VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2025, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 13, Number 4, October-December 2025 pp:159-176

# Public Perception of Chinese Language Education in Saudi Arabia: A Keyword-Enhanced Aspect-Based Sentiment Analysis of Social Media Discourse

**Farzeen Ashfaq** [1], **Noor Zaman Jhanjhi** [1], **Sayan Kumar Ray** [1], **Husham M. Ahmed** [2]

[1]School of Computer Science, Taylor's University, Subang Jaya, 47500, Malaysia; [2]College of Engineering, University of Technology Bahrain, Bahrain

***Keywords***: *Emerging Technologies, Personalized Learning, BERT, Chinese Language, Keyword Thresholding, LDA, Language adoption*

***Journal Info***:
*Submitted: November 10, 2025*
*Accepted: December 20, 2025*
*Published: December 31, 2025*

## Abstract

The increasing popularity of Chinese as a second language in Saudi Arabia offers a unique chance to study and explore public perceptions and opinions through computational methods. Natural Language Processing (NLP) is one such method, widely adopted in research for this type of analysis. It offers techniques to extract insights from large volumes of unstructured data. However, sentiment analysis on multilingual, culturally specific online discourse remains a challenging task in NLP. The aim of this research work is to address the problem of accurately detecting sentiments and topics in contexts influenced by culture; we discuss the adoption of Chinese language education in Saudi Arabia. For this, we implemented a transformer-based sentiment analysis model on a custom domain-specific dataset with LDA for topic modeling. In this way, we identified key thematic clusters related to globalization, education, and cultural exchange. The research results indicate that topics associated with globalization carry the most positive sentiment, reflecting optimistic public attitudes toward linguistic expansion. This work contributes to the field of applied NLP by demonstrating the feasibility of sentiment and topic modeling in low-resource, culturally diverse environments and contexts.

***Correspondence author email address:** noorzaman.jhanjhi@taylors.edu.my
DOI: 10.21015/vtse.v13i4.2307

## 1 Introduction

NLP has become an important way to understand how people communicate, express emotions, and share opinions online. It allows researchers to make sense of large amounts of multilingual and culturally varied data found across digital platforms. Techniques such as sentiment analysis and topic modeling help researchers explore how people feel about social, educational, and economic issues around the world. As online spaces continue to grow, more people are voicing their thoughts through social media posts, videos, and comment sections, creating huge collec-

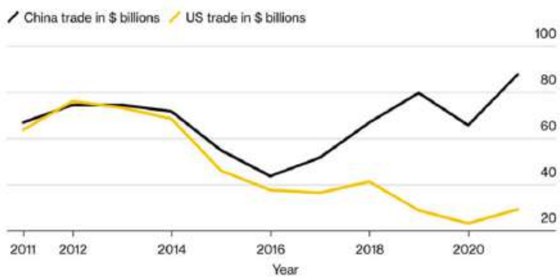


**Figure 1.** China Saudi Trade Volume Over Time

tions of unstructured text. When topic modeling is applied to these texts the analysis can reveal and uncover important trends in language adoption, for example showing how one language can be used in different communities and how learners adopt specific grammar structures influenced by global media. One example is to study public perceptions of bilingual education, international student mobility, and cross-cultural communication initiatives. Then another example is to analyze discussions surrounding language adoption in response to shifting international collaborations. As shown in Figure 1 [3] (source: Bloomberg), changing trade patterns between nations, such as the increasing economic interaction between Saudi Arabia and China, have coincided with a growing interest in Chinese language education [1, 2]. Understanding public perception of these developments by applying NLP for topic and sentiment analysis, provides a compelling opportunity to gain deeper insights into linguistic and cultural dynamics in evolving international contexts [4, 5].

Hence, this changing economic partnership between the two countries [6, 7] has also stimulated increasing interest in the new language among local populations [8]. Recent reports highlight the Saudi government's decision to introduce Mandarin Chinese in schools [9], supported by bilateral agreements between the respective education ministries to provide learning materials and curriculum support. This proactive approach and initiative will not only strengthen ties between the two nations but also will allow and equip students of entirely different culture with valuable language skills to engage with vibrant economy like China.

In the advancing linguistic landscape, understanding the public sentiment and perceptions surrounding this educational initiative is very important and significant. For example, analyzing reactions to the introduction of Mandarin provides valuable insights into the attitudes, beliefs, and expectations of various stakeholders, including students, educators, policymakers, and the wider community [36]. If we can analyze public discussions, such as social media comments and online conversations, using sentiment and topic analysis; we will be able to discover new perspectives and identify possible challenges and opportunities related to any language adoption in any culture. In figure 2 we show a generic view of the main benefits and drawbacks of introducing a language into a country's national education system.

Within the umbrella of NLP, sentiment analysis or opinion mining [10], is a useful computational technique which analyzes and categorizes subjective information expressed in text data [11]. It involves the extraction of sentiment, attitudes, emotions, and opinions from various sources, including social media platforms, online reviews, and discussion forums. Analyzing the sentiments and emotions allows us to measure and interpret the polarity of sentiments expressed in textual data, ranging from positive and negative to neutral [12, 13]. In this study, we applied sentiment analysis to understand how people in Saudi Arabia feel about introducing the Chinese language into the education system as a case study. In this we will examine the hidden and not so clearly evident sentiment trends. We will also be able to identify common themes and evaluate the overall public response to this linguistic initiative. We also implemented aspect based sentiment analysis to recognize and gain a deeper understanding of opinions related to language inclusion in educational curricula. ABSA goes beyond traditional sentiment analysis by looking more in depth and identifying specific aspects or topics within the text and analyzing the sentiment associated with each aspect individually [14]. When we did this we were able not only assess the overall sentiment towards language adoption but also to pinpoint and identify the sentiments related to different aspects


**Figure 2.** Pros and Cons of Introducing New Language

such as globalization, cultural exchange, geopolitical relations, and language learning [15]. The detailed opinions and feelings expressed in online comments were analyzed thus, giving useful insights into the different attitudes toward this language initiative. Statistical methods like LDA [16], which has ability to identify the clusters of words which occur more frequently together, can group dominant and discursive text into topics. Transformers after that can be brought into pipeline like BERT (Bidirectional Encoder Representations from Transformers) [17] to better pay attention to the main context and can produce more accurate sentiment results. Hence when they are fully integrated, we can explore public opinions in greater depth and analyze the gained insights much better than the traditional methods (solo statistical method) which might overlook. Some of the major contributions of our paper are:

1. We developed a domain-specific dataset by collecting YouTube comments related to the adoption of new languages in education, focusing on discussions about introducing Chinese in Saudi Arabian institutions.
2. We applied keyword thresholding to identify important words related to four main aspects: globalization, geopolitical relations, language learning, and cultural exchange.
3. We used LDA to categorize comments into topics based on these four aspects, providing a structured way to analyze sentiments.
4. We classified both positive and negative opinions with pre-trained BERT for sentiment analysis.
5. We used EDA (exploratory data analysis techniques) to visualize and understand the results, helping us study overall sentiment patterns and themes.

## 2 Literature Review

AI has recently changed technology in almost every domain including education and learning like personalized learning assistants, virtual tutors, automated grading systems, prediction system for student success and simulated AR/VR. In this section however, we will review previous studies only related to sentiment analysis, aspect-based sentiment analysis, topic modeling, and their applications in education and language learning contexts. After that we will discuss how these methods have been used in multilingual and culturally diverse settings.

### 2.1 Sentiment Analysis in NLP Research

There are different methods to detect sentiment polarity ranging from rule-based methods and lexicon-based approaches [18–20] to machine [21] and deep learning models [22] and [23]. The lexicon-based method operates predefined word lists like sentiment dictionaries [24] and emotion lexicon to determine whether a piece of text expressed positive, negative or neutral emotions. Although lexicon based methods are simple and interpretable, they often struggle with linguistic ambiguity, sarcasm, and context sensitivity.

