big data and cognitive computing

MDPI

Article

Fit Talks: Forecasting Fitness Awareness in Saudi Arabia Using Fine-Tuned Transformers

Nora Alturayeif [1,*,†], Deemah Alqahtani [1,†], Sumayh S. Aljameel [2], Najla Almajed [1], Lama Alshehri [1], Nourah Aldhuwaihi [1], Madawi Alhadyan [1] and Nouf Aldakheel [1]

[1] Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia; daalqahtani@iau.edu.sa (D.A.)
[2] Aramco Saudi Accelerated Innovation Lab (aramcoSAIL), Saudi Aramco, Dhahran 31311, Saudi Arabia
[*] Correspondence: nsalturayeif@iau.edu.sa
[†] These authors contributed equally to this work.

Abstract: Understanding public sentiment on health and fitness is essential for addressing regional health challenges in Saudi Arabia. This research employs sentiment analysis to assess fitness awareness by analyzing content from the X platform (formerly Twitter), using a dataset called Saudi Aware, which includes 3593 posts related to fitness awareness. Preprocessing steps such as normalization, stop-word removal, and tokenization ensured high-quality data. The findings revealed that positive sentiments about fitness and health were more prevalent than negative ones, with posts across all sentiment categories being most common in the western region. However, the eastern region exhibited the highest percentage of positive sentiment, indicating a strong interest in fitness and health. For sentiment classification, we fine-tuned two transformer architectures—BERT and GPT—utilizing three BERT-based models (AraBERT, MARBERT, CAMeLBERT) and GPT-3.5. These findings provide valuable insights into Saudi Arabian attitudes toward fitness and health, offering actionable information for public health campaigns and initiatives.

Keywords: natural language processing; transformer-based models; sentiment analysis; deep learning

check for updates

Academic Editor: Alberto Fernandez Hilario

Received: 6 December 2024
Revised: 10 January 2025
Accepted: 16 January 2025
Published: 23 January 2025

Citation: Alturayeif, N.; Alqahtani, D.; Aljameel, S.S.; Almajed, N.; Alshehri, L.; Aldhuwaihi, N.; Alhadyan, M.; Aldakheel, N. Fit Talks: Forecasting Fitness Awareness in Saudi Arabia Using Fine-Tuned Transformers. *Big Data Cogn. Comput.* **2025**, *9*, 20. https://doi.org/10.3390/bdcc9020020

1. Introduction

Recently, there has been an increase in interest in using natural language processing (NLP) and machine learning algorithms to assess feelings posted on social networks [1–3]. Sentiment analysis is an effective method for analyzing the public's views and sentiments. In social networks, sentiment analysis has been used to assess individuals' opinions on topics such as COVID-19 preventive measures [4,5], consumer perceptions of different brands, and predicting election results [6].

In the context of health and fitness, sentiment analysis was used to evaluate people's reviews of persuasive features of mobile fitness apps [7]. It was also implemented to analyze the public tweets of people who share their fitness insights [8]. Although sentiment analysis offers promising information on public views on health and fitness, relatively few studies have focused specifically on people's perceptions, opinions, and experiences related to their health and well-being. Saudi Arabia has witnessed a significant transformation in the lifestyle of people, with a greater emphasis on health and fitness [9]. As a result, it is critical to understand the standard views among individuals about the knowledge about health and fitness. Traditional survey methods are often time consuming and have limitations in capturing the changing nature of views [10]. Classical machine learning methods such as Support Vector Machines (SVM) can leverage real-time data processing and enhance accuracy in sentiment classification [11]. However, such methods often rely on manual feature extraction and selection. Recent advances in transformer-based models such as BERT, AraBERT, CAMeLBERT, T0, and T5 allowed the processing of large amounts of text data, automatic acquisition of contextual relationships and the extraction of valuable insights [12,13]. This has boosted the ability of sentiment analysis to produce more accurate findings [14].

This study investigates the efficacy of several transformer-based models in predicting people's health and fitness awareness in different regions in Saudi Arabia. Using the Arabic language resources and pre-trained models, we utilize embeddings to capture the nuances of sentiment in Arabic text. The proposed approach involves collecting an extensive dataset of X platform posts that contain online reviews and other textual data relevant to the awareness of health and fitness. The acquired data was pre-processed and fed into transformer-based sentiment analysis algorithms. The labeled data were used to fine-tune the models and capture sentiments about Saudi Arabia's cultural environment. The findings of this study have important implications for various stakeholders, including healthcare providers, policy makers, and experts from the fitness industry. Understanding each individual's sentiments could help establish customized tactics to increase health and fitness awareness, personalize healthcare services, and create compelling marketing efforts, thus inspiring and motivating the future of the fitness industry [15]. The contributions of this paper are as follows.

- We constructed Saudi Aware, the first publicly available dataset featuring 3593 geo-tagged posts from the X platform, focused on health and fitness awareness and categorized by Saudi Arabia's five main regions.
- We leveraged the Saudi Aware dataset to fine-tune two Transformer architectures to assess health and fitness awareness among individuals in Saudi Arabia.
- We deployed our model in a user-friendly interface for real-time sentiment tracking.

2. Background and Related Work

In this section, we first discuss transformer-based architectures, explaining their relevance and applications in NLP. We then review existing literature on sentiment analysis, specifically focusing on studies that have worked with health and fitness data, highlighting the methodologies and findings in this area.

