Sentiment Analysis for Women in STEM using Twitter and Transfer Learning Models

1st Shereen Fouad
School of Informatics and Digital Engineering
Aston University
United Kingdom
s.fouad@aston.ac.uk

2nd Ezzaldin Alkooheji
School of Informatics and Digital Engineering
Aston University
United Kingdom
@aston.ac.uk

***Abstract*—The science, technology, engineering and math (STEM) sector is integral to the nation's advancement and economy. However, the STEM workforce is perceived as male-dominant, and women are systematically tracked away from it. There has been a rising popularity of the gender disparity problem in STEM across social media platforms. Attitudes relating to women influence the careers women choose to pursue. It is thus timely and important to assess the public's opinion on this topic. This paper proposes a sentiment analysis classification framework that detects the sentiment of social media tweets in relation to women in STEM. To this end, we extracted more than 250,000 relevant tweets and used various open-language models to uncover insights into the perceptions of women in STEM using various open-language models. The study evaluates the performance of multiple machine learning and deep learning methods. We also study the performance of state-of-the-art transformer-based models, including bidirectional encoder representations from transformers (BERT), BERTweet, and TimeLMs (Time Language Models), which achieves 96% accuracy in sentiment detection. Results reveal that people's attitude in response to women in STEM is generally positive on the Twitter platform. However, we observed a significant correlation between positive sentiment in tweets and dates celebrating women's achievements (e.g. International Day of Women and Girls in Science, and International Women's Day). This finding demonstrates the impact of such campaigns on the public's opinion. Therefore, promoting these events among the public can encourage more females to pursue careers in STEM.**

***Index Terms*—Women In STEM, Machine Learning, Deep Learning, Transformers, Twitter, Sentiment Analysis, Natural Language Processing.**

I. Introduction

Science, Technology, Engineering and Mathematics (STEM) are one of the most significant fields contributing to the economy [1]. In the coming years, the STEM sector is projected to become one of the world's largest employers. In the UK, STEM makes up around half of the country's Gross Domestic Product (GDP), which measures the economy's health in a specific period [3], [4]. Similar impacts were also seen in the US, where STEM makes up to $13.5 trillion of its national GDP [5]. This has made the growth and demand for STEM jobs increase significantly.

Despite the flourishing industry, a gender disparity exists in the STEM field. Several studies have repeatedly reported that the STEM field is perceived as a male dominant, and women have been systematically tracked away from it [6], [7]. According to [8], women are underrepresented in the UK STEM workforce, making up only 24.4%, which is significantly lower than men's representation in the field. This issue is not only inclusive to the UK, but it affects many countries worldwide. For example, according to the American Association of University Women [9], women only comprise 28% of the US STEM workforce.

The widespread of social networking platforms, such as Twitter and Facebook, has enabled people to communicate their opinions and express their thoughts about various social topics. Due to the low cost of use and easy access, social media has become a powerful tool integrated into most people's everyday lives. Recently, there has been a rising popularity of the gender disparity problem in STEM across social media platforms [20]. This campaign resulted in the women in STEM movement (e.g. [2], [7], [8]). However, there have been reports of an increased spread of hateful and sexist content targeting women in STEM on social media platforms [10]. These negative posts can have an unfavourable impact on the attraction, retention and progression of girls and women in STEM studies and careers [7]. It is important to note here that many users form their opinions based on the information and contributions available on social media networks [11]. According to [12], teenage girls use social media platforms more than their male counter, with 50% of them being put into this classification, compared to 39% of teenage boys. With such a predominantly female audience, it is important to detect and address the negatives of social media posts, as they can play a significant role in females' career and education choices. This gives rise to the need to create a sentiment analysis method that can automatically identify sentiments about women in the STEM field on social media networks.

In this paper, we propose a sentiment classification approach to assess the general public opinion around women in STEM using Twitter's social media platform. To this end, we scrape around 250,000 tweets related to women in STEM. Twitter is one of the biggest social media platforms, with around 500 million tweets posted daily. This creates an information-rich reservoir of opinions shared by the public about this topic of interest. However, social media posts are often vague, heterogeneous, and hard to contextualise, which makes them challenging to analyse.

Sentiment analysis is a natural language processing (NLP) technique that determines a given text's sentiment classification and polarity [13]. Sentiment analysis is a key component of NLP, and it aims to recognize and detect the emotions included in subjective texts [15]. It is extensively used to analyse social media content and identify textual data's sentiment polarity. The sentiment polarity of a given text uncovers the emotions of the text's writer as being positive, negative, or neutral. The sentiment analysis task comprises three main steps: data pre-processing, feature extraction and classification. The classification step is often performed by Machine learning (ML) and deep learning (DL) techniques, which have proven to be effective in extracting the writer's sentiment as conveyed in the text. ML and DL tools have been used widely in the literature for the automatic analysis of social media content (e.g. [10], [13]–[16], [18]).

