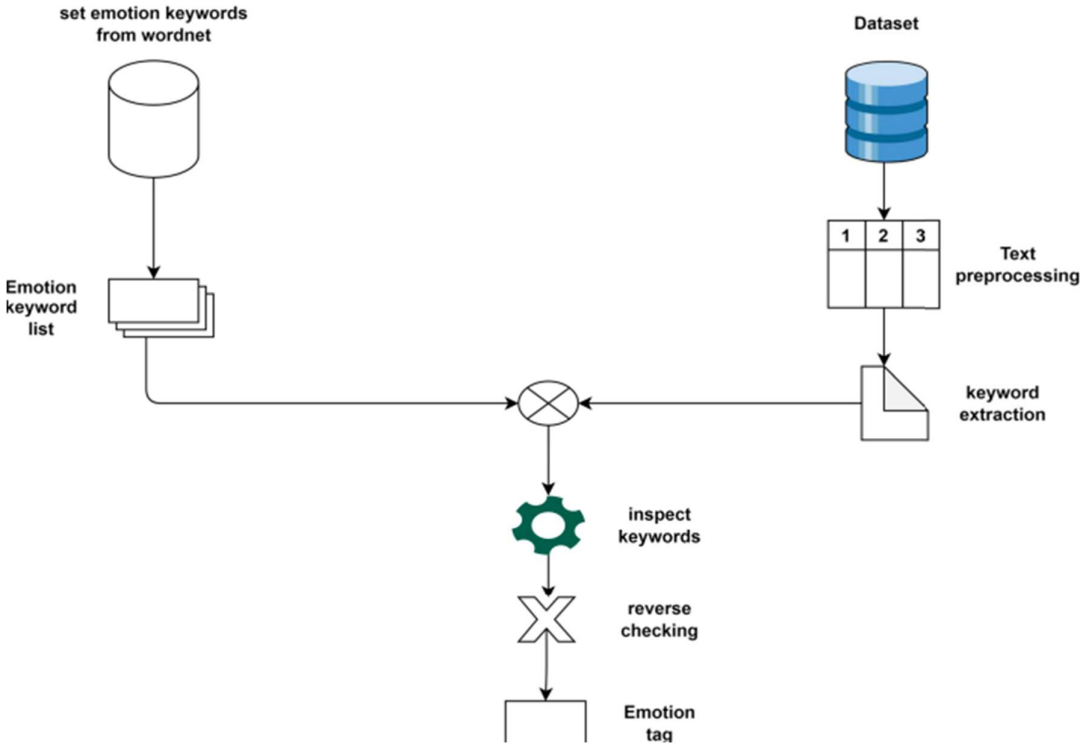


Fig3.1 keyword Based Approach

Advantages:-

1. **Simplicity and Ease of Implementation**: It is easy to use and does not need much computational resources.
2. **Speed**: It is fast in processing text as it only need to match keyword.
3. **Transparency**:It is transparent in nature i.e the method is simple in nature and also understandable.

Disadvantages

1. **Context Insensitivity**: This approach sometimes can not consider the context in which keywords are used which result in incorrect emotion detection.

2. **Negation and Sarcasm Handling**: Keywords is unable to handle some words like sarcasm or negation.
3. **Static and Inflexible**: There is need of regular updates of keyword dictionary so that new words,slang and evolving language use.

Applications

1. **Customer Feedback and Review Analysis**: They are helpful in customer feedback and review analysis.
2. **Social Media Monitoring**: They help in inspecting public opinion and brand perception using social media network.
3. **Market Research**: They are helpful in understanding the consumer attitudes towards product ,services and compaigns**.**

3.2 Lexicon-based Method

The lexicon-based method is one of the best method which is widely used for sentiment anlaysis and emotion detection.This method mainly depends upon the predefined set of words called sentiment lexicon which are annotated with sentiment scoresThese scores obtained helps in identifying the emotion conveyed in the given text. Following is the detailed explanation of how this method works and its applications:

Key Components

1. **Sentiment Lexicon**: Sentiment Lexicon is a collection of words which are associated with predefined sentiment scores(positive,negative,neutral) or emotional categories(joy,sadness,anger,etc). Examples include WordNet-Affect, SentiWordNet, and AFINN.

2. **Text Preprocessing**: This is the next step where cleaning of the text take place by removing stop words,punctuation and special characters as well as tokenizing the text into individual words or phrases.
3. **Sentiment Scoring**:Now here there is matching done with each word in text with the words in lexicon.If there is a matching then sentiment score is assigned to that word.
4. **Aggregation**: The individual sentiment scores or emotional categories are aggregated to compute an overall sentiment score or emotional profile for the entire text.