ML approaches were later introduced to improve accuracy and scalability. Supervised classification models like SVM, Naïve Bayes, and Logistic Regression [25, 26] have been widely used to classify sentiments in large datasets. Then with the advent of deep learning, models such as CNN, RNN and LSTM, a superior performance in capturing contextual relationships within text [27, 28] was demonstrated. BERT and GPT [29, 30] which are renowned transformer models, have achieved best accurate results because they effectively understand sentence level semantics and context.

### 2.2 Aspect-Based Sentiment Analysis (ABSA) and Topic Modeling

Aspect-Based Sentiment Analysis provides a more detailed understanding than the traditional sentiment analysis techniques because they identify the overall polarity of a text, by linking sentiments to specific aspects or themes [31–33]. In the context of education for example, ABSA can determine whether students express positive opinions about teaching methods or negative opinions about course materials. Researchers have used semi or unsupervised techniques [34], to identify and cluster topics before applying sentiment classification. This integration of topic modeling and ABSA provides structured insights into how people feel about different aspects of a broader issue. These hybrid approaches are particularly useful in exploring opinions in multilingual or informal online discussions, users often express complex and mixed sentiments.

### 2.3 Applications of Sentiment Analysis in Education and Language Learning

The adoption of the Chinese language in Saudi Arabia in recent years has gained quite significant attention from researchers and policymakers alike. The are numerous studies which explored the motivations and implications of the language adoption initiative. For example, Li et al. explores the implications of Saudi Arabia's decision to introduce the Chinese language into its national education system [35]. They discussed the potential development of Chinese studies in Saudi Arabia, examining Saudi Arabia's historical experiences with foreign language teaching. In another study [36], the authors identified challenges and recommendations for integrating Chinese as a foreign language (CFL). The obstacles such as learner motivation, cultural concerns, and pedagogical effectiveness, while suggesting solutions like modern teaching methods and incentives for learners were also emphasized by the authors. The state of the Chinese language instruction in Saudi Arabia was discussed by [37], its difficulties, and possible solutions were also highlighted by the authors. The importance was highlighted as Belt and Road initiative by the authors. A brief overview of Saudi Arabia's recent experience introducing Chinese language instruction mentioning the historical, political, and economic factors when teaching Chinese as a foreign or second language was discussed in [38]. The authors talked about the difficulties that come with teaching and learning, such as individual, psychological, and cognitive factors, in addition to challenges arising from the special characteristics of Chinese characters.

### 2.4 Techniques, Uses, and Trends for Sentiment Analysis of YouTube Comments

Researchers have implemented and employed various methodologies to analyze public sentiment across diverse contexts in the past years. Lexicon-based approaches [39–41], such as sentiment dictionaries [42] and emotion lexicons [43], have been widely used to categorize and quantify sentiment expressed in textual data [44]. ML and statistical approaches like SVM and neural networks, have also been applied to classify sentiment in large datasets with high accuracy [45] and [46]. Then further deep learning techniques, such as recurrent neural networks and convolution neural networks also have shown promising results in capturing complex sentiment patterns from unstructured text data [47, 48] and [49]. In recent years, sentiment analysis of YouTube comments (product reviews, political discussions, and social issues) also has emerged as a valuable approach for understanding public opinion and sentiment across various domains [50].

Studies such as [51] and [52] have investigated sentiment trends and user attitudes expressed in youtube comments related to consumer products which revealed deep insights into consumer preferences and satisfaction levels. Sentiment analysis in the domain of politics(political figures and policies) has been applied by [53] and [54]. Then [55] and [56] discussed analysis of youtube comments regarding social issues, perspectives and sentiments expressed by users on topics such as climate change, social justice, and healthcare. But very specifically in understanding learners' feedback on programming language tutorials the research was done by [57]. The authors collected and scrutinized 2,583 feedback comments on introductory video tutorials for Python and C++. Analyzing post-video comments from leading science channel was conducted by [58], findings revealed that youtube comments in this informal setting often transcended and exceeded information sharing, leading to argumentative negotiation and higher-level knowledge construction, surpassing proportions found in formal settings. Sentiment analysis at aspect level was performed by [59] with unsupervised and semi-supervised LDA techniques to identify aspects in students' reviews. These reviews were then segmented into sentences, and the system estimated the sentiment polarity of each aspect. Sentiment polarity of sentences was found using a custom lexicon resulting in positive polarity for learning aspects, lab facilities, job/career prospects, and grades/tests, and negative polarity for instructors, course content, course structure, fees, and teaching methods. At the end of this section, we can conclude that youtube comments' sentiment analysis is a popular research topic in the field of NLP. Different researchers proposed different approaches to evaluate, explore and examine the hidden feeling in the given comments on youtube videos. The aim of these studies is to use sentiment analysis to forecast the like ratio of a YouTube video by examining the emotive tone of viewer comments. We summarize our review of the literature in table 1.

Although sentiment analysis and ABSA have been applied extensively across domains, limited attention has been given to analyzing multilingual, culturally specific discourse related to language education initiatives. Existing studies have explored policy level discussions on the adoption of new languages, but few have used NLP methods to analyze public sentiment expressed in informal online spaces. This gap presents an opportunity to apply sentiment analysis and topic modeling techniques to examine public perceptions in such contexts. By focusing on multilingual domain specific youtube comments related to language adoption, this study contributes to the understanding of how computational methods reveals attitudes, beliefs, and cultural perspectives in low-resource linguistic environments.

## 3 Methodology

In this section, we outline the research process briefly including data collection and preprocessing. And then we discuss the techniques which we used in our experimentation and the evaluation methods to assess our results.

### 3.1 Advance Text Analysis Techniques

In text analysis there is not a single methods rather variety of methods which are used to extract meaningful information from raw text. Below in the subsections we describe some of the key techniques which were used by us to analyze large amounts of youtube data.

#### 3.1.1 Latent Dirichlet Allocation

LDA is the most popular probabilistic generative model for discovering topics (or themes) from a text corpus. This collection of text may include books, articles, reviews, comments, blog posts, tweets, etc [50, 51]. The baseline idea in LDA assumes that in any collection of text the topic itself is a collection of words and each document is collection of topics thus each word for the topic can be drawn from this collection as it is the probability distribution over words. Equation 1 represents the calculation of the probability of topic $z$ given document dd using Bayes' theorem

$$P(z \mid d) = \frac{P(z) \times P(d \mid z)}{P(d)} \tag{1}$$

Where:

- $P(z \mid d)$ is the probability of topic $z$ given document $d$.

**Table 1.** Summary of the Notable Studies

| Study | Purpose | Method / Algorithm |
|---|---|---|
| [60] | provides key insights for content creators on improving their content and audience engagement by understanding viewers' sentiments towards their videos. | SVM, Naïve Bayes, Logistic Regression, KNN, Decision Tree, and Random Forest. |
| [61] | analyze data from YouTube related to the 2024 Indonesian presidential election candidate to extract insights into viewer perception and interactions | machine learning algorithms |
| [62] | to evaluate user-posted content and likes/dislikes across multiple social media platforms in order to suggest YouTube videos that address climate change and global warming | machine learning algorithms and graph theory |
| [63] | sentiment analysis of YouTube comments related to government policies, specifically the construction of the Mengwi-Gilimanuk Toll Road in Bali Province during the Jokowi era | naïve bayes algorithm |
| [64] | the opinions expressed by viewers on YouTube in comments and on news and political vloggers' videos about the achievements of the two Philippine presidential candidates in the May 2022 election | semi supervised machine learning |
| [65] | study utilizes opinion mining to analyze video simulations to explore the use of virtual reality (VR) as a training tool in the construction industry | machine learning |
| [66] | examine the tone and language used to offer assistance in suicide prevention Public Service Announcements (PSAs) and the comments that accompany them on YouTube. | data mining and machine learning |
| [67] | analyzing user comments on popular YouTube channels dedicated to EFL learning | latent semantic analysis (LSA) |

- $P(z)$ is the prior probability of topic $z$.
- $P(d \mid z)$ is the likelihood $d$ given topic $z$.
- $P(d)$ is the marginal probability of document $d$.