2.1. Transformer-Based Architectures

BERT, or Bidirectional Encoder Representations from Transformers, revolutionized NLP [16]. Through bidirectional understanding, it gained widespread recognition for its ability to capture the intricate contextual nuances of words in sentences. Unlike previous models that processed text in a unidirectional or shallowly bidirectional manner, BERT made significant strides in various NLP tasks such as text classification, named entity recognition, sentiment analysis, and question answering. One of the key features of BERT is its encoder-only architecture [17]. Designed as a bidirectional model, BERT learns to predict missing words in sentences (Masked Language Model) and determine whether pairs of sentences are consecutive in the original text (Next-sentence prediction) during its pre-training phase [16]. This process allows BERT to develop a deep understanding of the language context. Its exceptional performance is achieved through a two-stage process that involves pre-training on extensive text data corpora followed by fine-tuning on specific downstream tasks. BERT's encoder-only architecture enables it to excel across various NLP tasks. Although BERT is language-agnostic by nature, various adaptations and variants have emerged to address language-specific challenges [16]. For instance, CAMeLBERT and AraBERT have been adapted for processing Arabic text. These models, equipped with BERT's architecture and pre-training techniques, effectively handle NLP tasks in Arabic by understanding the complexities of Arabic morphology, dialects, and orthography [18,19]. Similarly, MarBERT is a powerful deep bidirectional transformer-based model designed for diverse Arabic varieties. It is designed to achieve state-of-the-art results across multiple tasks, particularly excelling in social media tasks and topic classification. These variants of BERT signify significant progress in the field of NLP for their respective languages, successfully addressing the challenges arising from linguistic diversity and complexity. Researchers and developers can harness the robust encoder-only architecture and pre-training techniques of BERT to construct advanced language understanding systems tailored to specific languages and domains, thus empowering the development of sophisticated NLP applications [20].

GPT, or Generative Pre-trained Transformer, is a prominent type of neural network architecture introduced by OpenAI [21]. It has gained significant attention and has achieved remarkable performance in various NLP tasks. GPT is primarily known for its ability to generate coherent and contextually relevant text. At its core, GPT is based on the Transformer architecture, which comprises self-attention mechanisms and feedforward neural networks. The Transformer architecture has proven highly effective in capturing contextual relationships between words or tokens in a given input sequence. GPT leverages this architecture to model language and generate text. GPT is explicitly designed as a decoder-only architecture. Unlike the encoder-decoder structure commonly found in sequence-to-sequence models, GPT focuses on the decoding aspect of the Transformer, where it generates text based on the learned representation. The lack of an explicit encoder means that GPT is primarily used for tasks involving text generation, such as language modeling, text completion, and text generation conditioned on a given prompt. GPT training involves a two-step process: pre-training and fine-tuning. GPT is trained on a large corpus of unlabeled text data during the pre-training phase using a language modeling objective. It learns to predict the next word in a sequence of words given the context, thus acquiring a broad understanding of language patterns and structures. After pretraining, GPT can be fine-tuned on specific downstream tasks using task-specific labelled data. This fine-tuning process allows GPT to adapt its learned representation to the specific task at hand, enhancing its performance on tasks such as text classification, sentiment analysis, and question answering. GPT has achieved state-of-the-art results in various NLP benchmarks and has been widely used in research and practical applications. Its decoder-only architecture makes it particularly suitable for tasks that involve text generation or understanding based on context. Using the power of the Transformer architecture and its large-scale pre-training, GPT has demonstrated notable capabilities to generate human-like text and to understand complex language patterns [21].

In summary, BERT and GPT are two influential architectures in the field of NLP. BERT, with its encoder-only structure and bidirectional understanding, excels in capturing contextual nuances and performing a wide range of NLP tasks. On the other hand, GPT's decoder-only design focuses on text generation and understanding, making it ideal for tasks involving coherent and contextually relevant text. Both models have significantly advanced the state-of-the-art in NLP and have been widely adopted in various applications.

2.2. Sentiment Analysis in the Health and Fitness Domains

Sentiment analysis is a method that takes text input and uses machine learning and NLP to find and extract subjective information. Many studies have used sentiment analysis to understand public opinions and emotional responses in areas such as marketing, healthcare care, and social networks [22,23,23]. In the fields of health and fitness, sentiment analysis has been used to track public perceptions of fitness-related live-streaming [24], spot patterns in health-related activities [25], and determine the emotional resonance of wellness-related content posted on social media [26].

Vickey et al. [8] examined fitness tweets linked to mobile apps, analyzing sentiment and online influence using Klout scores, retweets, and follower counts. They found that most tweets were positive, with retweets and content being stronger influence predictors. Polimis [27] examined demographic disparities in attitudes towards physical activity by analyzing 830,000 tweets, finding women less positive than men. The study noted biases, lack of accuracy metrics, and ethical concerns about the use of social media data. Liu et al. [28] analyzed 442 million tweets to explore using Twitter data to track physical activity. They found links between geotagged tweets and activity levels. The study highlights the potential of social networks to monitor physical activity, but calls for further research to increase accuracy and overcome data biases.

L [29] used machine learning algorithms, including SVM and Naive Bayes, to classify mental health status based on Twitter data. By analyzing emotional phrases and patterns in tweets, the study showed the potential of social media and machine learning to identify mental health issues. Lee et al. [30] analyzed Twitter data from 2010 to 2016 to look at trends in health technology using sentiment analysis and ontology-based techniques. The study reported that “mHealth” is the most talked about topic, and most of the posts reflect positive sentiment. Limitations included demographic biases in Twitter users and the lack of detailed performance metrics for sentiment analysis. Pimenta et al. [31] reviewed fitness and nutrition mobile apps, coding features using Behavior Change Techniques (BCT) taxonomy. They also performed a sentiment analysis on 20,492 user reviews. Positive sentiments were related to the framing/reframing technique, while negative sentiments often involved reward and threat techniques. However, the study lacked details on the performance of sentiment analysis and was limited by potential reviewer bias.

ŞAHİN et al. [32] analyzed the impact of the COVID-19 pandemic on tweets related to physical activity in Turkey. Using sentiment analysis, the study compared the non-COVID and COVID periods. They found that the volume of positive and negative sentiment tweets increased during the COVID period. The study acknowledged the need for multidimensional sentiment analysis. Musleh et al. [33] employed machine learning techniques to detect depression in Arabic tweets. They found that Random Forest achieved the best performance with an accuracy of 82.39%. The authors noted difficulties with Arabic dialects and recommended working with larger datasets and exploring additional classifiers. Wanniarachchi et al. [34] examined fat stigma and body objectification on social media through sentiment analysis, word co-occurrence mapping, and qualitative thematic analysis. The findings of the sentiment analysis indicated that the discussions were mostly negative, with feelings such as disgust, sadness, and anger frequently directed at both genders. Although less common, positive sentiments included remarks that were encouraging or empathetic.