Process:-

1. **Preprocessing the Text**: The text is cleaned and tokenized. For example, "I am extremely happy today!" would be tokenized into ["I", "am", "extremely", "happy", "today"].
2. **Word Matching**: Each tokenized word is matched against the sentiment lexicon. In our example, "happy" might have a positive score in the lexicon.
3. **Score Aggregation**: The sentiment scores of individual words are summed up to get an overall sentiment score. If the scores are:
    - "happy": +3 The total score would be +3, indicating a positive sentiment.
4. **Emotion Detection**: Similar to sentiment scoring, but using a lexicon annotated with emotional categories. For instance, "happy" might be tagged as "joy".

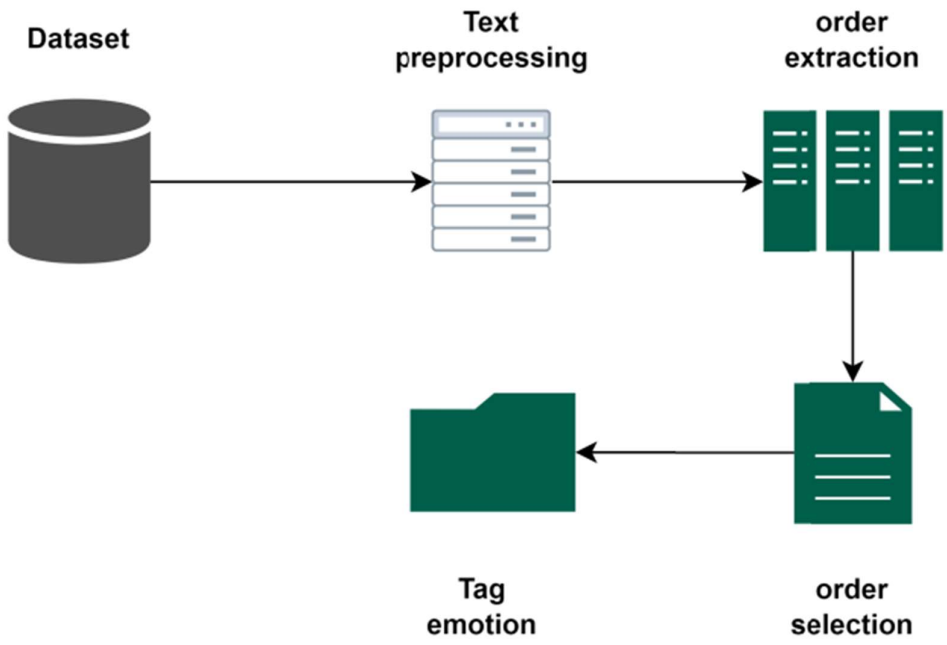


Fig 3.2:-Lexical-based approached work flow

Advantages

1. **Simplicity**: It is simple and easy to understand.
2. **Efficiency**:It does not need much computational resources as compared to machine learning models.
3. **Domain Adaptability**:They are able to adapt any specific domain by creating or customizing tokens.

Disadvantages

1. **Limited Context Understanding**: This approach sometimes can not consider the context in which keywords are used which result in incorrect emotion detection.
2. **Static Lexicon**: They depend upon station set of words so they cant account for evolving language.

3. **Negation Handling**: There is need of special rule to handle negation. (e.g., "not happy" should be negative, but may be interpreted as positive without additional rules).

Applications

1.Customer Feedback Analysis: They are helpful in customer feedback and review analysis.

2.Social Media Monitoring:They take the help of social media platform to understand the customer behaviour towards their product and services..

3.Emotion Detection in Text: They are helpful in in detecting emotions in text for applications such as sentiment-based recommendation systems, mental health monitoring, and interactive chatbots.

4.3 Machine Learning Apporoach:

Machine learning (ML) approaches is one of the best approach for sentiment analysis and emotion detection which require large dataset and to automatically classify the sentiment or emotion expressed in text.They try to incorporate many features and models to achieve high accuracy. Here's an detailed look at how ML-based sentiment analysis and emotion detection work:

Key Components

1. **Dataset**:
    - **Labeled Data**: In Sentiment analysis and emotion detection, every text is associated with correct sentiments like (positive, negative, neutral) or (joy, sadness, anger) and also they need large dataset.

- **Popular Datasets**: The popular dataset is IMDb movie reviews.These dataset have a lot of samples with sentiment lables which provdes a base for training and evaluation models.