Next in LDA, the joint probability distribution represents the likelihood of observing both the documents in the corpus and the latent topic according to the model's generative process. The generative process assumes that for each document, a distribution over topics is drawn from a Dirichlet distribution, and for each word in the document, a topic is sampled from this distribution over topics. Subsequently, a word is sampled from the topic's distribution over words. Equation 2 shows the calculation of joint probability in LDA as the product of the probabilities of each step in the generative process.

$$\begin{aligned} P(\text{documents, latent variables}) = &\prod_{d=1}^{M} P(\theta_d) \prod_{k=1}^{K} P(\phi_k) \\ &\times \prod_{d=1}^{M} \prod_{n=1}^{N} P(z_{d,n} \mid \theta_d) \\ &\times P(w_{d,n} \mid z_{d,n}) \end{aligned} \tag{2}$$

In equation 2, $P(z_{d,n} \mid \theta_d)$ represents the probability of assigning topic $z$ to word $n$ in document $d$ given the document's topic distribution $\theta_d$. This probability is calculated based on the distribution of topics in the document $\theta_d$, which is determined by the model during the training process. Similarly, $P(w_{d,n} \mid z_{d,n})$ represents the probability of observing word $w$ given its assigned topic $z$ and the word distribution for topic $z$, $\phi_{z,n}$. Again, this probability is determined by the word distribution for the assigned topic $\phi_{z,n}$, which is learned during training.

#### 3.1.2 BERT

One of the latest enhancements for NLP under the umbrella of deep learning was the release of BERT which is Bidirectional Encoder Representations from Transformers by Google in 2018. When utilizing a pre-trained BERT model for sentiment classification, the model architecture comprises several key components tailored to effectively process and analyze textual data as shown in Figure 3. Firstly, the input layer of the model accepts tokenized text data, allowing for efficient representation of the input sequence. Following this, the pre-trained BERT model, which typically consists of a multi-layer bidirectional transformer encoder, is employed to encode and capture contextualized representations of each token in the input text. These representations effectively capture the nuanced contextual information present in the text, enabling the model to understand the sentiment expressed. Additionally, a pooling layer may be incorporated to aggregate the contextualized representations of all tokens into a single pooled representation of the entire input sequence. This pooled representation is then fed into the classification layer, which often comprises one or more dense (fully connected) layers followed by an appropriate activation function such as softmax for multi-class classification or sigmoid for binary classification. The output layer of the model generates the final predictions or probabilities for each sentiment class, providing valuable insights into the sentiment expressed in the input text. Incorporating a pre-trained BERT model for sentiment classification enables robust and accurate analysis of textual data, facilitating various downstream applications in natural language processing tasks.

## 4 Our Proposed Model

Our proposed model utilizes TF-IDF vectors and one-hot encoded topic keyword vectors to enhance LDA for more accurate topic modeling. The pipeline of our methodology is depicted in Figure 4, illustrating the seamless integration of these components for improved topic inference.

### 4.1 Dataset Collection

We collected 4022 comments related to the topic of language adoption. The comments were collected from 10 you tube videos discussing the topic using beautifulSoup library from python. This library enabled parsing and extraction of textual content from HTML-based sources. To maintain the privacy and anonymity of data no personal information such as usernames, IP addresses or other such meta data was collected. Some part of the dataset used in this study collected by the authors has been made publicly available [68, 69].

#### 4.1.1 Data Preprocessing

All the collected comments were stripped of any HTML tags, numbers, or special characters, and preprocessed to only include relevant linguistic information. For the data pre processing we applied and followed a detailed and comprehensive pipeline so as to make sure that the comments which we have collected are cleaned and ready for analysis. We did following steps:

1. We stripped HTML tags using beautifulSoup.
2. We removed all special characters (non alpha numeric) using regular expression.
3. We converted all text to lowercase for maintaining the uniformity.
4. We tokenized all the comments using the $word_tokenize$ function from the NLTK library.
5. We removed stop words( common English words like is, are, the, in etc.) using the stop words corpus.
6. Any punctuation marks were eliminated from the tokens list.
7. Empty tokens were discarded after punctuation and stop words removal.

### 4.2 Deductive Thematic Mapping

This step serves as a labeling or data annotation step in our proposed model pipeline. Once we cleaned the corpus and before we go for the data split we performed this annotation step. As a part of this

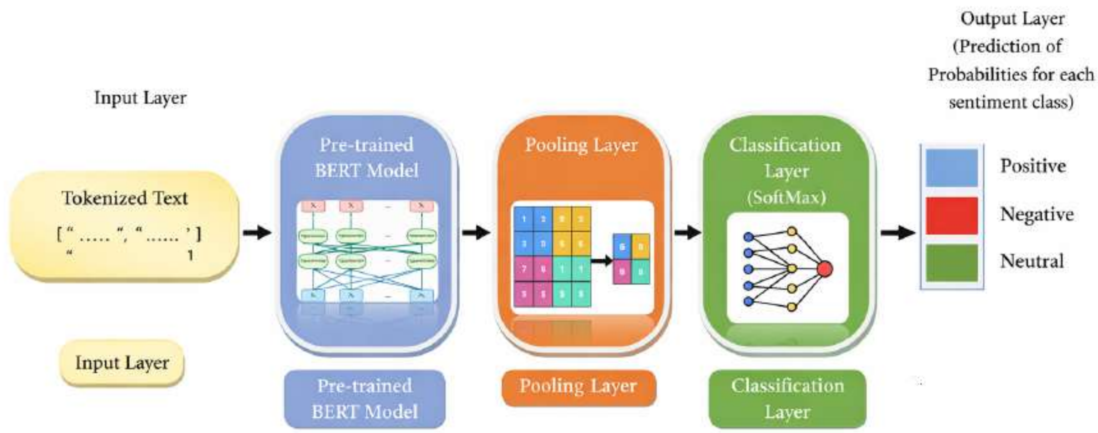


**Figure 3.** Sentiment Classification using a pre-trained BERT model

step we developed a domain specific dictionary to incorporate a predefined taxonomy to categorize comments into specific research-relevant aspects. We defined four primary aspects: *Globalization*, *Language Learning*, *Cultural Exchange*, and *Geopolitical Relations*. This mapping served as a theoretical bridge between our raw data and the keywords we use to enhance the LDA.

To ensure the integrity of the annotated dataset a post annotation check was performed to identify any unannotated comments. Through iterative refinement of the keyword list, 100% of the 4,022 comments were successfully mapped to at least one thematic aspect. In figure 5 we present the distribution of the comments across these aspects. We can see from the bar chart that *Globalization* and *Geopolitical Relations* were the most prominent themes in the discourse. Interpreting these aspect distribution of comments we can say that globalization received most of the comments which means that people think that language adoption is part of larger global trend where countries are connecting and exchanging ideas, cultures, and business practices more than ever. Geopolitical relations as the figure shows is the second important discussed topic. This means that people also see the adoption of some language as a way to promote and foster mutual understanding between any two nations. Language learning (i.e. how people see any language as a useful skill to acquire, perhaps for personal development, education, or business) and cultural exchange (how language adoption facilitate and foster the sharing of social values, religious commonalities, and historical identities between nations) were having nearly equal presence. This means there is a balanced engagement with these themes in the extracted dataset of comments. For the qualitative validation of the mapping process we generated the word clouds for each aspect based on the frequency of terms within the annotated subsets. Figure 6 displays the most frequent terms for each category. For instance, the *Language Learning* aspect is dominated by terms such as "long term," "economic goals," and "Saudi Arabia," while *Geopolitical Relations* word cloud is highlighting "international," "current global," and "execution matters." From the word cloud visualizations we are able to confirm that the deductive mapping correctly isolated the sub-topics present in the raw corpus.