ML have been widely in the literature as a sentiment analysis tool that analyzes texts for polarity, from positive to negative. This is done by training ML algorithms with examples of emotions in text, and machines automatically learn how to detect sentiment without human input [10], [13], [15], [16], [18]. Supervised ML techniques require that the model be trained on labelled data before being evaluated on unseen data for determining the model's performance. The most commonly used labelling techniques for NLP tasks are VADER, TextBlob, and RoBERTa [23], [24]. However, many researchers argued that VADER provides a more granular sentiment than TextBlob as it considers capitalisation, repeated words, and emoji when evaluating the text's sentiment. On the other hand, RoBERTa is considered the state-of-the-art tool for sentiment labelling as it can catch the deep meaning of a text rather than individual words. Hence, it is the most effective tool, especially if the sentence does not exceed the model tokenisation limit [25].

DL uses artificial neural networks with multiple hidden layers to high-level model abstractions in data. DL algorithms have demonstrated impressive results in NLP applications. Two main DL methods are used in NLP applications, the feed-forward networks (FFN) and recurrent neural networks (RNNs), and both can be combined in some cases. The FFN uses multi-layer perceptions that take fixed-sized inputs or variable-length inputs. This allows MLPs to accept linear models, which would benefit from the neural network non-linearity and the ability to integrate pre-trained word embeddings, resulting in higher classification accuracies. The RNNs are used on sequential data; they accept many inputs to perform multiple tasks and output a fixed-size vector. RNNs are used as input-transformers trained to format output data for the FFN that operates on top of the RNN. Therefore, they are used in many NLP applications due to their impressive ability to understand language modelling and perform word predictions [13], [14]. LSTM is an RNN architecture that is used in several NLP applications. Unlike standard FFN, that process only single data points, LSTM has feedback connections and can process the entire sequences of data [19]. Bidirectional long-short-term memory (Bi-LSTM) is an extension of the LSTM, which allows the neural network to read the sequence of information in both directions, backward or forward.

The transformer neural network is a DL architecture that is now considered the state-of-the-art technique in the field of NLP [17], [22], [24], [26], [30]. Similar to RNNs, transformers aim to solve sequence-to-sequence tasks, such as natural language translation or text summarization. However, unlike RNNs, they can handle long-range dependencies with ease. They include an encoder-decoder model that leverages attention mechanisms to compute better embeddings and to better align output to input without using sequence-aligned RNNs. The attention mechanism provides context for any position in the input sequence.

This study investigates the performance of ML, DL and transformers in predicting the sentiment of social media tweets about women in STEM. We also study the impact of two important women's celebration days (International Day of Women and Girls in Science, and International Women's Day) on public opinion on Twitter platforms. To the best of our knowledge, investigating the public views pertaining directly to women in STEM on social media platforms using NLP and AI-related techniques has not been studied in the literature. This study is based on the work submitted in [31].

II. Related Work

Sentiment analysis uses NLP techniques to analyse a person's opinion and emotion, and it is also known as opinion mining [13]. In recent years, researchers have shown an increased interest in conducting sentiment analysis on social issues regarding women on social media networks. For example, [15] conducted a sentimental analysis of people's opinions on social issues of women on Twitter. They extracted a total of 150,000 tweets using relevant keywords such as #Women and #MeToo. Tweets were processed and labelled as positive, negative or neutral using open language models. They applied several ML models such as Logistic Regression, Random Forest, Naïve Bayes and Support Vector Machine classifiers and achieved accuracy scores between 80% to 95%. [10] has developed a dataset of sexist expressions and attitudes on Twitter in Spanish (MeTwo) and applied ML and DL techniques to detect sexist behaviours on social media platforms. They detected clear signs of sexism on social networks, that is discrimination based on sex or gender, especially against women and girls. [16] used an ensemble of classifiers, such as logistic regression and tree ensemble models for detecting hate speech against women in English language tweets. They built a binary classification model to detect misogynous from non-misogynous tweets.

Sentiment analysis for women in STEM field has not been studied widely in the literature. Only a few researchers have studied this problem in a non-social media context. For example, [20] performed sentiment analysis on STEM-related data obtained from the comments of 450 YouTube videos. They have collected 23,005 comments from these videos, up to 100 comments per video. However, this number decreases significantly when the scope is turned to women in STEM. The article found that female-hosted channels accumulated more comments per view when compared with male channels, but their comments were hostile, negative and sexist. The research used Univariate Analysis of Variance (ANOVA) to conduct its sentiment analysis. This statistical method compares the mean value of two or more groups and decides whether it is different. However, this paper doesn't explore the polarity or sentiment scores of the studied data.

III. Methodology

The overview of our proposed framework is shown in Fig. 1, with each component of the framework explained in the following.