Although sentiment analysis has been used in various fields, there is still a noticeable gap in its application to understand awareness of health and fitness, especially among Arabic-speaking populations. Most existing studies either focus on broad health issues or apply sentiment analysis in completely different cultural contexts, which means we still lack localized insights into how people in Saudi Arabia perceive fitness. Traditional approaches like surveys are often slow and do not reflect people’s real-time thoughts, making it harder to keep up with constantly changing public sentiment. In addition, analyzing Arabic text presents unique challenges due to the variety of dialects and the informal language commonly used on social media. Many existing models struggle to handle these complexities, which can lead to inaccurate sentiment classification. To bridge this gap, we developed the Saudi Aware dataset and fine-tuned advanced transformer-based models specifically for Arabic sentiment analysis. By focusing on a culturally and linguistically relevant approach, our study fills an important gap and provides a practical way to track and understand fitness awareness in Saudi Arabia.

3. Methodology

The proposed methodology uses sentiment analysis techniques to estimate individuals' health and fitness awareness in Saudi Arabia. Understanding how emotions are represented in public discourse, particularly on the X platform, is considered essential as health and well-being become increasingly emphasized. By applying these methods to a collection of Arabic posts, tailored to the Saudi Arabian context, we aim to gain insights into the population's awareness and debates surrounding health and fitness.

The proposed methodology for sentiment analysis involves multiple stages, which are illustrated in Figure 1. The process begins with data collection from the X platform (formerly known as Twitter). The raw data then undergoes preprocessing steps that include normalization and tokenization to ensure consistency and compatibility with machine learning models. Next, the preprocessed data is fed into various transformer-based models, including AraBERT, CAMeLBERT, MARBERT, and GPT. These models are fine-tuned on the sentiment analysis task, and their performances are evaluated to classify the sentiments into three categories: positive, negative, and neutral. The final stage involves identifying the best-performing model based on performance metrics, which will be used for further analysis and application. The following sections will provide a detailed explanation of the data collection, preprocessing, annotation, and model development.

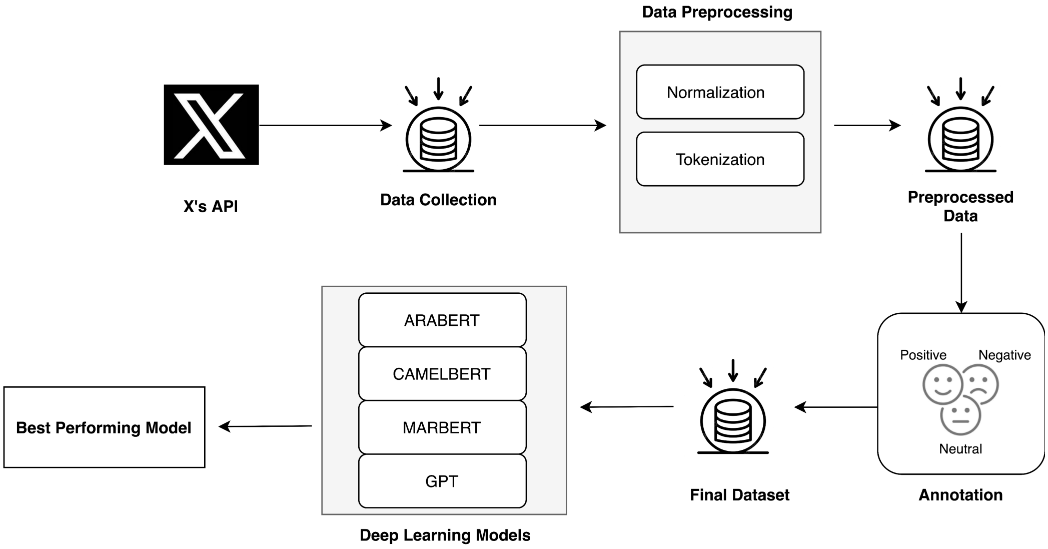


Figure 1. Overview of the proposed methodology.

4. Saudi Aware Dataset

In this section, we detail the process of collecting a dataset of opinions related to health and fitness. We also describe the crowdsourcing setup used for sentiment annotation and present the statistics of the collected dataset, which we refer to as the Saudi Aware dataset.

4.1. Data Collection

Our objective is to create a dataset that contains positive, negative, and neutral posts. We used the APIFY platform for collecting our dataset, which includes tools and infrastructure for data extraction, web scraping, and automation. We started scraping data using a set of keywords, such as الصحة (health), تمرين (exercise), النادي (gym), خطوات (steps), سعرات حرارية (calories), focusing on posts geotagged to specific regions in Saudi Arabia. For example, in the northern region, we targeted cities like Hail and Tabuk. Additionally, we specified that the language of the posts is Arabic. The total number of collected posts amounted to 33,438. Figure 2 illustrates the data collection process and the regions targeted.

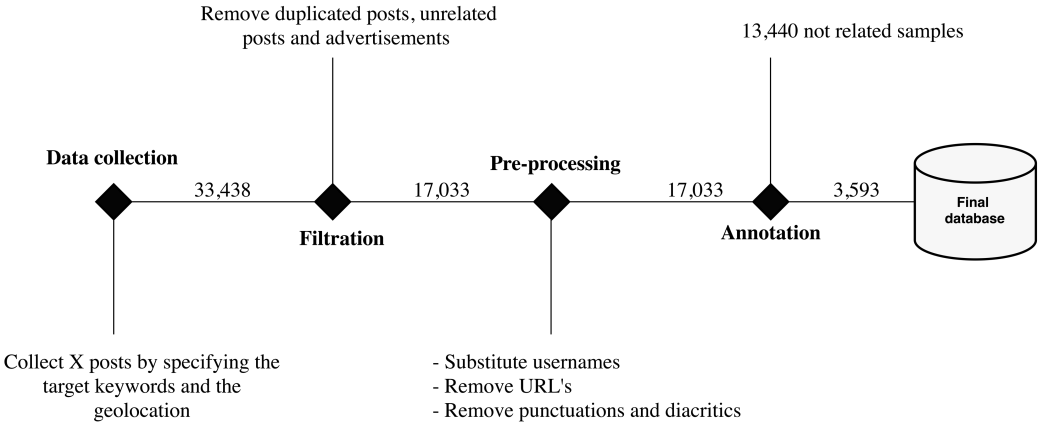

Remove duplicated posts, unrelated posts and advertisements
13,440 not related samples
Data collection
Pre-processing
33,438
17,033
17,033
3,593
Final database
Filtration
Annotation
Collect X posts by specifying the target keywords and the geolocation
- Substitute usernames
- Remove URL's
- Remove punctuations and diacritics

Figure 2. Data collection workflow for the Saudi Aware dataset.