The importance of these word clouds is that they visualize the most frequent terms within the corpus, allowing for a rapid understanding of the dominant themes within each aspect. The word cloud for the 'Cultural Exchange' aspect showed a prominent emphasis on terms such as 'clearly divided,' 'everyone,' and 'shaping education,' indicating a discourse that is heavily debated among the public regarding how cultural shifts affect future generations. In the word cloud for 'Geopolitical Relations,' words such as 'ex-

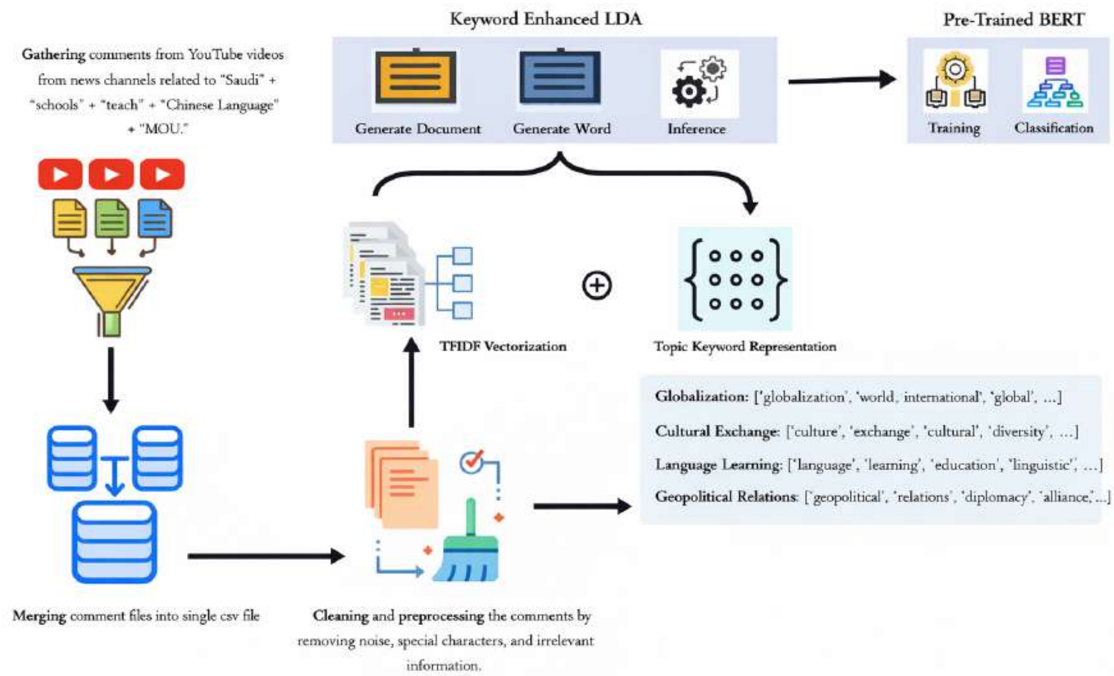


**Figure 4.** Our Proposed Methodology

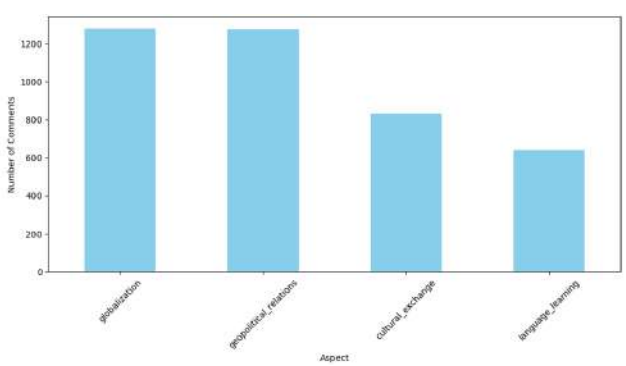


**Figure 5.** Distribution of Comments Across Four Deductive Aspects

ecution matters,' 'international,' and 'current global' stand out, highlighting discussions centered on the practical implementation of strategic decisions. For the 'Language Adoption' aspect, terms like 'language missing,' 'makes sense,' and 'younger generation' feature prominently, suggesting a discourse around how the public views language adoption not just as a school subject, but as a strategic, long-term commitment tied to the nation's broader economic objectives. In the 'Globalization' aspect, terms such as 'long term,' 'others remain skeptical,' and 'economic goals' are indicative and prevalent. This indicates that the global shifts are perceived as a logical necessity for the future.

## 4.3 Dataset Split

The total dataset consists of 4022 comments, which were split into training and validation subsets. Due to the nature of this dataset, we used a 70/30 train-test split to evaluate model performance. The training set contains 3218 comments, while the validation set consists of 804 comments. This split ensured that the model is trained on a diverse range of data, while also allowing for independent evaluation.

## 4.4 Keyword Enhanced LDA

In our hybrid approach we refer to a different application of LDA where keywords are pre-assigned to topics and then the model infers the probabilities of these keywords within each document, rather than inferring topics directly from the documents them-


(a) Geopolitical Relations


(b) Cultural Exchange


(c) Globalization


(d) Language Learning

**Figure 6.** Word clouds validating the thematic consistency of the annotated aspects.

selves. In this case, the traditional formulation of LDA is indeed modified. Instead of inferring topics directly from the documents and then assigning keywords to these topics post-inference, the model incorporates the pre-assigned keywords as part of the inference process. In this modified formulation, the model still aims to estimate the probability distributions of topics within documents and words within topics, but it does so while considering the fixed assignments of keywords to topics. This modification allows the model to incorporate prior knowledge or domain expertise in the form of predefined keywords. In the traditional LDA model, the generative process for a document involves sampling topics from a distribution over topics, and then sampling words from a distribution over words for each topic. The parameters of these distributions are learned from the data during the training process. However, in our modified version where keywords are pre-assigned to topics, the generative process would change slightly. Following sections describe the high-level process of how our keyword enhanced LDA is modified:

### 4.4.1 Generative Process for Documents

Instead of sampling topics directly from a distribution over topics, the topics are pre-assigned based on the predefined keywords associated with each topic. So, the generative process would involve selecting a topic for each document based on the predefined keywords. Formally, the generative process for documents can be expressed as:

**Algorithm 1.** Modified Generative Process For Documents

---

**Require:** Set of predefined keywords $K = \{K_1, K_2, \vdots, K_T\}$ for $T$ topics

1: **for** each document $d$ **do**

2: Determine the set of predefined keywords $K_d$ associated with each topic

3: Assign the topic $z_d$ to document $d$ based on the keywords in $K_d$

4: **end for**

Mathematically, this process can be represented as:

$$z_d = \mathbf{argmax}_k(P(z = k \mid K_d)) \tag{3}$$

In equation 3, $z_d$ is the assigned topic for document $d$, $k$ iterates over all possible topics, and $(P(z = k \mid K_d))$ is the probability of topic $k$ given the set of keywords $K_d$ associated with document $d$.

#### 4.4.2 Generative Process for Words

Once the topics are assigned to each document, the words are generated based on the distribution over words for each topic. In the context of our modal, the generative process for words remains consistent with the traditional LDA model, even if we pre-assigned the topics based on predefined keywords. The process involves generating each word in a document based on the distribution over words for the assigned topic.

