Binary Battle: Leveraging Machine Learning and Transfer Learning Models to Distinguish between Conspiracy Theories and Critical Thinking

Sidharth Mahesh[†], Sonith Divakaran[†], Kavya Girish[†] and Hosahalli Lakshmaiah Shashirekha[*,†]

Department of Computer Science, Mangalore University, Mangalore, Karnataka, India

Abstract

In the context of automatic content moderation Natural Language Processing (NLP) has a complex task when it comes to distinguishing between conspiracy theories and critical thinking. While conspiracy theories present complex narratives attributing significant events to covert actions by powerful and malicious entities, critical thinking involves scrutinizing decisions without resorting to any sinister explanations. Making this distinction is essential to avoid the mislabeling of valid criticism as conspiracy, which may unintentionally lead people to join conspiracy communities. Conspiratorial and critical narratives are both examples of oppositional thinking, which is important in public debate, particularly in controversial areas like public health. In this direction, "Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives"- a shared task organized at PAN 2024, invites the research community to address the challenges of distinguishing between conspiracy and critical texts in English and Spanish languages. To explore the strategies for distinguishing between critical and conspiracy texts in English and Spanish on social media platforms, in this paper, we - team MUCS, describe the models proposed for Subtask-1: "Distinguishing between critical and conspiracy texts" of the shared task. We explored machine learning models trained with Term Frequency-Inverse Document Frequency (TF-IDF) of char n-grams in the range (1, 5) and transfer learning techniques using several BERT variants fine-tuned with the given English and Spanish datasets, to classify the given unlabeled English and Spanish text into one of the two categories - 'CONSPIRACY' or 'CRITICAL'. Among the proposed models, English_BERT and Spanish_BERT models obtained Matthews Correlation Coefficient (MCC) scores of 0.7162 and 0.6293 for English and Spanish languages respectively.

Keywords

Oppositional thinking analysis: Conspiracy vs Critical Narratives, Oppositional Thinking, Conspiracy Theories, Machine Learning, Transfer Learning

CEUR-WS.org/Vol-3740/paper-266.pdf

1. Introduction

Conspiracy theories and critical thinking are two forms of oppositional thinking which are common especially on contentious issues and analyzing oppositional thinking entails scrutinizing narratives that question mainstream perspectives. Further, understanding the impact of different types of oppositional thinking on public opinion and behavior is crucial [1]. While critical thinking fosters constructive and democratic debate characterized by reasoned questioning without unfounded explanations, conspiratorial thinking can lead to misinformation and social conflict attributing the significant events to hidden malevolent forces[2]. In social and political discourse, conspiracy theories can have detrimental effects on an individual as well as on organisations or the entire society. These theories suggest that major social (encompass gatherings and activities involving people) and political events (activities related to government and leadership) with careful planning by powerful and malevolent entities can spread false information and stir social unrest. Conspiracy theories have been associated with violence, war, terrorism, prejudice, poor health choices, and denial of climate change [3]. In contrast, critical

CLEF 2024 – Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France

[*] Corresponding author.

[†] All authors contributed equally.

sidharthmaheshedu@gmail.com (S. Mahesh); sonithksd@gmail.com (S. Divakaran); kavyamujk@gmail.com (K. Girish); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

thinking involves analyzing and questioning decisions, particularly in areas like public health, without attributing events to hidden conspiracies.

Table 1
Sample text and their corresponding labels in English dataset

Sample Text	Category
2021: They wanted to know your vaccination status and see your papers to be allowed to go to restaurants. 2023: They don't want you to know the vaccination status of someone who died suddenly.	CRITICAL
"Victoria Premier Dan Andrews threatening construction workers that they have a week to prove they 've been injected with the experimental covid vaccine or he 'll close down their entire industry! ",	
Living in Europe after grad school and working in corporate finance for adidas Europe when the Euro went into effect, when NO EUROPEANS wanted it, was enough to cause me to start digging. Then you find the NWO, then that leads to the luciferians, and then it all ultimately leads to the children. It's always about the children. You can never unknow, unsee, unheard, unlearn. # GodWins	CONSPIRACY
If you have n't yet, I would HIGHLY recommend for anyone unfamiliar to read ALL of the Q intelligence drops, especially the earlier ones posted while still on 4Chan … You can start here: https :// qalerts. app You will see just how EPICALLY it 's all falling into place right now, and you will be able to appreciate the DESTRUCTION of the Old Guard NWO in a whole different light! Despite the fact that this world belongs to Satan, God ALWAYS wins.	