4.2. Data Annotation

In this step, we utilized the Appen platform to label the posts as positive, negative, neutral, or unrelated. We assigned a minimum of three and a maximum of five annotators to evaluate each post and assign one of the labels based on their perspectives. With a confidence score set at 0.7, each post was presented to three random annotators for judgment. If there were differing opinions among the annotators, the post would be automatically redirected to an additional annotator. If we reached the maximum number of annotators without achieving agreement, the post would be discarded. Below, Table 1 showcases sample posts with annotations, providing insights into the annotation process.

Table 1. Examples illustrating the annotations of the Saudi Aware dataset.

	Post	Sentiment
1	صباح الخير بعد احلى كارديو مع الوالد Good morning after the best morning cardio with my father	Positive
2	الكيتو نظام غذائي علاجي وليس نظام غذائي صحي Keto is a therapeutic diet, not a healthy diet	Neutral
3	كل المقاسات متوفرة تسوي دايت ليش All sizes are available why would you go on a diet	Negative
4	اول خطوات السلام النفسي هي التغاضي First step to find inner peace is forgiveness	Not Related

4.3. Dataset Statistics

After finalizing the dataset, which includes 17,033 posts, we conducted an analysis to determine the distribution of sentiment across different regions of Saudi Arabia. In each region—Eastern, Middle, Northern, Southern, and Western—we identified varying numbers of positive, negative, neutral, and unrelated posts (Table 2). After aggregating the results across all regions and excluding unrelated posts, the dataset comprised 3046 positive posts, 383 negative posts, and 164 neutral posts. Our analysis revealed that the western region has the highest number of posts across all sentiment categories—positive, negative, and neutral—while the eastern region has the highest percentage of positive sentiment, reflecting a strong attitude toward health and fitness. The distribution of labels across each region is presented in Figure 3. These statistics offer valuable insights into the prevailing sentiments and the level of health and fitness awareness throughout Saudi Arabia.

Table 2. Data statistics of Saudi Aware dataset.

	Positive	**Negative**	**Neutral**	**Not Related**
North	341	27	12	2254
South	186	20	10	1034
East	482	99	23	1293
West	1290	143	88	4272
Middle	747	94	31	4587
Total	3046	383	164	13,440

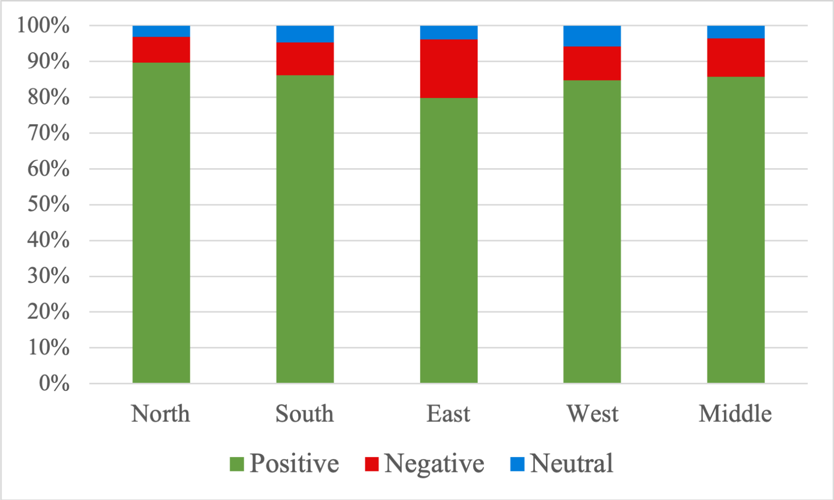


Figure 3. Labels' distribution in Saudi Aware dataset across five regions of Saudi Arabia.

5. Experimental Setup and Model Fine-Tuning

This section outlines the experimental setup used to fine-tune and evaluate transformer-based models for sentiment analysis in the context of fitness and health awareness in Saudi Arabia. We describe the models chosen for the task, the dataset used for training and testing, the hyperparameters selected to optimize performance, as well as the evaluation metrics used to assess the model's performance. The code used for fine-tuning, along with the training dataset, is publicly available (https://github.com/MadawiYousef/FitTalks-Sentiment-Analysis-for-health-and-fitness-awareness-in-Saudi-Arabia, accessed on 15 January 2025), ensuring transparency and reproducibility of the results.

Our goal is to fine-tune and assess the performance of multiple models on the Saudi Aware dataset, which includes sentiment-labeled posts. Specifically, we fine-tuned three BERT-based models—AraBERT-twitter, MARBERT, and CAMeLBERT—along with GPT-3.5. While the selected transformer models are pre-trained on extensive and diverse corpora, they are not optimized for domain-specific tasks such as Arabic sentiment analysis in the health and fitness context. Fine-tuning pre-trained transformer models is essential to adapt them to task-specific features like regional dialects and health-related terminology.

We fine-tune all models on the Saudi Aware dataset, which is specifically curated for sentiment analysis in the health and fitness domain. The dataset was divided into training and testing subsets, with 80% of the data used for training and 20% for testing. While the Saudi Aware dataset dataset consists of 3593 records, it is tailored specifically to Arabic sentiment analysis in the health and fitness domain. The dataset underwent rigorous preprocessing and annotation to ensure quality and reliability. Additionally, transformer-based models such as AraBERT and GPT-3.5, which were used in this study, have been shown to perform well even on modestly sized datasets, thanks to their extensive pre-training on large corpora. These factors collectively make the dataset suitable for fine-tuning in this specific context.