**Algorithm 2.** Generative Process for Words

**Require:** Set of predefined keywords $K = \{K_1, K_2, \ldots, K_T\}$ for $T$ topics

1: **for** each document $d$ **do**
2: Determine the topic $z_d$ assigned to document $d$ based on the predefined keywords
3: **for** each word position $n$ in document $d$ **do**
4: Sample a word $w_{d,n}$ from the word distribution associated with topic $z_d$
5: **end for**
6: **end for**

#### 4.4.3 Inference

During inference, the goal is to estimate the posterior distribution over topics for each document given the observed words. This process would involve updating the posterior distribution based on the pre-assigned topics (keywords) and the observed words in the document. In the modified LDA with keyword-assigned topics, the inference process involves updating the posterior distribution over topics for each document given the observed words and the preassigned topics based on keywords. This requires incorporating the information from the keywords into the inference process, which may involve adjusting the likelihood function or incorporating additional constraints to reflect the pre-assigned topics.

$$p(z_d \mid w_{d,n}, \alpha, \beta, K_d) \propto p(w_{d,n} \mid z_d, \beta) \times p(z_d \mid \alpha) \times p(z_d \mid K_d) \tag{4}$$

In equation 4, $p(z_d \mid w_{d,n}, \alpha, \beta, K_d)$ represents the posterior distribution over topics for document $d$ and $p(w_{d,n} \mid z_d, \beta)$ represent the likelihood function representing the probability of observing word $w_{d,n}$ given topic $z_d$. The prior distribution over topics for document $d$ is represented by $p(z_d \mid \alpha)$. The distribution over topics based on the pre-assigned keywords $K_d$ is shown by $p(z_d \mid K_d)$. The posterior distribution over topics for each document $d$ given the observed words $w_{d,n}$ and the pre-assigned topics based on keywords $z_d$ is updated using Bayes' theorem. $\alpha$ and $\beta$ are the Dirichlet hyperparameter controlling the document-topic distribution, and the topic-word distribution respectively.

**Algorithm 3.** Inference Process in Keyword-Enhanced LDA

**Require:** Number of iterations $N$

1: Initialize topic assignments for each word in the document
2: Initialize posterior distribution of topic proportions for each document
3: **for** $i = 1$ to $N$ **do**
4: Update topic assignments for each word in the document based on the predefined keywords $K_d$
5: Estimate posterior distribution of topic proportions for each document:

$$p(\theta_d \mid \mathbf{z}_d, \alpha)$$

6: **end for**

#### 4.4.4 PreTrained BERT

Once the Keyword Enhanced LDA has assigned topics to each YouTube comment based on the predefined keywords, the next step in our proposed model involves utilizing a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to classify the sentiment polarity of these comments. The comments are first preprocessed to ensure compatibility with the BERT model. This preprocessing step involves tokenization of the comments into tokens that BERT understands. Following the preprocessing step, the pre-trained BERT model (we used the nlptown/bert-base-multilingual-uncased-sentiment variant, which supports a range of languages and has been pretrained on sentiment analysis tasks) is used to perform sentiment analysis on each tokenized comment. The BERT model is fine-tuned for multilingual sentiment analysis, enabling it to accurately classify the sentiment polarity of comments in various languages. By analyzing the sentiment polarity, the model assigns a sentiment label to each comment, such as "positive," "negative," or "neutral." We then computed a compound score for each comment, which represented the overall sentiment intensity. The compound score was calculated based on the probabilities of positive and negative sentiments predicted by the model.

$$\text{compound\_score} = P(\text{positive}) + P(\text{very positive}) - P(\text{negative}) - P(\text{very negative}) \tag{5}$$

where $P(\cdot)$represents the probability assigned by the model to each sentiment category.

By considering both the sentiment label and the compound score, our proposed model provided a comprehensive understanding of the sentiment expressed in the YouTube comments. Integrating pre-trained BERT for sentiment analysis enhances the capabilities of our proposed model by providing a sophisticated analysis of sentiment polarity in the comments. This facilitates a deeper understanding of the acceptance level of Chinese as a language in Saudi academic institutes among Arab users, shedding light on sentiment trends and public opinion regarding this linguistic integration.

## 5 Results and Discussion

Our primary objective in this study was to analyze sentiments hidden inside the youtube comments and as case study we picked linguistic dynamics between Saudi Arabia and China. Through a KEABSA (Keyword Enhanced Aspect-Based Sentiment Analysis) we tried to gain insights into the attitudes and opinions of users regarding various topics related to Saudi-Chinese relations. We show in figure 7, the header of the processed corpus shows the KEABSA model successfully assigned sentiment scores ranging from -0.74 (very negative) to 0.86 (very positive).

For instance, comments regarding "shaping education" often triggered negative scores when tied to "divided" opinions, while comments supporting the "China-Muslim alliance" resulted in high positive polarity. This granular level of analysis demonstrates the model's ability to distinguish between general sentiment and specific aspect-based opinions.

The sentiment analysis figures, as depicted in figure 8 and figure 9 are shown as the sentiment polarity surrounding the introduction of the Chinese language in Saudi schools and colleges. This is actually the depiction of public responses for adopting a new language into an existing education system. Figure 8 presents a simplified binary sentiment density view where we see only negative and positive sentiment polarities to highlight the overall trend within each aspect. On the other hand, figure 9 provides a more comprehensive and aggregated multi-class sentiment breakdown providing more granular and detailed distribution including categories such as very negative, negative, neutral, positive, and very positive. The two visuals together shows both understanding of public favorability and a detailed analysis of the intensity of user opinions across the four thematic aspects.

KDE or Kernel Density Estimation is a statistical technique used to estimate the probability density function of a continuous random variable. In the context of the provided figure 10, KDE is applied to visualize the distribution of polarity scores across our assigned aspects. The plot shows the distribution of polarity scores for each aspect. The peaks observed in the KDE plot represent regions of high density which indicated areas where polarity scores are more concentrated and the most frequent sentiment intensities within each category. For instance, the globalization aspect (blue curve), the primary peak is observed around a polarity score of 0.70, with a secondary, lower-density concentration near -0.70. This bimodal distribution suggests that while the majority

```python
def scorer(score):
  if score >= 0:
    return 'positive'
  elif score < 0:
    return 'negative'

df['Sentiment'] = df['polarityscore'].apply(lambda i: scorer(i))
df.head(30
        )
```

| | comments | preprocessed_comment | assigned_aspect | sentiment | polarityscore | Sentiment |
|---|---|---|---|---|---|---|
| 0 | I learned arabic in saudi for years and chin... | [learned, arabic, saudi, years, chinese, puton... | globalization | very negative | -0.346632 | negative |
| 1 | This one thing I would love to see A Saud... | [one, thing, would, love, see, saudi, man, spe... | language_learning | very negative | -0.036114 | negative |
| 2 | dont see problems with it, go meet their milli... | [dont, see, problems, go, meet, millions, arab... | geopolitical_relations | very negative | -0.506627 | negative |
| 3 | but next time if abhinandan go to pakistan , h... | [next, time, abhinandan, go, pakistan, hardly,... | cultural_exchange | very negative | -0.743522 | negative |
| 4 | moroad they know lol even saudis teaches y... | [moroad, know, lol, even, saudis, teaches, yog... | language_learning | very positive | 0.467437 | positive |
| 5 | China will learn arabic | [china, learn, arabic] | globalization | positive | 0.312751 | positive |
| 6 | Long live china muslim alliance | [long, live, china, muslim, alliance] | globalization | very positive | 0.709120 | positive |
| 7 | Good learning language of other countries is a... | [good, learning, language, countries, handy, g... | globalization | positive | 0.859616 | positive |
| 8 | I thought the saudis were nationalist enough t... | [thought, saudis, nationalist, enough, promote... | cultural_exchange | very negative | -0.680727 | negative |
| 9 | Wow it is great \nWill ne the language of t... | [wow, great, ne, language, world, future, agre... | language_learning | very positive | 0.862899 | positive |