2. **Feature Extraction**:
   - **Bag of Words (BoW)**: Bag of Words mainly counts the frequency of each words For instance, a review "The city was great and fantastic" might be represented as a vector indicating the frequency of words like "city", "great", and "fantastic".
   - **TF-IDF (Term Frequency-Inverse Document Frequency)**: This mainly tells the relationship between words and corpus how word is related to corpus.It helps in bringing down the common words and making highlight the important one.
   - **Word Embeddings**: There are many dense vector representation like Word2Vec,GloVe.They capture the semantic meaning of the text.

3. **Model Selection**:
   - **Naive Bayes**: A probabilistic classifier based on Bayes' theorem. It assumes independence between features, which simplifies computation but may not always be accurate.
   - **Support Vector Machines (SVM)**: An algorithm that finds the hyperplane which best separates data points of different classes. SVMs are effective in high-dimensional spaces and are robust against overfitting.
   - **Logistic Regression**: A linear model used for binary and multiclass classification. It estimates the probability that a given input belongs to a particular class.
   - **Deep Learning Models**: Neural networks, especially recurrent neural networks (RNNs) and transformers like BERT, can capture

complex patterns and dependencies in text. RNNs are good for sequential data, while transformers handle long-range dependencies and context effectively.

4. **Training and Validation**:
    - **Training**: The model is trained on a labeled dataset, where it learns to associate text features with sentiment or emotion labels.
    - **Validation**: A separate validation set is used to fine-tune hyperparameters and prevent overfitting. Cross-validation techniques, such as k-fold cross-validation, are often used to ensure the model generalizes well to unseen data.
5. **Evaluation**:
    - **Metrics**: The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics help in understanding the model's effectiveness in classifying sentiments and emotions.

Detailed Process

1. **Data Preprocessing**:
    - **Cleaning**: There are many noise in the text so try to remove it like stop words,punctuation,numbers and special characters.
    - **Normalization**:Sice text are in uppercase so try to convert them into lowercase and perform stemming to bring words to their base form.
    - **Tokenization**: Split the text into individual words or phrases (tokens).

2. **Feature Extraction**:
    - **Bag of Words (BoW)**: Collect all the unique words and keep them in a vocalbulary of dataset.Make each document as vector of word count.
    - **TF-IDF**:for each word kept in vocabulary ,try to calculate the term frequency and inverse document frequency.Now just to get TF-IDF we need to multiple the TF and IDF.
    - **Word Embeddings**: Use pre-trained embeddings (e.g., Word2Vec) . We need to express each word as dense vector in 3D space.
3. **Model Training**:
    - **Naive Bayes**: Calculate the probabilities of each word given a class to train the model. Make use of these probabilities to define new texts.
    - **Support Vector Machines (SVM)**: Train the SVM to find the optimal hyperplane that separates the classes. Use kernels to handle non-linear separation if necessary.
    - **Logistic Regression**: Fit a logistic regression model by estimating the parameters that maximize the likelihood of the training data.
    - **Deep Learning Models**: Train an RNN or transformer model. For RNNs, feed the tokenized text sequences to the network, and for transformers, use the attention mechanism to process the text.
    -
4. **Prediction**:
    - **Naive Bayes**: Calculate the posterior probabilities for each class and assign the class with the highest probability to the new text.
    - **SVM**: Determine the side of the hyperplane the new text falls on and assign the corresponding class.
    - **Logistic Regression**: Compute the probability of each class and assign the class with the highest probability.

- **Deep Learning Models**: Use the trained network to predict the sentiment or emotion of new text inputs.

5. **Evaluation and Fine-Tuning**:
   - **Cross-Validation**: Perform k-fold cross-validation to validate the model's performance. Adjust hyperparameters such as learning rate, regularization strength, and the number of hidden layers in neural networks.
   - **Hyperparameter Tuning**: Use grid search or random search to find the optimal hyperparameters. Evaluate the model on the validation set to avoid overfitting.

Example

Suppose we have a dataset of movie reviews labeled as positive or negative:

1. **Dataset**: "The movie was fantastic!", "I hated the film."
2. **Preprocessing**:
   - Tokenization: ["The", "movie", "was", "fantastic"], ["I", "hated", "the", "film"]
   - Remove stop words: ["movie", "fantastic"], ["hated", "film"]
   - Lowercase: ["movie", "fantastic"], ["hated", "film"]
3. **Feature Extraction**:
   - **Bag of Words**:
     - Vocabulary: {"movie": 1, "fantastic": 2, "hated": 3, "film": 4}
     - Vector representation: [1, 2], [3, 4]
   - **TF-IDF**:
     - Calculate TF-IDF scores for each word in each document.
   - **Word Embeddings**:
     - Use pre-trained embeddings to get vectors for "movie", "fantastic", "hated", "film".