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art model that has significantly advanced various NLP tasks such as text classification, machine translation, and question answering [17]. BERT is pre-trained on a large corpus of text data and is later fine-tuned for specific downstream tasks. Several variants of BERT have been developed to address specific languages and domains. In our experiments, we fine-tuned the following three BERT-based models:

- AraBERT-twitter is a BERT-based model that extends the training of AraBERT (v0.2) on 60 million Arabic tweets, enhancing its capabilities for sentiment analysis and other language tasks in social media contexts [19].
- MARBERT is a BERT-based model that focuses on both Modern Standard Arabic (MSA) and Dialectal Arabic (DA), trained on a substantial dataset comprising 15.6 billion tokens from 1 billion Arabic tweets, which enhances its ability to capture linguistic variations across dialects [35].
- CAMeLBERT is a BERT-based model that addresses MSA and DA, emphasizing a balanced training approach with 5.8 billion tokens, which allows it to perform well across various Arabic language tasks by integrating both dialectal and standard forms [18].

Additionally, GPT (Generative Pre-trained Transformer), developed by OpenAI, is another prominent language model that, although not specifically tailored for Arabic dialects, has been trained on a vast corpus of text data, including MSA, DA, and Classical Arabic [21]. GPT is significantly larger than the BERT-based models, featuring an impressive 175 billion parameters. In contrast, the largest of the BERT variants, such as AraBERT, MarBERT, and CAMeLBERT, range from 135 million parameters to 163 million parameters. This substantial difference in model size highlights the extensive resources required for GPT, enabling it to capture complex linguistic patterns across various languages, including Arabic.

In the fine-tuning process of both the BERT-based models and GPT-3.5, we adapt pre-trained transformer models to our specific task of sentiment classification. For the BERT-based models, the hidden representation of the [CLS] token, which captures the context of the entire input sequence, is passed through a feed-forward neural network followed by a Softmax function. This generates probabilities for each sentiment class (positive, negative, neutral), allowing the model to classify posts based on the given input. The fine-tuning involves updating the model's weights using a labeled dataset specific to our task, thereby optimizing its performance for sentiment analysis.

For GPT-3.5, we utilized the OpenAI API to fine-tune the model for our sentiment classification task. Although GPT-3.5 is primarily a generative model, it can adapt to classification tasks by generating outputs that correspond to specific sentiment categories. During the fine-tuning process, we set the temperature parameter to 0, as shown in Table 3. This ensures that the model produces deterministic and consistent outputs, focusing on accuracy and eliminating randomness in its predictions.

For both BERT-based models and GPT-3.5, the final classification is made by feeding the hidden representations into a Softmax layer, which outputs the predicted sentiment labels. Fine-tuning enables the models to capture task-specific linguistic patterns and optimize their performance for classifying health and fitness-related sentiments.

The key hyperparameters employed for training the BERT models are presented in Table 4. For the GPT-3.5 model, we utilized the hyperparameters shown in Table 3.

Table 3. Hyperparameters for training the GPT 3.5 model.

Parameter	Value
Batch size	5
Number of epochs	3
Temperatur	0
Learning rate	2

Table 4. Hyperparameters for training the BERT models

Parameter	Value
Maximum sequence length	128
Feature dimension	768
Batch size	32
Number of epochs	20
Dropout rate	0.1
Early stop patience	20
Optimizer	AdawW
Learning rate	2×10^{-5}
Weight decay	1×10^{-5}

Following the fine-tuning process, we evaluated each model's performance using Macro F1 scores, precision, recall, and accuracy, which are essential metrics for classification tasks [36,37]. The Macro F1 score is the harmonic mean of precision and recall, averaged across all classes, and provides a balanced measure of a model's ability to correctly classify instances from each class, regardless of class imbalance. Precision indicates the proportion of true positive predictions among all positive predictions made by the model, reflecting its accuracy in identifying relevant instances. Recall, on the other hand, measures the proportion of true positive predictions among all actual positive instances, highlighting the model's ability to capture all relevant instances. Accuracy represents the proportion of correctly classified instances out of the total instances. While a detailed breakdown of performance by individual sentiment classes (positive, negative, neutral) could add granularity, it was omitted to maintain the clarity and readability of the results table. The Macro F1 score sufficiently encapsulates the comparative performance of the models and aligns with the study's focus on overall classification effectiveness.

6. Experimental Results and Discussion

In this section, we present the results of our experiments and analyze the performance of the fine-tuned transformer-based models. We focus on four main metrics—precision, recall, Macro F1 score, and accuracy—to evaluate the models' ability to classify sentiments related to health and fitness awareness. Our analysis not only highlights the comparative performance of the models but also explores their strengths and limitations in the context of sentiment analysis for Arabic text. Finally, we discuss the real-world implications of the results and potential applications in public health awareness.

The results in Table 5 provide a clear comparison of the performance of the fine-tuned models on the Saudi Aware dataset. Among the BERT-based models, AraBERT-twitter and MARBERT show very similar performance, with AraBERT-twitter slightly outperforming MARBERT in terms of Macro F1 (30.82 vs. 30.79) and accuracy (85.95% vs. 85.81%). Both models demonstrate relatively low precision and recall, indicating that they struggle to achieve high classification performance, despite performing reasonably well in terms of accuracy. CAMeLBERT-da, while exhibiting a higher precision (30.65) compared to the other BERT models, shows a noticeable drop in accuracy (59.39%) and the lowest Macro F1 score (27.46). This suggests that CAMeLBERT-da may not generalize as effectively to this specific sentiment analysis task, particularly in capturing health and fitness awareness sentiments.

Table 5. Performance results for the fine-tuned models and non-transformer models on the Saudi Aware dataset.

Model	Precision	Recall	Macro F1	Accuracy
Naïve Bayes	22.34	24.12	23.19	57.89
SVM	25.67	27.45	26.54	58.34
AraBERT-twitter	28.65	33.33	30.82	85.95
MARBERT	28.64	33.28	30.79	85.81
CAMeLBERT-da	30.65	31.00	27.46	59.39
GPT-3.5	**55.21**	**58.13**	**55.87**	**89.23**

In contrast, the GPT-3.5 model significantly outperforms all three BERT-based models across all metrics, with a Macro F1 score of 55.87 and an accuracy of 89%. The GPT model's precision (55.21) and recall (58.13) are also much higher than those of the BERT models, demonstrating its superior ability to classify sentiments in a balanced and accurate manner. This strong performance likely stems from GPT's larger model size and extensive pre-training, which enable it to capture more nuanced patterns in the text data. The relatively lower performance of the BERT-based models can be attributed to their smaller size and the complexity of the dataset, which includes a mix of Arabic dialects and MSA. These models may struggle to generalize across diverse linguistic variations, leading to lower precision and recall.