**Figure 7.** Code to Assign Sentiment Polarity Score

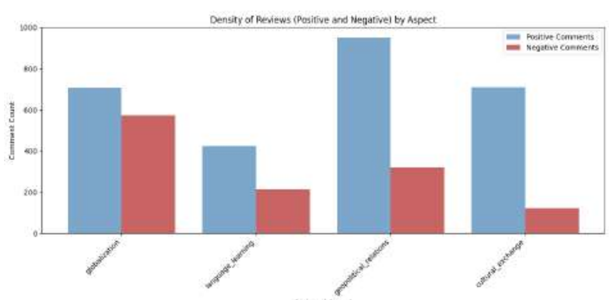


**Figure 8.** Binary Sentiment Density (Positive vs. Negative) by Aspect

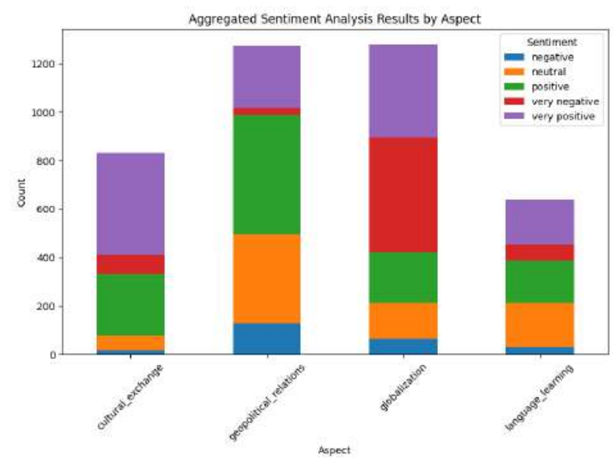


**Figure 9.** Aggregated 5-Class Sentiment Distribution by Aspect

of globalization-related discourse is highly positive, there is a distinct, significant group of highly negative opinions. The red curve for cultural aspect shows the highest density peak in the entire dataset, reaching a value above 1.5 at approximately 0.70 polarity. This high peak indicates extremely high concentration of strongly positive sentiment regarding cultural interactions. For language learning and geo-political relations (orange curve and green curve respectively) we see the highest densities in the positive region (between 0.50 and 0.80). However, language learning and adoption show a wider spread in neutral-to-negative regions (around -0.20), meaning a greater spread of the language in the neutral-to-negative regions (around -0.20).

While the KDE plot (Figure 10) shows the smooth probability per aspect, this histogram 11 shows the raw frequency of scores for the entire 4,022-comment dataset. The histogram clearly shows that the bars are much taller on the right side (positive) than on

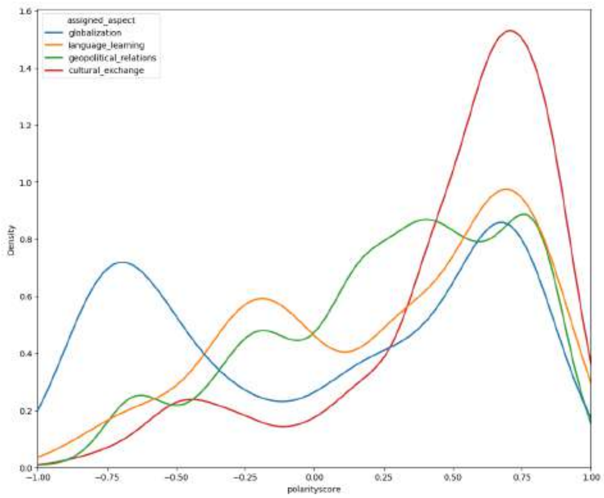


**Figure 10.** KDE Plot showing Sentiment Polarity Distribution

the left (negative). This shows that the general public consensus is favorable and dataset is polarized i.e. people generally have a strong opinion (either positive or negative) and are rarely "indifferent" about the language policy. A wide spread from -1.0 to +1.0 proves that our corpus is diverse and contains a full spectrum of human emotion.

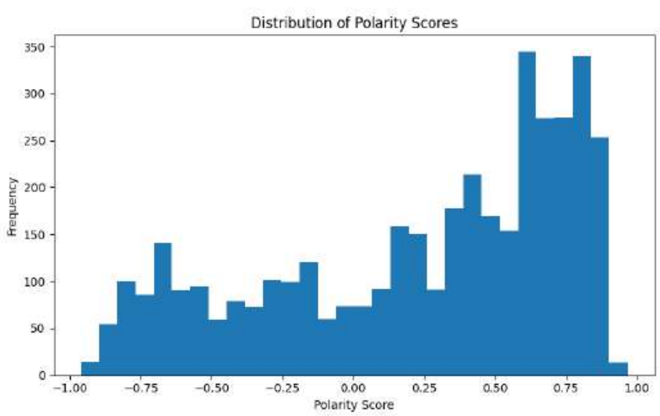


**Figure 11.** Polarity Score Distribution Histogram

The figure 12 shows the correlation heatmap(Pearson analysis) which shows the statistical relation between different sentiment labels in our corpus. This figure help us understand the relationship between different sentiment categories within our corpus. The high correlation values between 'negative,' 'neutral,' and 'positive' categories (ranging from 0.81 to 0.92)

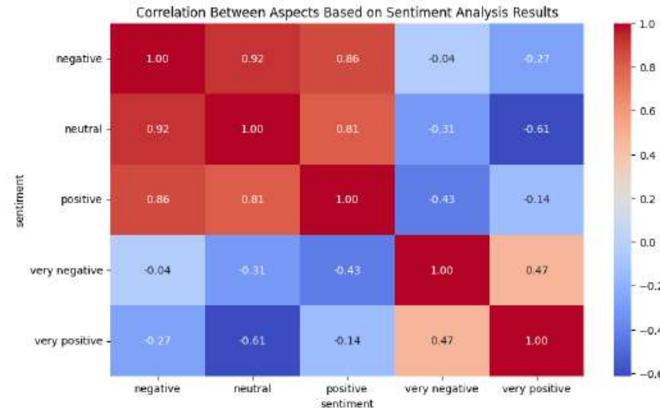


**Figure 12.** Correlation Heatmap of Sentiment Categorie

indicate that many user comments are multi-faceted, containing a mix of different emotional tones within a single entry. Conversely, the negative correlation between 'very positive' and 'neutral' (-0.61) suggests that highly enthusiastic comments tend to be polarized and rarely contain neutral or ambiguous language.

This analysis proves the complexity of the public's opinion in our case study. The high correlation between positive and negative sentiments suggests that users aren't just on any one side rather they are writing long, detailed comments that weigh both pros and cons. This also justifies why we used an Aspect-Based model. Since sentiments are so highly correlated, a simple global sentiment score would fail so we needed the aspect-based approach to separate which part of the comment is positive and which part is negative.