Mislabeling critical discourse as conspiratorial can suppress healthy debate and alienate individuals who are merely questioning decisions and mislabeling critical conversation as conspiracy can stifle constructive disagreement and alienate people who are only challenging judgements [3, 4]. This leads to a climate of mistrust and inhibits the exchange of open and sincere ideas. On the other hand, failing to identify and address conspiratorial narratives allows the spread of misinformation, leading to societal division and mistrust. Thus, accurate differentiation between the two forms of oppositional thinking by content moderation systems is essential to avoid marginalizing legitimate criticism and preventing individuals from being drawn into conspiracy communities by mistaking valid critique for conspiracy theories [5]. Additionally, distinguishing between these forms of oppositional thinking is essential for public discourse and social harmony.

Distinguishing between the oppositional forms of thinking is challenging due to the nuanced and overlapping content, the context-dependent nature of oppositional statements, and oppositional attitudes [6]. Further, differentiating between well-founded critical and unfounded conspiracy theories require advanced and context-aware NLP techniques and developing such techniques is crucial for improving the accuracy and fairness of content moderation systems. To address the challenges of distinguishing between the oppositional forms of thinking, "Oppositional thinking analysis: Conspiracy theories vs Critical thinking narratives" shared task organized at PAN 2024 [7], invites the research community to develop models to distinguish between conspiracy and critical thinking texts in English and Spanish languages. The shared task has two subtasks in English and Spanish languages and we - team MUCS participated in only Subtask-1: 'Distinguishing between critical and conspiracy texts'. This subtask having two categories - 'CONSPIRACY' and 'CRITICAL', is modeled as a binary classification problem. In this paper, we describe various machine learning models trained with TF-IDF of char n-grams in the range (1, 5) and transfer learning techniques using several BERT variants fine-tuned with the given Engish and Spanish datasets, to classify the given unlabeled English and Spanish text into one of the two categories - 'CONSPIRACY' or 'CRITICAL'. The sample text from the given datasets for English and Spanish are shown in the Tables 1 and 2 respectively.

The rest of the paper is organized as follows: Section 2 describes the recent literature on the two forms of oppositional thinking and Section 3 focuses on the description of the proposed models followed by the experiments and results in Section 4. The paper concludes with future works in Section 5.

2. Related Work

Conspiracy theories involve elaborate, unverified claims driven by cognitive biases and emotional needs, often rejecting official explanations without sufficient evidence. In contrast, critical thinking is characterized by objective analysis, logical reasoning, and the evaluation of evidence from multiple sources. Research highlights that while conspiracy beliefs are linked to cognitive biases and feelings of

Table 2
Sample text with the English translation and corresponding labels in Spanish dataset

Sample Text	Translated Text	Category
Siento ya tdas las vacunas vienen contaminadas mi sobrina hace un mes le pusieron la vacuna de la hepatitis a. El fin de semana apareció erupciones. A mi hermano le dijeron es varicela	I feel like all the vaccines are contaminated, my niece was given the hepatitis A vaccine a month ago. Rashes appeared over the weekend. They told my brother it's chickenpox	CRITICAL
Buenas , he buscado en la lupa y no he encontrado nada relativo a espondilitis . Algún vacunado en el grupo o que conozca a alguien que la padezca ? Estoy muy interesada en saber de efectos adversos o si los especialistas recomendaron su inoculación . Sobre psoriasis sí he vistazo varios brotes en el grupo . Gracias	Hello, I have searched through the magnifying glass and I have not found anything related to spondylitis. Anyone vaccinated in the group or who knows someone who suffers from it? I am very interested in knowing about adverse effects or if specialists recommended inoculation. Regarding psoriasis, I have seen several outbreaks in the group. Thank you	
Dieciochoañero nos expresa amablemente su vivencia de la # Plandemia , sus razones para vacunarse sin ningún tipo de información y su desconocimiento acerca de la # censura y de la existencia de una disidencia culta y respetable . # Despierta # Respiracionistas # NWO # YoNoMeVacuno # LosNiñosNoSeTocan # Libertad # AMíNoMePinchais	Eighteen-year-old kindly expresses to us his experience of the # Pandemic, his reasons for getting vaccinated without any information and his ignorance about # censorship and the existence of a cultured and respectable dissidence. # Wake up # Breathists # NWO # I don't get vaccinated# ChildrenDon'tTouch # Freedom # Don'tPinchMe	CONSPIRACY
Por suerte vivo en un pueblo alejado de 110 habitantes, rodeada de bosques y campos (la mejor medicina) creo que soy la única no vacunada , la antena està a 4kms , i creo tener mi casa y nuestros cuerpos bien protegidos	Luckily I live in a remote town with 110 inhabitants, surrounded by forests and fields (the best medicine) I think I am the only one not vaccinated, the antenna is 4km away, and I think I have my house and our bodies well protected	

powerlessness, critical thinking fosters informed decision-making and intellectual humility [3].