To provide a more comprehensive comparison, we also evaluated two non-transformer models, Naïve Bayes and Support Vector Machines (SVM), on the Saudi Aware dataset using TF-IDF features. As shown in Table 5, these models demonstrated noticeably lower performance compared to transformer-based models, particularly in terms of Macro F1 score and accuracy. While simpler models like SVM and Naïve Bayes require fewer computational resources, their inability to effectively capture the complexities of Arabic text, including dialectal variations, limits their performance. These findings underscore the effectiveness of transformer-based architectures for sentiment classification tasks.

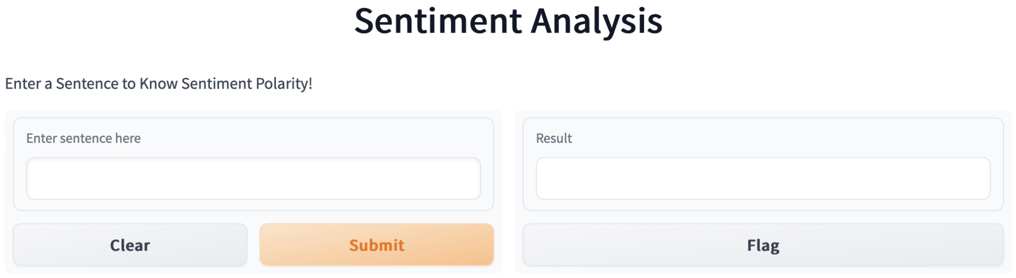
To further validate the generalizability of our models, we tested the top-performing models, AraBERT-twitter and GPT-3.5, on the "Mawqif" dataset [38]. The Mawqif dataset focuses on pressing social issues, including topics such as the "COVID-19 vaccine", "digital transformation", and "women's empowerment". Collected from the X platform, the dataset contains 4121 labeled Arabic posts, making it a suitable benchmark for sentiment analysis models. This dataset allows us to assess the robustness of the models in sentiment analysis tasks that extend beyond health and fitness. The performance results of AraBERT and GPT-3.5 on this external dataset are shown in Table 6. The results confirm the same conclusion drawn from the Saudi Aware dataset: GPT-3.5 consistently outperforms AraBERT-twitter across all metrics, with significantly higher precision, recall, Macro F1, and accuracy. This consistent performance reinforces GPT-3.5's advantage over BERT-based models for sentiment analysis in both health and social issues.

Table 6. Performance results for the fine-tuned models on the Mawqif dataset.

Models	Precision	Recall	Macro F1	Accuracy
AraBERT-twitter	42.9	41.2	36.2	41.2
GPT-3.5	**73.15**	**74.31**	**73.24**	**74.23**

Given GPT-3.5's superior performance in both accuracy and Macro F1 score, it was selected for deployment in a user-friendly application designed to provide real-time sentiment analysis of Arabic text. Leveraging Gradio, an open-source Python framework, we developed an interactive interface that enables users to input Arabic text and receive real-time sentiment predictions—categorized as positive, negative, or neutral. The interface, illustrated in Figure 4, prioritizes simplicity, allowing users to input text and obtain results with just one click. Designed to be highly intuitive, the tool is ideal for non-technical users such as public health officials and fitness professionals who wish to analyze public sentiment without requiring expertise in machine learning.

In addition, the interface includes a feature to flag erroneous predictions, enabling users to report inaccuracies, which can facilitate continuous model refinement. This deployment highlights the practical utility of advanced NLP models for real-world applications, offering a valuable tool for monitoring and analyzing fitness and health-related sentiments in Saudi Arabia.


Sentiment Analysis

Enter a Sentence to Know Sentiment Polarity!

Figure 4. Sentiment analysis interface.

7. Practical Implications for Public Health Campaigns

The insights gained from this study can be used effectively by public health officials to design region-specific campaigns that promote fitness and health awareness. Since the results revealed regional variations in public sentiment, such as the higher percentage of positive sentiment in the eastern region, campaigns can be tailored to reinforce positive attitudes in regions that already show strong interest while addressing potential barriers in regions with more neutral or negative sentiments. For example, in the eastern area, where positive sentiment is higher, campaigns can build on this enthusiasm with community events or fitness challenges to encourage participation. However, in regions with neutral or negative sentiment, efforts should focus on addressing barriers, such as limited access to facilities or misconceptions about fitness, through targeted education and awareness campaigns.

Additionally, the sentiment analysis tool developed in this study provides a practical way to monitor public sentiment in real time. By tracking how different regions respond to health initiatives, public health officials can fine-tune their messaging and strategies to ensure greater engagement and more effective promotion of fitness and health. This data-driven approach could lead to more targeted and culturally sensitive campaigns, ultimately improving public health outcomes throughout Saudi Arabia.

8. Limitations

This study, while providing valuable insights into health and fitness awareness in Saudi Arabia, has several limitations that should be considered when interpreting the results. Firstly, the dataset used, Saudi Aware, is based on geotagged posts from the X platform (formerly Twitter) and is limited to the content available on social media. As the data is self-reported by users, it may not fully represent the broader population's sentiment, particularly because the dataset is composed of posts that are publicly available. This means that the sample may be skewed toward individuals who are more active on social media, potentially overlooking the views of less engaged segments of the population. In addition, The dataset consists mainly of Arabic posts in MSA and various regional dialects. Although the study leverages Arabic-specific models like AraBERT, MARBERT, and CAMeLBERT, these models may still struggle with the complexities of dialects and informal language typical in social media. Dialectal variations are often underrepresented in training data, which, despite fine-tuning, can lead to potential misclassifications and reduced accuracy in capturing the nuanced sentiments expressed in slang and emotive language.

The decision to categorize sentiments into three classes—positive, negative, and neutral—was made to align with standard practices in related sentiment analysis studies and to simplify model training and evaluation. While effective, this approach may overlook subtle sentiment variations, such as degrees of positivity or negativity. Future research could explore the use of finer-grained sentiment categories (e.g., "slightly positive", "moderately negative") or multi-dimensional sentiment analysis frameworks to capture more nuanced sentiment patterns.