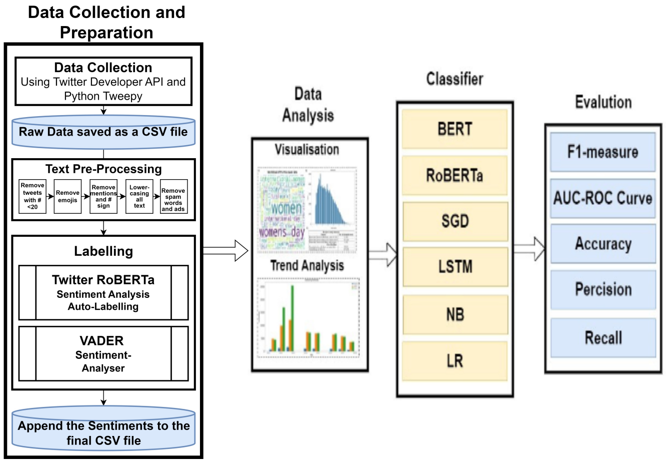

Fig. 1. Overview of the proposed framework

A. Data Collection and analysis

We use the Twitter Academic API to obtain a large sample of tweets related to women in STEM covering the period from January 2022 to August 2022. The data set (tweets) has been collected using Tweepy, an official Python Twitter API library. Based on our initial analysis, Twitter carries an array of different hashtag phrases related to women in STEM field, as well as phrases related to the individual STEM components (i.e., Science, Technology, Engineering and Mathematics). To maximize the size of the extracted data, we collected relevant tweets for each of these components individually and in combination with each other. The hashtags/keywords used for our data collection are shown in Table I. We collected a total of 254,281 tweets. The data set was stored as a CSV (comma-separated values) file using the Pandas library (python library for data analysis).

B. Data pre-processing

Text pre-processing is a method to clean text data and make it ready to feed data to the model. In this study, we applied a range of data pre-processing techniques to achieve this objective.

- Noisy data were removed from Tweets by removing URLs, @users, punctuation and numbers

TABLE I
Hashtags/keywords used as search parameters during data extraction phase

Hashtags	Associated Hashtag
#womenInSTEM	#girlsInSTEM
#womenInScience	#girlsInScience
#womenInTech	#girlsInTech
#womenWhoCode	#girlsWhoCode
#womenInEngineering	#girlsInEnginnering
#womenInMaths	#girlsInMaths

- Remove tweets that contained more than 20 hashtags from the dataset. Twitter guidelines recommend that using a maximum of 2 hashtags per tweet is more effective for sentiment analysis.
- Tokenisation was applied, breaking a given tweet (sentence) into a collection of words (tokens).
- Following the pre-processing step, the tweets were grouped by their word count. There was a total of 4,843 tweets with three words or less. These tweets were removed because they hold little to no information.
- Retweets were filtered to prevent duplicates.
- To safeguard the users' privacy, we removed all data from personally identifiable information and any meta-data about users.

Overall, the described pre-processing procedures resulted in removing 137,172 tweets from 254,281, leaving a pre-processed dataset of 141,000 tweets. Statistics of the highly frequently used keywords and wordcloud are shown in Figs. 2 and 3. We found that people are talking mainly about events that celebrate women's success, such as women's international day and women in science.

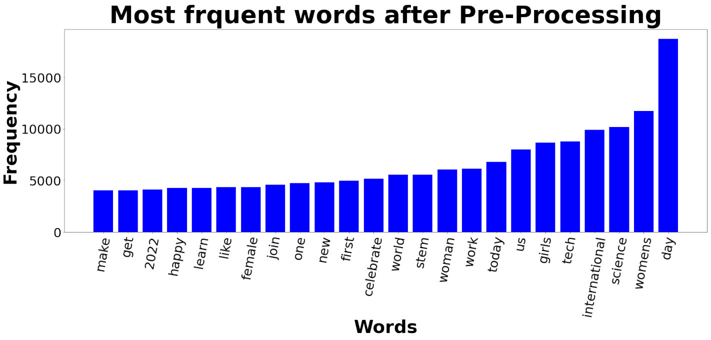

Fig. 2. Most frequent words in the pre-processed dataset

C. Labeling the sentiment data set

The labelling process is essential in any supervised machine-learning task for sentiment analysis. According to different literature in the field (e.g. [13]), the sentiments of a text can be mainly classified into three groups: positive, negative, and neutral. However, the manual annotation of sentiments is extremely difficult due to the large size of the collected tweets. Owing to this reason, we adopt two sentiment

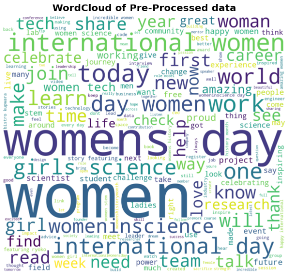

Fig. 3. Wordcloud of the pre-processed dataset

annotation methods that researchers have recently used for similar tasks.

1. VADER labelling method We use the Valence Aware Dictionary (VADER) sentiment analyzer to calculate the intensity of the sentiment as either positive, neutral, or negative [23]. VADER combines a dictionary, which maps lexical features to emotional intensity. Each word in the text is given a sentiment score measured on a scale from -4 to +4, where -4 is the most negative, 0 is neutral, and +4 is the most positive. Then, a compound score C is estimated by computing the sum of the valence scores of each word, where it will be normalised to be between -1 (negative) and +1 (positive). To extract the tweet's final sentiment, VADER uses the compound score C to classify the sentences as positive, neutral or negative as follows:

$$S_{Ti} = \begin{cases} Negative, & C \leq -0.05. \\ Positive, & C \geq 0.05. \\ Neutral, & -0.0 < C < 0.05. \end{cases} \tag{1}$$

Where S_{Ti} denotes a sentiment label assigned to a data record (tweet) T_i. VADER polarity/sentiment results for our dataset are shown in Fig. 4. The figure displays the average compound, positive, neutral and negative scores generated for each sentiment class.