4. **Model Training**:
    - **Logistic Regression**: Fit the model to the vectorized text data.
    - **SVM**: Train the SVM to classify the reviews based on their vector representations.
    - **Deep Learning Model**: Train an RNN or transformer model on the tokenized sequences.
5. **Prediction**:
    - For a new review "The film was awful":
        - Preprocess: ["film", "awful"]
        - Vectorize: [4, vector_for_awful]
        - Predict sentiment using the trained model (e.g., logistic regression, SVM, RNN).
6. **Evaluation**:
    - **Accuracy**: Percentage of correctly classified reviews.
    - **Precision**: Proportion of true positive predictions out of all positive predictions.
    - **Recall**: Proportion of true positive predictions out of all actual positives.
    - **F1-Score**: Harmonic mean of precision and recall.

Machine Learning is one the powerful tool to analyze text data for sentiment analysis and emotion detection.They can achieve high accuracy and can also handle complex pattern in the dataset. They need expertise in many field but their adaptability make them better in application of customer feedback analysis,market research. They follow many process like process data preprocessing, feature extraction, model training, prediction, and evaluation which ensures a effective analysis of sentiment and emotions in text

3.3 Challenges:-

Emotion and sentiment analysis is a kind of field which need to be explored a lot. For complex human emotions ,there is need of some potential to create new system and improving the existing one

Text-based Emotion Network Detection

Now, researcher mainly focus on social media to know the emotion of users. However, We can know the emotion of user by just seeing the comment and replies of user over the social media network.It is useful information as it can help in identifying the emotion over the social media network and can check the spread of specific emotions.

Multiple Emotion Detection

In major study, the main aim is to identify the primary emotion in the text. Since there can be multiple emotion in a sentence and they are mainly ignored.. For instance, if anyone writes, “I was happy this evening but now I am sad or “This makes me excited and anger at the same time,” the system should be able to idemtify the difference between both the emotion.

CHAPTER 4

CONCLUSION

The main aim of this thesis is to review those recent work on text-based emption and sentiment analysis and also highlighting the area which needs improvement. Researcher have used many approaches like keyword-based approach ,Lexicon based approach and machine learning techniques. They have used these approached and techniques on different dataset and have explored the emotion and sentiments. Support Vector Machine and other ensemble system have been used but ensemble system seemed to be perform better.

Despite all recent work ,determing accurately impilicit emotion is a tough task.It is a challenging one.Sentiment detection which define text as positive ,negative or neutral proved to be more accurate. But there can be sarcasm in the text so identifying them is big task and still need of advancements. Some topics which are emerging like identify emotion by intensity,detection of multiple emotion in a single text,identify the cause of emotion and linking emotion and sentiment to social to individual factors.

Future research should conducted on these important emerging areas so that there can be improvement in accuracy of sentiment and emotion detection.With the help of this review a clear thought of different methodologies ,features and output generated by various researcher can be find out.

Emotion detection is an important factor to know how humans interact with computers.Researchers have tried their best to identify all those factor which can detect the emotion accurately including various model with advantage and

disadvantage.If we have to detect emotion from the text then machine learning and deep learning are the best model.

In this paper, we have tried to detect the progress in detection emotion from the text based on existing research.Many effective method have been developed for sentiment analysis and researcher have came to know the some important specification to make better decision.This has resulted in focussing on emotion detection which differenciate between positive and negative. This has led them to focus on emotion detection, which can differentiate between various positive and negative emotions. Additionally, with the rise of social media over the past few years so much textual data has been generated as people share their thoughts, opinion ,feeling about any event ,product or idea.

REFERENCES

[1]. Barros L., Rodriguez P., Ortigosa A. “Automatic Classification of Literature Pieces by Emotion Detection A Study on Quevedo’s Poetry” , Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on ,2-5 Sept. 2013, pp.141-146.

[2]. Pang B., Lee L., and Vaithyanathan S., “Thumbs Up? Sentiment Classification Using Machine Learning Techniques,” Proceedings Of The Conference On Empirical Methods In Natural Language Processing, 2002, pp. 79-86.

[3]. He Z.S., Liang W.T., Li L.Y., Tian Y.F., “SVM-Based Classification Method for Poetry Style” International Conference on Machine Learning and Cybernetics,volume-5,2007, pp. 2936 – 2940.

[4]. Kumar V. and Minz S., “Poem classification using machine learning” Proceedings published in Advances Volume 236, 2014, pp. 675-682.