So far the results we showed discusses the general shape and trends of the data and sentiment densities, figure 13 provides the exact numerical intersection between the selected themes and the intensity of the public's emotion. It shows the granular numerical breakdown of the KEABSA results. In figure we see that Cultural Exchange aspect got the highest concentration of extreme enthusiasm, with 421 comments categorized as 'very positive.' Globalization is the most discussed topic, it also contains the highest volume of 'very negative' sentiment (475 comments). We can interpret this as as the scale of discussion moves from cultural ties to global structural changes, public

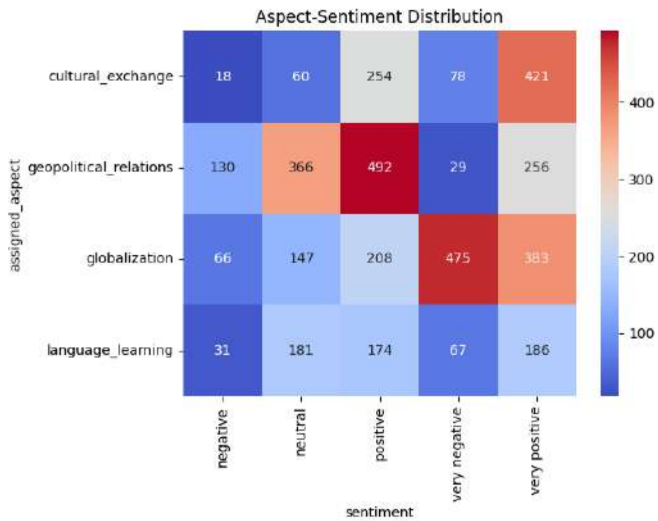


**Figure 13.** Aspect-Sentiment Distribution Heatmap showing numerical frequency of sentiment intensities across the four thematic aspects.

opinion becomes significantly more polarized.

## 6 Conclusion

This research demonstrated the application of advanced natural language processing techniques to analyze multilingual social media data within a low-resource and culturally nuanced domain. By integrating keyword-enhanced aspect classification with BERT-based sentiment analysis, we established an effective framework for extracting topic-specific sentiment trends from noisy and unstructured YouTube comments. The approach enabled automated identification of correlations between sentiment categories and the distribution of polarity scores across different thematic aspects, as illustrated in our correlation and KDE analyses. From an NLP perspective, this work contributes to the growing body of research on aspect-based sentiment analysis in underrepresented languages and domains. Although we see some challenges of handling dialectal variation, imbalanced sentiment classes, and contextual ambiguity in user-generated Arabic content. In the future the work may focus on improving domain adaptation for transformer models, refining aspect extraction through unsupervised topic modeling, and expanding datasets to support more robust cross-lingual sentiment detection. Our findings highlights the significance and potential of AI-driven language models in advancing multilingual sentiment understanding and enhancing computational methods for opinion mining in emerging digital contexts.

## Compliance with Ethical Standards

It is declared that all authors do not have any conflict of interest. Furthermore, informed consent was obtained from all individual participants included in the study.

## Author Contributions:

**Farzeen Ashfaq**: Writing—original draft preparation and methodology. **NZ Jhanjhi**: Supervision, reviewing, and editing. **Sayan Kumar Ray**: Conceptualization. **Husham M. Ahmed**: Investigation, data curation, and visualization.

## References


[1] N. Al-Tamimi, “China and Saudi Arabia,” in *Routledge Handbook on China–Middle East Relations*. Routledge, 2021.

[2] S. Foley, “Saudi Arabia’s Relations with China,” in *Routledge Companion to China and the Middle East and North Africa*. Routledge, 2023.

[3] E. Olander, “China-Saudi Trade is Booming... And Not Just Because of Oil,” *China Global South*, Dec. 2022. [Online]. Available: https://chinaglobalsouth.com/2022/12/07/china-saudi-trade-is-booming-and-not-just-because-of-oil/

[4] G. She, “Changing Focus While Maintaining Balance: Strategic Adjustments Behind the Developing Sino-Saudi Relations,” *Mediterranean Politics*, pp. 1–10, 2023.

[5] L. Zhou, “Chinese Investments in the Special Economic Zones in the Gulf Region: New Structural Economics Perspective,” in *Social Change in the Gulf Region: Multidisciplinary Perspectives*, pp. 531–546, 2023.

[6] S. Niu and D. Wang, “Three Summits’ and the New Development of China-Arab States Relations in the New Era,” *Asian Journal of Middle Eastern and Islamic Studies*, vol. 17, pp. 15–30, 2023.

[7] Oxford Analytica, “Centralised Saudi Rule Will Help Meet Economic Targets,” *Emerald Expert Briefings*, 2023.

[8] J. Fulton, "Situating Saudi Arabia in China's Belt and Road Initiative," *Asian Politics & Policy*, vol. 12, pp. 362–383, 2020.

[9] WION Video Team, "Saudi Schools to Teach Chinese Language; China, Saudi Education Ministries Sign MoU," 2022. [Online]. Available: https://www.wionews.com/videos/saudi-schools-to-teach-chinese-language-china-saudi-education-ministries-sign-mou-542953

[10] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer Nature, 2022.

[11] M. Wankhade, A. Rao, and C. Kulkarni, "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artificial Intelligence Review*, vol. 55, pp. 5731–5780, 2022.

[12] M. Humayun, M. Khalil, S. Almuayqil, and N. Jhanjhi, "Framework for Detecting Breast Cancer Risk Presence Using Deep Learning," *Electronics*, vol. 12, p. 403, 2023.

[13] N. Jhanjhi, "Comparative Analysis of Frequent Pattern Mining Algorithms on Healthcare Data," in *2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2024, pp. 1–10.

[14] M. Trușcă and F. Frasincar, "Survey on Aspect Detection for Aspect-Based Sentiment Analysis," *Artificial Intelligence Review*, vol. 56, pp. 3797–3846, 2023.

[15] N. Jhanjhi, "Investigating the Influence of Loss Functions on the Performance and Interpretability of Machine Learning Models," in *International Conference on Mathematical Modeling and Computational Science*, 2025, pp. 480–491.

[16] H. Li, "Latent Dirichlet Allocation," in *Machine Learning Methods*, pp. 439–471, 2023.

[17] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework for Sentiment Analysis of Tweets," *Sensors*, vol. 23, no. 1, p. 506, 2023.

[18] M. Asghar, A. Khan, S. Ahmad, M. Qasim, and I. Khan, "Lexicon-Enhanced Sentiment Analysis Framework Using Rule-Based Classification Scheme," *PLOS ONE*, vol. 12, p. e0171649, 2017.

[19] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule-Based Approach: A Case Study," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, 2020, pp. 537–542.

[20] P. Berka, "Sentiment Analysis Using Rule-Based and Case-Based Reasoning," *Journal of Intelligent Information Systems*, vol. 55, pp. 51–66, 2020.

[21] M. Ahmad, S. Aftab, S. Muhammad, and S. Ahmad, "Machine Learning Techniques for Sentiment Analysis: A Review," *International Journal of Multidisciplinary Sciences and Engineering*, vol. 8, p. 27, 2017.

[22] N. Dang, M. Moreno-García, and F. Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, p. 483, 2020.

[23] A. Yadav and D. Vishwakarma, "Sentiment Analysis Using Deep Learning Architectures: A Review," *Artificial Intelligence Review*, vol. 53, pp. 4335–4385, 2020.

[24] D. Yuan, Y. Zhou, R. Li, and P. Lu, "Sentiment Analysis of Microblog Combining Dictionary and Rules," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014, pp. 785–789.

[25] A. Rahat, A. Kahir, and A. Masum, "Comparison of Naive Bayes and SVM Algorithm Based on Sentiment Analysis Using Review Dataset," in *2019 8th International Conference on System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 266–270.

[26] K. Han, W. Chien, C. Chiu, and Y. Cheng, "Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter Dataset," *Applied Sciences*, vol. 10, p. 1125, 2020.

[27] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for Situations Understanding Based on Sentiment Analysis of Twitter Data," *Procedia Computer Science*, vol. 111, pp. 376–381, 2017.

[28] J. Wang, L. Yu, K. Lai, and X. Zhang, "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 225–230.

[29] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer Based Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis," *Future Generation Computer Systems*, vol. 113, pp. 58–69, 2020.

[30] K. Tan, C. Lee, K. Anbananthen, and K. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis with Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21 517–21 525, 2022.