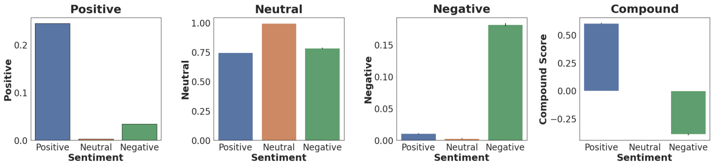

Fig. 4. Average VADER scores across the three sentiment classes

2. RoBERTa Based Classifier After BERT's revolutionary progress, a group of researchers at Facebook modified the BERT model by doubling the size of its hidden layers. They have re-train the BERT model using 160GB of text, and modified key hyperparameters allowing it to process out-of-vocabulary words better; this model was called RoBERTa (A Robustly Optimized BERT Pretraining Approach), and it was built on top of Pytorch [24].

Each data record (tweet) was assigned the highest compound score sentiment label. Table II shows RoBERTa polarity/sentiment results for example of tweets from our dataset.

It is noticed that the labelling results generated by RoBERTa were slightly different from the ones generated by VADER. In particular, RoBERTa has classified more tweets (15%) as neutral, and fewer tweets as positive and negative sentiment when compared to VADER (see Fig 5).

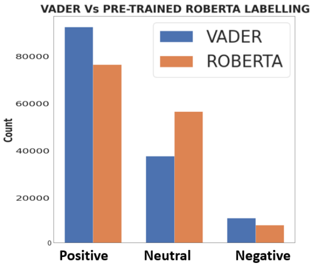

Fig. 5. Sentiments labels generated by VADER (blue) and RoBERTa (orange)

It is noticeable that VADER labelling is limited to a small range of values, whereas in RoBERTa, the scores are not limited to a range and are more generalised. This is because, Unlike the VADER model, RoBERTa's model can extract context. Owing to these reasons and based on our experimental observations, we used the dataset labelled by the RoBERTa method to train and test a series of machine learning models.

D. Sentiment Timeline Analysis

The average number of the extracted posts relevant to women/girls in STEM was 30,000 tweets per month (from Jan 2022 to August 2022), except for February and March, where there was a massive increase of around 50% and 100%, respectively. Figure9 shows the sentiment timeline (generated by the RoBERTa labelling method), which helps to understand trends in positive, neutral and negative sentiment over time of the extracted tweets. It was observed that people expressed their positive sentiments more often in February and March compared to the other months. In addition, the positive and neutral sentiments were almost equal in January, April, May, June, July and August. It was found that February and March correspond to two annual international days celebrating women. The first event was the International Day of Women and Girls in Science, which is an annual celebration (on February 11) that aims to achieve gender equality and empower women and girls in science [21]. The first event was International Women's Day, an annual celebration (on March 8) that supports women's rights and brings attention to issues of gender equality in various domains, including the STEM domain [21]. These initiatives aim to encourage girls to participate more in STEM subjects and empower women to pursue careers in STEM fields. Recent statistics show that these campaigns have contributed to a 0.9% increase in the percentage of women in STEM from 2019 to 2020 [2].

IV. Sentiment Classification Models

Here, we aim to design and train a supervised classification model that can automatically detect the sentiment of tweets towards women in STEM. The classification model will be

TABLE II
SENTIMENT ANALYSIS SCORES USING ROBERTA

Sample Text	Negative	Neutral	Positive	NegativeSentiment
Women who work in STEM fields disgust me; I wish they had disappeared	0.93	0.06	0.01	Negative
I work as an engineer, but it is a very exhausting job for me as a woman	0.94	0.052	0.004	Negative
Girls love to work in STEM roles	0.0	0.02	0.98	Positive
happy new month our dearest queens yay feb is finally here and yes we are excited	0.001	0.005	0.99	Positive
this image was shot on the campus at the Stanford school of engineering	0.026	0.95	0.025	Neutral
The woman is a scientist.	0.01	0.58	0.41	Neutral

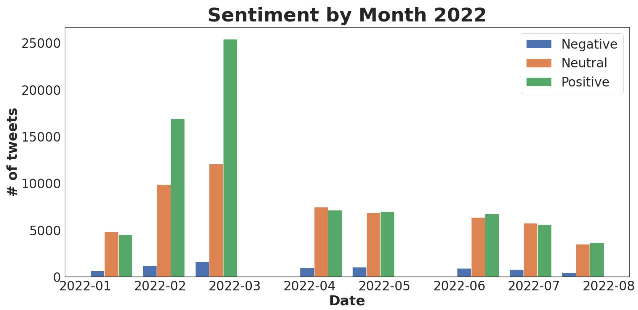

Fig. 6. Sentiment Timeline Analysis generated by RoBERTa method including two international days celebrating women (International Day of Women And Girls in Science)

applied to a corpus of 141,000 preprocessed tweets. True sentiment labels were assigned by the RoBERTa model.