[5]. Tan S. and Zhang J. “An empirical study of Sentiment Analysis for Chinese documents” Expert Systems with Applications: An International Journal, Volume 34 Issue 4, May, 2008, pp. 2622-2629.

[6]. Poria S., Cambriab E.,Wintersteinc G., Huanga G.B. “Sentic Patterns: Dependency-Based Rules for ConceptLevel Sentiment Analysis”, Knowledge-Based Systems, 2014, pp. 1–32.

[7]. Al Amrani Y, Lazaar M, El Kadiri KE (2018) Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput Sci 127:511–520.

[8]. Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* **55**, 5731–5780 (2022).

[9]. Ahmad Fakhri Ab. Nasir *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **769** 012022.

[10]. Safari, F., Chalechale, A. Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope. *Artif Intell Rev* **56** (Suppl 3), 3273–3297 (2023). https://doi.org/10.1007/s10462-023-10603-3.

[11]. Ullah, A., Khan, S.N. & Nawi, N.M. Review on sentiment analysis for text classification techniques from 2010 to 2021. *Multimed Tools Appl* **82**, 8137–8193 (2023). https://doi.org/10.1007/s11042-022-14112-3

[12]. Sailunaz, K. (2018). Emotion and Sentiment Analysis from Twitter Text Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/32714 http://hdl.handle.net/1880/107533.

[13]. Alswaidan, N., Menai, M.E.B. A survey of state-of-the-art approaches for emotion recognition in text. *Knowl Inf Syst* **62**, 2937–2987 (2020). https://doi.org/10.1007/s10115-020-01449-0.

[14]. Balakrishnan, V., Lok, P.Y. & Abdul Rahim, H. A semi-supervised approach in detecting sentiment and emotion based on digital payment reviews. *J Supercomput* **77**, 3795–3810 (2021). https://doi.org/10.1007/s11227-020-03412-w.

[15]. Shah, B., Shah, M. (2021). A Survey on Machine Learning and Deep Learning Based Approaches for Sarcasm Identification in Social Media. In: Kotecha, K., Piuri, V., Shah, H., Patel, R. (eds) Data Science and Intelligent Applications. Lecture Notes on Data Engineering and Communications

Technologies, vol 52. Springer, Singapore. https://doi.org/10.1007/978-981-15-4474-3_29.

[16]. T. Almanie, A. Aldayel, G. Alkanhal, L. Alesmail, M. Almutlaq and R. Althunayan, "Saudi Mood: A Real-Time Informative Tool for Visualizing Emotions in Saudi Arabia Using Twitter," *2018 21st Saudi Computer Society National Computer Conference (NCC)*, Riyadh, Saudi Arabia, 2018.

PAPER NAME

thesis_manoj_swe.pdf

WORD COUNT	CHARACTER COUNT
7095 Words	**39491 Characters**
PAGE COUNT	FILE SIZE
41 Pages	**1.8MB**
SUBMISSION DATE	REPORT DATE
May 27, 2024 12:08 PM GMT+5:30	**May 27, 2024 12:08 PM GMT+5:30**

Soniba 27/05/2024

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 6% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)

Summary

Scanned with OKEN Scanner

How much of this submission has been generated by AI?

0%

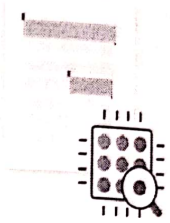
of qualifying text in this submission has been determined to be generated by AI.

Caution: Percentage may not indicate academic misconduct. Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

What does the percentage mean?
The percentage shown in the AI writing detection indicator and in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was generated by AI.



Our testing has found that there is a higher incidence of false positives when the percentage is less than 20. In order to reduce the likelihood of misinterpretation, the AI indicator will display an asterisk for percentages less than 20 to call attention to the fact that the score is less reliable.

However, the final decision on whether any misconduct has occurred rests with the reviewer/instructor. They should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in greater detail according to their school's policies.

How does Turnitin's indicator address false positives?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be AI-generated will be highlighted blue on the submission text.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

What does 'qualifying text' mean?
Sometimes false positives (incorrectly flagging human-written text as AI-generated), can include lists without a lot of structural variation, text that literally repeats itself, or text that has been paraphrased without developing new ideas. If our indicator shows a higher amount of AI writing in such text, we advise you to take that into consideration when looking at the percentage indicated.

In a longer document with a mix of authentic writing and AI generated text, it can be difficult to exactly determine where the AI writing begins and original writing ends, but our model should give you a reliable guide to start conversations with the submitting student.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify both human and AI-generated text) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Scanned with OKEN Scanner