Furthemore, the study presents a snapshot of sentiment in Saudi Arabia at a particular point in time. A longitudinal study examining sentiment trends over a more extended period could provide a deeper understanding of how public attitudes toward health and fitness evolve, particularly in response to public health interventions or societal changes. While the Mawqif dataset provided valuable insights into the generalizability of our models, it is not specific to health-related topics. To the best of our knowledge, there are no publicly available Arabic datasets focused on health and fitness sentiment analysis. This limitation restricts our ability to validate our models across more domain-specific tasks. Addressing these limitations in future research, such as by expanding the dataset, incorporating more dialectal variations, exploring alternative model architectures, and developing or accessing health-related Arabic datasets, could further enhance the robustness and applicability of sentiment analysis in this field.

9. Conclusions

This study investigated the use of transformer-based models for sentiment analysis to evaluate health and fitness awareness in Saudi Arabia. We developed the Saudi Aware dataset, comprising posts from various regions, to examine public sentiment and regional differences. Our analysis revealed that positive sentiments about health and fitness outweigh negative ones across the dataset, with the western region having the highest volume of posts in all sentiment categories. In contrast, the eastern region demonstrated the highest percentage of positive sentiment, indicating a particularly strong interest in health and fitness. By constructing the Saudi Aware dataset and fine-tuning BERT-based models (AraBERT-twitter, MARBERT, CAMeLBERT) and GPT-3.5, we demonstrated the effectiveness of these models in classifying Arabic sentiments related to fitness. Our findings highlight the superior performance of GPT-3.5 across all metrics, significantly outperforming the BERT-based models in both the Saudi Aware and Mawqif datasets. This performance can be attributed to the larger model size and extensive pre-training of GPT-3.5, which enables it to capture more nuanced patterns in Arabic text.

The deployment of GPT-3.5 in a user-friendly sentiment analysis interface underscores the practical application of this model for real-time sentiment monitoring. This tool provides valuable insights for public health officials and fitness professionals, enabling targeted campaigns and interventions to promote fitness awareness.

Future work can focus on expanding the dataset to cover more regions and exploring additional aspects of public health sentiment. Further model enhancements, such as adapting GPT models for specific dialects, may improve performance in nuanced language contexts. Overall, our approach presents a powerful method for understanding public sentiment and promoting health awareness through the application of advanced NLP techniques.

Author Contributions: Conceptualization: N.A. (Nora Alturayeif), D.A. and S.S.A.; methodology: N.A. (Nora Alturayeif) and D.A.; software, validation, formal analysis, investigation: N.A. (Nora Alturayeif), N.A. (Najla Almajed), L.A., N.A. (Nourah Aldhuwaihi), M.A. and N.A. (Nouf Aldakheel); resources: N.A. (Nora Alturayeif), D.A., N.A. (Najla Almajed), L.A., N.A. (Nourah Aldhuwaihi), M.A. and N.A. (Nouf Aldakheel); data curation: N.A. (Najla Almajed), L.A., N.A. (Nourah Aldhuwaihi), M.A. and N.A. (Nouf Aldakheel); writing—original draft preparation: N.A. (Najla Almajed), L.A., N.A. (Nourah Aldhuwaihi), M.A. and N.A. (Nouf Aldakheel); writing—review and editing: N.A. (Nora Alturayeif) and D.A.; visualization: N.A. (Najla Almajed), L.A., N.A. (Nourah Aldhuwaihi), M.A. and N.A. (Nouf Aldakheel); supervision: N.A. (Nora Alturayeif), D.A. and S.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in GitHub at https://github.com/MadawiYousef/FitTalks-Sentiment-Analysis-for-health-and-fitness-awareness-in-Saudi-Arabia, accessed on 15 November 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Naresh, A.; Venkata Krishna, P. An efficient approach for sentiment analysis using machine learning algorithm. *Evol. Intell.* **2021**, *14*, 725–731. [CrossRef]
2. Kavitha, M.; Naib, B.B.; Mallikarjuna, B.; Kavitha, R.; Srinivasan, R. Sentiment analysis using NLP and machine learning techniques on social media data. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April 2022; pp. 112–115.
3. Pandey, K.K.; Thorat, M.; Joshi, A.; Srinivas, D.; Hussein, A.; Alazzam, M.B. Natural Language Processing for Sentiment Analysis in Social Media Marketing. In Proceedings of the 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 12–13 May 2023; pp. 326–330.
4. Aljameel, S.S.; Alabbad, D.A.; Alzahrani, N.A.; Alqarni, S.M.; Alamoudi, F.A.; Babili, L.M.; Aljaafary, S.K.; Alshamrani, F.M. A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. *Int. J. Environ. Res. Public Health* **2021**, *18*, 218. [CrossRef] [PubMed]
5. Comito, C. How Do We Talk and Feel About COVID-19? Sentiment Analysis of Twitter Topics. In Proceedings of the International Conference on Big Data, Delhi, India, 7–9 December 2023; pp. 95–107.
6. Ahad, M. A Review Article: Use of Sentiment Analysis in Social Media. *Int. J. Eng. Appl. Sci. Technol.* **2023**, *7*, 171–176. [CrossRef]
7. Nutrokpor, C.; Ekpezu, A.O.; Wiafe, A.; Wiafe, I. Exploring the impact of persuasive system features on user sentiments in health and fitness apps. In Proceedings of the 9th International Workshop on Behavior Change Support Systems, BCSS 2021, Aachen, Germany, 12 April 2021.
8. Vickey, T.; Breslin, J.G. Online influence and sentiment of fitness tweets: Analysis of two million fitness tweets. *JMIR Public Health Surveill.* **2017**, *3*, e8507. [CrossRef] [PubMed]
9. Saudi Vision 2030. Health Sector Transformation Program. 2024. Available online: https://www.vision2030.gov.sa/en/explore/programs/health-sector-transformation-program (accessed on 15 January 2025).
10. Das, A.; Prajapati, A.K.; Zhang, P.; Srinath, M.; Ranjbari, A. Leveraging Twitter Data for Sentiment Analysis of Transit User Feedback: An NLP Framework. *arXiv* **2023**, arXiv:2310.07086.