To provide a comprehensive analysis, we study the performance of multiple supervised classification systems for detecting sentiment in women in STEM. This includes ML classifiers, DL models, and transfer learning (TL) approaches. We finally conduct a comparative analysis of the results obtained from all methods. The following subsections describe the classification task in detail.

A. Feature extraction

Feature extraction involves the transformation of raw text data into numerical features. After completing the data preprocessing phase, we extracted a set of relevant numerical features from textual tweets. This was performed using CountVectorizer (CV) and Term Frequency and Inverse Documents Frequency (TF-IDF) methods. Both functions are built on top of the sklearn Python library.

Firstly, CV was used to turn the text into vectors depending on the word frequency resulting in a matrix that contains a unique representation of each word as its columns, and each row will have the count of the words in each tweet. Then the matrix will be passed into TF-IDF. The term frequency is set to the number of times a word is prevalent in a given block of text. The inverse document frequency is applied to a specific word, calculated by dividing the number of documents by the number of documents that contain a specific word. TF-IDF is an efficient information retrieval technique representing a specific word or phrase's importance to a given document. Hence, the TF-IDF does not only focus on the frequency of words present but also the significance of those words in the context of the entire data set.

B. Classification performance measures

Our classification results were evaluated using the following performance measures:

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Recall = TP/(TP +FN)
- Precision = TP/(TP + FP)
- F1-score = 2*(Recall * Precision)/(Recall + Precision)

The above performance measures are derived from the counts of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). Note that, if an instance is positive and it is classified as positive, it is defined as TP. If the instance is negative and it is classified as positive, it is FP. While a negative instance classified as negative is TN and if it is classified as positive, it is called FN.

C. Machine Learning Models

In this study, we evaluate the performance of three commonly used classifiers, and the results will be used as a baseline for the rest of the models. First, we studied the performance of the Naive Bayes Classifier (NB) algorithm [27]. It is a probabilistic classifier that applies the Bayes theorem with strong independence assumptions between the features. The Stochastic gradient descent classifier (SGD) was also investigated here [28]. It is an iterative method that computes the gradient descent using a single sample.

The third model applied here was Logistic Regression [29] which is a linear statistical model for discriminant analysis.

D. Deep Learning Models

In this study, we evaluate the performance of three DL models, namely Long-Short Term Memory (LSTM), Bidirectional-LSTM (Bi-LSTM), and Hybrid Bi-LSTM [19]. LSTM is an RNN architecture that has been used widely in the literature to solve NLP-related problems. This is due to their impressive ability to understand language modelling and perform word prediction. They are mainly used for sequential data as they can remember their input, due to internal memory. However, LSTM and Bi-LSTM differ from RNNs in their ability to recognize the relationship between values from the beginning to the end of a sequence by maintaining long-range connections. Bi-LSTMs are an extension of LSTMs. Unlike LSTM, which only receives inputs from one direction, the Bi-LSTM uses another layer of LSTM that receives the reversed input to understand the full context of the sentence.

The two above LSTM-based models were trained to identify the sentiment of the women in STEM tweets. The TensorFlow sequential model has been used here to configure the LSTM models. The first layer was assigned as the word's embeddings. Next, a 1-D dropout layer of 0.5 was used for regularization. Then, the LSTM layer was added with fixed values set on the dropout and recurrent dropout options. The LSTM cell's output is connected to dense layers with ReLU as an activation function, followed by a softmax activation function configured to output three values. The model was compiled using the Adam optimizer and the categorical cross-entropy loss function with accuracy metrics. Similarly, Bi-LSTM used the same architecture as the standard LSTM, except for changing the standard LSTM layer to Bi-LSTM. Finally, a hybrid model that contains a layer of Bi-LSTM and another layer of LSTM was also trained and tested to determine the effectiveness of an additional LSTM layer.

E. Transformers Models

In this study, we adopt the transfer learning (TL) techniques, for classifying public opinions on women STEM-related tweets. TL is an ML technique that applies prevailing knowledge to problems in other domains and presents advanced prediction results [18].

BERT (Bidirectional Encoder Representations from Transformers) is an open-source NLP pre-trained (transformer) model developed by the Google AI Language team in 2018 [17], [22]. It was trained with large language datasets, such as the Wikipedia Corpus and Common Crawl, and can be fine-tuned for a specific task. It is a contextual and bidirectional language model. In contextual language representation, the unstructured date is converted into a structured one by taking into account the relationship between words. In BERT models, words are predicted based on both the before and after contexts of the words surrounding it. In this study, we fine-tune the BERT-base model, which uses 12 transformer encoders, to build a Sentiment classifier that can detect the sentiment of women in STEM.