[31] H. Do, P. Prasad, A. Maag, and A. Alsadoon, “Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review,” *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019.

[32] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, “A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, pp. 11 019–11 038, 2022.

[33] M. Mowlaei, M. Abadeh, and H. Keshavarz, “Aspect-Based Sentiment Analysis Using Adaptive Aspect-Based Lexicons,” *Expert Systems with Applications*, vol. 148, p. 113 234, 2020.

[34] J. Yu and D. Chauhan, “Trends in NLP for Personalized Learning: LDA and Sentiment Analysis Insights,” *Education and Information Technologies*, vol. 30, pp. 4307–4348, 2025.

[35] S. Li, “Scenarios for the Development of China Studies in Saudi Arabia in Correspondence with Contemporary Saudi Values,” *Asian Journal of Middle Eastern and Islamic Studies*, vol. 14, pp. 581–593, 2020.

[36] H. Alshammari, “Chinese Language in Saudi Arabia: Challenges and Recommendations,” *English Language Teaching*, vol. 13, pp. 75–85, 2020.

[37] W. Li and M. Zhang, “Chinese Language Teaching in Saudi: Status Analysis, Problems and Countermeasures,” in *Proceedings of the 6th International Conference on Modern Management and Education Technology (MMET 2021)*, 2021, pp. 237–241.

[38] M. Al Roomy, “Chinese Language Teaching in a Saudi Context: A Brief Review,” in *Conference on Multidisciplinary Research (MyRes)*, 2021, p. 130.

[39] R. Catelli, S. Pelosi, C. Comito, C. Pizzuti, and M. Esposito, “Lexicon-based sentiment analysis to detect opinions and attitude towards COVID-19 vaccines on Twitter in Italy,” *Computers in Biology and Medicine*, vol. 158, p. 106 876, 2023.

[40] B. Ainapure *et al.*, “Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches,” *Sustainability*, vol. 15, p. 2573, 2023.

[41] H. Hoiriyah, H. Mardiana, M. Walid, and A. Darmawan, “Lexicon-Based and Naive Bayes Sentiment Analysis for Recommending the Best Marketplace Selection as a Marketing Strategy for MSMEs,” *Jurnal Pilar Nusa Mandiri*, vol. 19, pp. 65–76, 2023.

[42] A. Raza, A. Habib, J. Ashraf, B. Shah, and F. Moreira, “Semantic orientation of crosslingual sentiments: Employment of lexicon and dictionaries,” *IEEE Access*, vol. 11, pp. 7617–7629, 2023.

[43] E. Schubert, “The aesthetic emotion lexicon: A literature review of emotion words used by researchers to describe aesthetic experiences,” *Empirical Studies of the Arts*, vol. 42, pp. 3–37, 2024.

[44] A. Saffar, T. Mann, and B. Ofoghi, “Textual emotion detection in health: Advances and applications,” *Journal of Biomedical Informatics*, vol. 137, p. 104 258, 2023.

[45] A. Alslaity and R. Orji, “Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions,” *Behaviour & Information Technology*, vol. 43, pp. 139–164, 2024.

[46] A. Mostafa, M. Aljasir, M. Alruily, A. Alsayat, and M. Ezz, “Innovative Forward Fusion Feature Selection Algorithm for Sentiment Analysis Using Supervised Classification,” *Applied Sciences*, vol. 13, p. 2074, 2023.

[47] A. Ghorbanali and M. Sohrabi, “A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis,” *Artificial Intelligence Review*, vol. 56, pp. 1479–1512, 2023.

[48] G. Kaur and A. Sharma, “A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis,” *Journal of Big Data*, vol. 10, p. 5, 2023.

[49] M. Başarslan and F. Kayaalp, “MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis,” *Journal of Cloud Computing*, vol. 12, p. 5, 2023.

[50] R. Alhujaili and W. Yafooz, “Sentiment analysis for YouTube videos with user comments,” in *Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 814–820.

[51] L. Mai and B. Le, “Joint sentence and aspect-level sentiment analysis of product comments,” *Annals of Operations Research*, vol. 300, pp. 493–513, 2021.

[52] R. Bose, R. Dey, S. Roy, and D. Sarddar, “Sentiment analysis on online product reviews,” in *Information and*

*Communication Technology for Sustainable Development (ICT4SD 2018)*, 2020, pp. 559–569.

[53] R. Sandoval-Almazan and D. Valle-Cruz, "Sentiment analysis of Facebook users reacting to political campaign posts," *Digital Government: Research and Practice*, vol. 1, pp. 1–13, 2020.

[54] N. Ashraf, A. Zubiaga, and A. Gelbukh, "Abusive language detection in YouTube comments leveraging replies as conversational context," *PeerJ Computer Science*, vol. 7, p. e742, 2021.

[55] M. Khan *et al.*, "Sentiment analysis on Bengali Facebook comments to predict fans' emotions towards a celebrity," *Journal of Engineering Advancements*, vol. 2, pp. 118–124, 2021.

[56] H. Hosseinmardi *et al.*, "Examining the consumption of radical content on YouTube," *Proceedings of the National Academy of Sciences*, vol. 118, p. e2101967118, 2021.

[57] R. Bringula, J. Victorino, M. De Leon, and M. Estuar, "Cluster and sentiment analyses of YouTube textual feedback of programming language learners," in *Proceedings of the Future Technologies Conference (FTC 2019), Volume 2*, 2020, pp. 913–924.

[58] I. Dubovi and I. Tabak, "An empirical analysis of knowledge co-construction in YouTube comments," *Computers & Education*, vol. 156, p. 103 939, 2020.

[59] J. Melba Rosalind and S. Suguna, "Predicting students' satisfaction towards online courses using aspect-based sentiment analysis," in *International Conference on Computer, Communication, and Signal Processing*, 2022, pp. 20–35.

[60] D. Musleh *et al.*, "Arabic Sentiment Analysis of YouTube Comments: NLP-Based Machine Learning Approaches for Content Evaluation," *Big Data and Cognitive Computing*, vol. 7, p. 127, 2023.

[61] A. Wisnubroto *et al.*, "Opinion-based sentiment analysis related to 2024 Indonesian Presidential Election on YouTube," in *5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 2022, pp. 318–323.

[62] K. McGarry, "Analyzing Social Media Data Using Sentiment Mining and Bigram Analysis for the Recommendation of YouTube Videos," *Information*, vol. 14, p. 408, 2023.

[63] I. Udayana, I. Indrawan, and I. Putra, "Decision Support System for Sentiment Analysis of YouTube Comments on Government Policies," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 5, pp. 27–37, 2023.

[64] R. Bringula, R. Tabo, F. Alcazar, J. Delica, and J. Sayson, "YouTube Videos on the Achievements of Presidential Candidates: Sentiment and Content Analysis," *Journal of Political Marketing*, pp. 1–17, 2023.

[65] K. Patil *et al.*, "Analysis of YouTube Comments to Inform the Design of Virtual Reality Training Simulations to Target Emotional Arousal," *Journal of Construction Engineering and Management*, vol. 149, p. 04023077, 2023.

[66] M. Ftanou *et al.*, "Suicide prevention public service announcements (PSAs): Examples from around the world," *Health Communication*, vol. 32, pp. 493–501, 2017.

[67] H. Alawadh, A. Alabrah, T. Meraj, and H. Rauf, "English Language Learning via YouTube: An NLP-Based Analysis of Users' Comments," *Computers*, vol. 12, p. 24, 2023.

[68] F. Ashfaq, *YouTube Comments Dataset for Sentiment Analysis*, Zenodo, 2024, doi: 10.5281/zenodo.18150330.

[69] F. Ashfaq, *YouTube Comments Dataset for Sentiment Analysis (Version 2)*, Zenodo, 2025, doi: 10.5281/zenodo.18210804.