11. Rana, S.; Kanji, R.; Jain, S. Comparison of SVM and Naïve Bayes for Sentiment Classification using BERT data. In Proceedings of the 2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, India, 26–27 November 2022; pp. 1–5.
12. Bello, A.; Ng, S.C.; Leung, M.F. A BERT framework to sentiment analysis of tweets. *Sensors* **2023**, *23*, 506. [CrossRef] [PubMed]
13. Bashiri, H.; Naderi, H. Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowl. Inf. Syst.* **2024**, *66*, 7305–7361. [CrossRef]
14. Tan, K.L.; Lee, C.P.; Anbananthen, K.S.M.; Lim, K.M. RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access* **2022**, *10*, 21517–21525. [CrossRef]
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need *arXiv* **2017**, arXiv:1706.03762.
16. Koroteev, M.V. BERT: A Review of Applications in Natural Language Processing and Understanding. *arXiv* **2021**, arXiv:2103.11943.
17. Sushil, M.; Šuster, S.; Daelemans, W. Are We There Yet? Exploring Clinical Domain Knowledge of BERT Models. Technical Report. 2021. Available online: https://aclanthology.org/2021.bionlp-1.5/ (accessed on 16 December 2024).
18. Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The Interplay of Variant, Size, and Task Type in Arabic Pre-Trained Language Models. 2021. Available online: https://aclanthology.org/2021.wanlp-1.10/ (accessed on 15 January 2025).
19. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-Based Model for Arabic Language Understanding. 2020. Available online: https://aclanthology.org/2020.osact-1.2/ (accessed on 10 January 2025).
20. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. 2020. Available online: https://aclanthology.org/2021.acl-long.551/ (accessed on 15 December 2024).
21. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Technical Report. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 15 January 2025).
22. Hossain, I.; Puppala, S.; Alam, M.J.; Talukder, S.; Talukder, Z. A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries. In Proceedings of the International AAAI Conference on Web and Social Media, Buffalo, NY, USA, 3–6 June 2024; Volume 18, pp. 596–609.
23. Zhang, Q.; Yang, J.; Niu, T.; Wen, K.H.; Hong, X.; Wu, Y.; Wang, M. Analysis of the evolving factors of social media users' emotions and behaviors: A longitudinal study from China's COVID-19 opening policy period. *BMC Public Health* **2023**, *23*, 2230. [CrossRef] [PubMed]
24. Tian, R.; Yin, R.; Gan, F. Exploring public attitudes toward live-streaming fitness in China: A sentiment and content analysis of China's social media Weibo. *Front. Public Health* **2022**, *10*, 1027694. [CrossRef] [PubMed]
25. Raggatt, M.; Wright, C.J.; Carrotte, E.; Jenkinson, R.; Mulgrew, K.; Prichard, I.; Lim, M.S. "I aspire to look and feel healthy like the posts convey": Engagement with fitness inspiration on social media and perceptions of its influence on health and wellbeing. *BMC Public Health* **2018**, *18*, 1–11. [CrossRef]
26. Auxier, B.; Buntain, C.; Golbeck, J. Analyzing sentiment and themes in fitness influencers' twitter dialogue. In Proceedings of the International Conference on Information, Munich, Germany, 15–18 December 2019; pp. 429–435.
27. Polimis, K. Can Social Media Assess Demographic Variations in Physical Activity Attitudes? Technical Report. Available online: https://paa.confex.com/paa/2017/mediafile/ExtendedAbstract/Paper16293/Polimis_Twitter_Health_Attitudes_Demographic_Differences.pdf (accessed on 15 January 2025).
28. Liu, S.; Chen, B.; Kuo, A. Monitoring physical activity levels using twitter data: Infodemiology study. *J. Med. Internet Res.* **2019**, *21*, e12394. [CrossRef]
29. Rakshitha, C.L.; Gowrishankar, S. Machine Learning based Analysis of Twitter Data to Determine a Person's Mental Health Intuitive Wellbeing. *Int. J. Appl. Eng. Res.***2018**, *13*, 14956–14963.
30. Lee, J.; Kim, J.; Hong, Y.J.; Piao, M.; Byun, A.; Song, H.; Lee, H.S. Health information technology trends in social media: Using twitter data. *Healthc. Inform. Res.* **2019**, *25*, 99–105. [CrossRef]
31. Pimenta, F.; Lopes, L.; Gonçalves, F.; Campos, P. Designing Positive Behavior Change Experiences: A Systematic Review and Sentiment Analysis based on Online User Reviews of Fitness and Nutrition Mobile Applications. ACM International Conference Proceeding Series. In Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia, Essen, Germany, 22–25 November 2020; pp. 152–161. [CrossRef]
32. Şahin, T.; Gümüş, H.; Gencoglu, C. Analysis of Tweets Related with Physical Activity During COVID-19 Outbreak. *J. Basic Clin. Health Sci.* **2021**, *5*, 42–48. [CrossRef]
33. Musleh, D.A.; Alkhales, T.A.; Almakki, R.A.; Alnajim, S.E.; Almarshad, S.K.; Alhasaniah, R.S.; Aljameel, S.S.; Almuqhim, A.A. Twitter arabic sentiment analysis to detect depression using machine learning. *Comput. Mater. Contin.* **2022**, *71*, 3463–3477. [CrossRef]
34. Wanniarachchi, V.U.; Scogings, C.; Susnjak, T.; Mathrani, A. Fat stigma and body objectification: A text analysis approach using social media content. *Digit. Health* **2022**, *8*, 20552076221117404. [CrossRef]

35. Abdul-mageed, M.; Zhang, C.; Hashemi, A.; Moatez, E.; Nagoudi, B. AraNet: A Deep Learning Toolkit for Arabic Social Media. *arXiv* **2020**, arXiv:1912.13072.
36. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; pp. 1015–1021.
37. Zeng, G. On the confusion matrix in credit scoring and its analytical properties. *Commun. Stat.-Theory Methods* **2020**, *49*, 2080–2093. [CrossRef]
38. Alturayeif, N.S.; Luqman, H.A.; Ahmed, M.A.K. Mawqif: A Multi-label Arabic Dataset for Target-specific Stance Detection. *Assoc. Comput. Linguist.* **2022**, *12*, 174–184.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.