We also fine-tuned another model called BERTweet [30], which is based on RoBERTa architecture and relies on a masked language modelling objective. The model has been pre-trained on 80GB of streamed tweets from January 2012 to March 2020, equivalent to 850 million tweets.

The third model used here was TimeLMs [26], which uses a recurrent learning strategy that enhances Twitter-based language models' capacity to deal with future and out-of-distribution tweets. The model was pre-trained on 124M tweets from 2018 to 2021. The model was benchmarked using TweetEval, which has outperformed several models in most tests, such as RoBERTa and BERTweet.

For all the above transformers, the architecture was as follows: The model included three layers with the first layer containing one of the above pre-trained models, defined initially using the transformer library. The second layer was the classifier which will reduce the outputs to three. The final layer was BCEWithLogitsLoss, which combines a BCEloss function with a sigmoid function in one layer. Finally, the Adam optimizer was used to allow the model to converge more efficiently. The Fig. 7 illustrates the architecture that was used with all the transformers models.

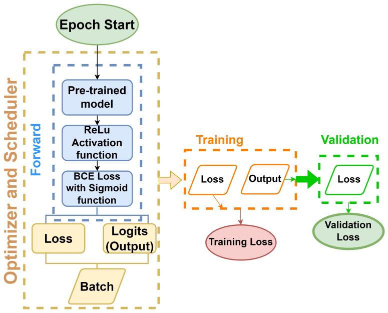

Fig. 7. Architecture of transformers models

V. Experiments and Results

This section presents the experimental set-up results of the ML, DL and TL methods, discussed above, using the extracted tweets. Dataset has been randomly split into three subsets distributed as: 75% training, 22.5% testing and 2.5% validation. In all experiments, (hyper-)parameters of all classification algorithms were tuned via 5-fold cross-validation on the validation set. For fair comparisons, each algorithm was run five times, and we report the average results over the five runs. All experiments were run using Python and Jupyter hosted on Google's Colab platform. The code for all methods has been made available on this Colab notebook

To retrain the BERT, BERTweet, and TimeLMS models, we used the Pytorch Lightning framework. The models were configured as follows: batch size was 128, the number of epochs was set to 25 (with an early stopping callback function that can automatically stop the training process by checking the change in validation loss), and AdamW optimizer was used with a 20% warmup percentage and weight decay of 0.01. The learning rate was configured using a cosine scheduler that peaked at 1e-5. Model Checkpointing callback function was added, which saves the best-trained version of the transformer models depending on the validation loss despite the number of epochs. Finally, the models were trained using Google Colab cloud computing eight cores TPU [32].

Table III illustrates the results generated by all the studied models using the collated data. It can be observed that the results obtained from the TL models are generally better than the results obtained from DL and ML models. In particular, the results indicate that TimeLMs yield the best performance across all metrics with an accuracy of 96% closely followed by BERTweet with an accuracy of 92%, and BERT with an accuracy of 88%. It can also be observed that the results obtained from the DL are generally outperforming the ML models, with Bi-LSTM being the best-performing DL model with an accuracy of 85%.

The results obtained by the DL and TL algorithms have been selected for further investigation as they outperform the ML algorithms. The accuracy obtained for the positive, neutral and negative sentiments is presented in receiver operating

TABLE III
SENTIMENT ANALYSIS CLASSIFICATION RESULTS FOR THE MACHINE LEARNING, DEEP LEARNING AND TRANSFER LEARNING METHODS

Model	Accuracy	Precision	Recall	F1-score
Machine Learning				
Naive Bayes	68%	81%	44%	43%
Logistic Regression	82%	75%	66%	68%
Stochastic Gradient Descent	79%	71%	68%	69%
Deep Learning				
LSTM	83%	77%	71%	73%
Bi-LSTM	85%	76%	75%	75%
Hybrid Bi-LSTM	82%	75%	73%	74%
Transfer Learning				
BERT	88%	85%	80%	82%
BERTweet	92%	90%	88%	89%
TimeLMs	96%	96%	95%	95%

characteristic (ROC) graphs, which is a probability curve that maps the true positive rate (TPR) against the false positive rate (FPR) at multiple threshold values for each sentiment. When the skewness of the curve is high to a genuine positive score, then the classifier is considered to have higher efficiency. The RoC curve acts as a reliable performance measure in the ML literature. Fig. 8. shows a comparative analysis of ROC curves between LSTM, BiLSTM, Hybrid Bi-LSTM, BERT, BERTweet, and TimeLMs models. It is noticed that the curves generated by the TL-based models (two bottom images) are closer to the TPR, indicating better performance than the DL-based models (two top images).

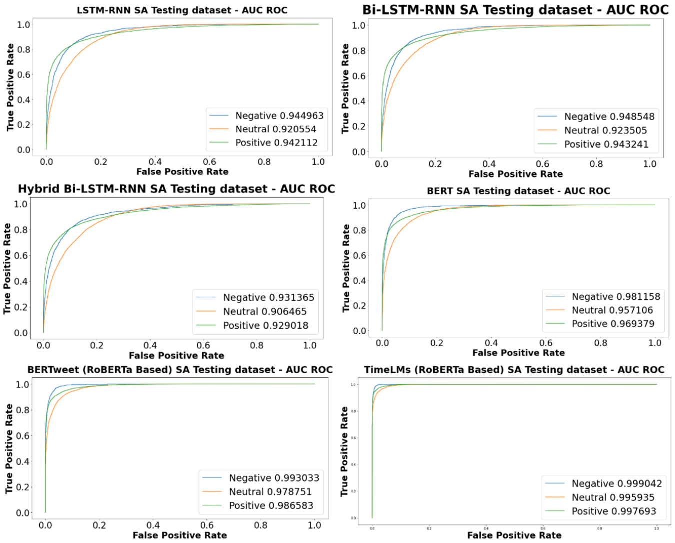

Fig. 8. Receiver operating characteristic (ROC) graphs generated by DL and TL models

A. *Investigate the impact of women's celebration days on sentiment results*

As mentioned in section III-D, the number of extracted tweets relevant to women in STEM was larger during February and March compared to the rest of the studied months. This is because these two months include two international days celebrating women (International Day of Women and Girls in Science [21]).

To investigate this further, we extracted all the tweets posted during these two days and performed some statistical analysis. Results are displayed in see Fig. 9, and they reveal that the International Day of Women And Girls in Science had 10,730 tweets, with 76.8% of these tweets being positive, 21% neutral, and 2.1% negative. Furthermore, International Women's Day had a total of 13,609 tweets with 82.3% positive, 14.8% neutral, and 2.9% negative.

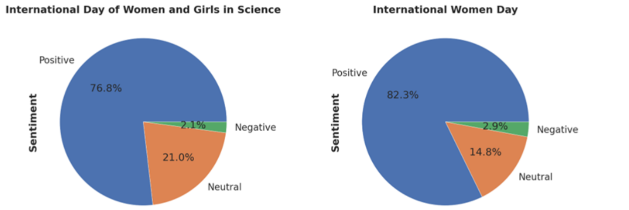

Fig. 9. The pie chart to the right illustrates the overall sentiment statistics for the International Day of Women And Girls in Science (2022-02-11), and on the right, the overall sentiment statistics for International Women's Day (2022-03-08).

In order to study the impact of these two annual celebration days on the generated sentiment, we have excluded all the tweets posted during these two days specifically from our dataset. Then, we applied RoBERTa again to generate the tweets' sentiments (true labels). As seen in Fig. 10, the ratio of positive tweets has reduced from 64.9% to 55.7% in March and from 60.5% to 50.3% in Feb. On the other hand, the total number of negative tweets has increased from 4.1% to 6.3% in March and from 4.2% to 5.6% in Feb. it is worthwhile mentioning that the number of positive sentiments in February and March are still higher when compared to the other studied months of the year. This is because people tend to post positive tweets about women in general, and specifically about women in science (STEM), within the months of February and March due to the celebration days. These findings illustrate the impact of the annual celebration days, which commemorates women's achievements, on the public's opinion of women in STEM.

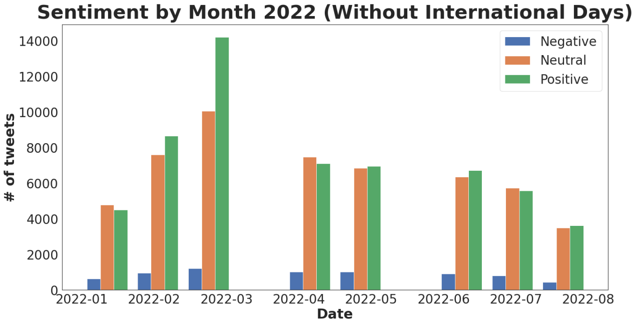

Fig. 10. Sentiment Timeline Analysis generated by RoBERTa method after excluding tweets posted during the two annual celebration days (International Day of Women And Girls in Science, and International Women's Day)

Much work is yet to be done to bridge the gender disparity prevalent in the STEM workforce making it an ongoing endeavour. However, the results obtained from this study suggest a promising future in achieving this goal with the opinions of the public being of a positive nature, thus aligning with the women in STEM goal of encouraging and empowering more females to pursue careers in STEM.

VI. Conclusion

This paper proposes a novel idea using advanced sentiment analysis classification methods to investigate views related to women in STEM on Twitter. We scraped and pre-processed more than 250k tweets relating to women in STEM, and conducted an exploratory analysis. We utilized several state-of-the-art NLP tools for pre-processing and labelling the textual data. Then, we applied a series of machine learning and deep learning methods to perform the sentiment classification using the collected dataset. Furthermore, variants of BERT transformer models were used, including BERTweet, and TimeLMs to achieve the best performance. Our results revealed that TimeLMs outperformed all the studied algorithms and achieved 96% accuracy in sentiment detection.

This study revealed that people's attitude in response to women in STEM is generally positive on the Twitter platform, with few people spreading negative posts (55% positive, 40% neutral and 5% negative). This result suggests that most people encourage women to participate in the STEM fields. However, it was observed that positive communication increases dramatically during the months that celebrate women's achievements (International Day of Women And Girls in Science, and International Women's Day). Positive communications can potentially influence women's aspirations to enrol in STEM majors in higher education, leading to more women pursuing careers in STEM fields, and hence filling the employment gap in the STEM market. Therefore, promoting such events, which celebrate women's success, will contribute effectively to addressing the gender inequality problem in the STEM domain.

References

[1] Rothwell, J., 2013. The hidden STEM economy. Washington, DC: Metropolitan Policy Program at Brookings.

[2] WISE. (2020). [online] Available at: https://www.wisecampaign.org.uk/

[3] Industry profiles and skills needs. (2016). [online] https://www.stem.org.uk/sites/default/files/pages/downloads/Industry-profiles-and-skills-needs.pdf

[4] Bank Of England (2022). What is GDP? [online] https://www.bankofengland.co.uk/knowledgebank/what-is-gdp

[5] Hrabowski III, F.A., 2012. Broadening participation in the American STEM workforce. BioScience, 62(4), pp.325-326.

[6] Makarova, E., Aeschlimann, B. and Herzog, W., 2019, July. The gender gap in STEM fields: The impact of the gender stereotype of math and science on secondary students' career aspirations. In Frontiers in Education (Vol. 4, p. 60). Frontiers Media SA.

[7] Kahn, S. and Ginther, D., 2017. Women and STEM (No. w23525). National Bureau of Economic Research.

[8] Smith, E., 2011. Women into science and engineering? Gendered participation in higher education STEM subjects. British Educational Research Journal, 37(6), pp.993-1014.

[9] American Association of University Women (AAUW). (2022). The STEM Gap: Women and Girls in Science, Technology, Engineering and Mathematics. [online] https://www.aauw.org/resources/research/the-stem-gap/

[10] Rodriguez-Sanchez, F., Carrillo-de-Albornoz, J. and Plaza, L. (2020). Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. IEEE Access, 8, pp.219563–219576.

[11] Burbach, L., Halbach, P., Ziefle, M. and Calero Valdez, A., 2020. Opinion formation on the internet: The influence of personality, network structure, and content on sharing messages online. Frontiers in Artificial Intelligence, 3, p.45.

[12] Lenhart, A., 2009. Adults and social network websites.

[13] Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[14] Prabha, M.I. and Srikanth, G.U., 2019, April. Survey of sentiment analysis using deep learning techniques. In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) (pp. 1-9). IEEE.

[15] Kaur, C. and Sharma, A., 2020, October. Social issues sentiment analysis using python. In 2020 5th International Conference on Computing, Communication and Security (ICCCS) (pp. 1-6). IEEE.

[16] Ahluwalia, R., Soni, H., Callow, E., Nascimento, A. and De Cock, M., 2018. Detecting hate speech against women in english tweets. EVALITA Evaluation of NLP and Speech Tools for Italian, 12, p.194.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[18] Palicki, S.K., Fouad, S., Adedoyin-Olowe, M. and Abdallah, Z.S., 2021, March. Transfer learning approach for detecting psychological distress in brexit tweets. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 967-975).

[19] Yu, Y., Si, X., Hu, C. and Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), pp.1235-1270.

[20] Amarasekara, I. and Grant, W.J., 2019. Exploring the YouTube science communication gender gap: A sentiment analysis. Public Understanding of Science, 28(1), pp.68-84.

[21] National Day Calendar, 2022. [online] https://nationaldaycalendar.com/register-a-national-day/

[22] Alaparthi, S. and Mishra, M., 2020. Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127.

[23] Hutto, C. and Gilbert, E., 2014, May. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[24] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[25] Ghasiya, P. and Okamura, K., 2021. Investigating COVID-19 news across four nations: a topic modeling and sentiment analysis approach. Ieee Access, 9, pp.36645-36656.

[26] Loureiro, D., Barbieri, F., Neves, L., Anke, L.E. and Camacho-Collados, J., 2022. Timelms: Diachronic language models from twitter. arXiv preprint arXiv:2202.03829.

[27] Berrar, D., 2018. Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 403.

[28] Bottou, L., 2012. Stochastic gradient descent tricks. In Neural networks: Tricks of the trade (pp. 421-436). Springer, Berlin, Heidelberg.

[29] Kirasich, K., Smith, T. and Sadler, B., 2018. Random forest vs logistic regression: binary classification for heterogeneous datasets. SMU Data Science Review, 1(3), p.9.

[30] Nguyen, D.Q., Vu, T. and Nguyen, A.T., 2020. BERTweet: A pre-trained language model for English Tweets. arXiv preprint arXiv:2005.10200.

[31] Alkooheji, E., Fouad S., 2022, Explainable Sentiment Analysis for women in STEM using Twitter. MSc Dissertation in Artificial Intelligence with Business Strategy, Engineering and Applied Science, Aston University, UK

[32] Akkalyoncu Yilmaz, Z., Clarke, C.L. and Lin, J., 2020, July. A lightweight environment for learning experimental IR research practices. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2113-2116).