To explore different strategies for the conspiracy textual content classification in social media, Moosleitner and Murauer [8] employed machine learning models (Support Vector Machines (SVM), Multinomial Naive Bayes (MNB), and Extremely randomized Trees) trained with TF-IDF of character, word and Document Term n-grams based features and BERT models (BERT-base, RoBERTa, and DistilBERT) for English text. For the three tasks: Task 1 - Text-Based Misinformation Detection, Task 2 - Text-Based Conspiracy Theories Recognition, and Task 3 - Text-Based Combined Misinformation and Conspiracies Detection, their proposed BERT-base models outperformed all other models in all three tasks obtaining MCC scores of 0.3184, 0.3624, 0.3347 for Tasks 1, 2, and 3 respectively. For detecting fake news during covid-19 pandemic, Tahat et al. [9] proposed a hybrid analysis using Structural Equation Modelling (SEM) and machine learning classification algorithms such as BayesNet, AdaBoostM1, LWL, Logistic, J48, and OneR, for English dataset. Among the proposed models J48 classifier outperformed other machine learning classifiers with an F-Measure score of 0.863. Peskine et al. [10] proposed transformer model (composed of an ensembling of CT-BERT models) and a node embedding-based techniques (node2vec + Multilayer Perceptron (MLP) classification head) to detect COVID-19-related conspiracy theories in tweets in English language which consists of three subtasks: Task 1 - Text-Based Misinformation and Conspiracies Detection, Task 2 - Graph-Based Conspiracy Source Detection, and Task 3 - Graph and Text-Based Conspiracy Detection. Their proposed CT-BERT ensembling model obtained a MCC score of 0.710 and 0.719 for Task 1 and Task 3 respectively, and node2vec + MLP model obtained a MCC of 0.355 for Task 2.

Giachanou et al. [2] performed a comparative analysis of various profiles and psychological and linguistic characteristics in social media texts of users who share posts about conspiracy theories. The authors then compared the effectiveness of these characteristics for predicting whether a user is a conspiracy propagator or not by proposing ConspiDetector, a model that is based on a Convolutional Neural Network (CNN) which combines word embeddings with psycho-linguistic characteristics extracted from the tweets of users to detect conspiracy propagators. Recordare et al. [11] implemented various machine learning classifiers (Logistic Regression (LR), k-Nearest Neighbours (kNN), Naive Bayes (NB), SVM, Decision Trees (DT), Random Forest (RF), Gradient Boosting: XGBoost and LightGBM, Quadratic Discriminant Analysis, MLP, Ridge Classifier, and Linear Discriminant Analysis, trained with Bidirectional Auto-Regressive Transformers (bart)-large-Multi-Genre natural language inference features for identifying users who propagate conspiracy theories based on a rich set of 871 features in

English language. Among all the proposed models, LightGBM classifier outperformed other models with a macro F1 score of 0.87. To identify whether an article belongs to conspiracy theory or not in English language, Ghasemizade and Onaolapo [12] proposed machine learning classifiers (RF, SVM, k-NN, NB) trained with TF-IDF of word unigrams and deep learning model trained with padded and embedded text sequences. Using their respective tokenizers, tokenized and padded text sequences were taken as inputs to train the transformer models - BERT and RoBERTa. Their proposed RoBERTa model outperformed other models with a macro F1 score of 87%.

The above literature highlights extensive research efforts aimed at detecting conspiracy theories utilizing a range of machine learning, deep learning, and transfer learning models. These studies offer valuable insights into the detection of conspiracy theories. However, they do not specifically address the distinction between conspiracy theories and critical thinking within the framework of oppositional thinking. This gap suggests a need for further research to effectively distinguish between these concepts, encouraging the creation of new models for this specific type of application.

3. Methodology

We have explored machine learning and transfer learning models for distinguishing between critical and conspiracy texts in English and Spanish and the steps involved in the construction of these models are explained in the following subsections.

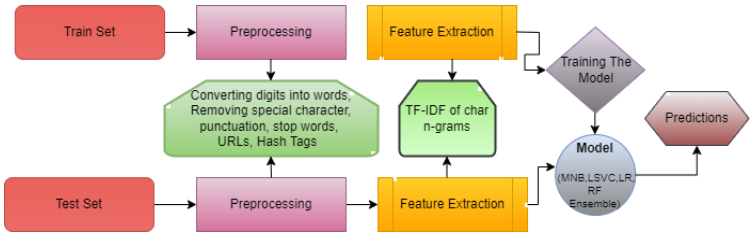

Figure 1: Framework of the proposed machine learning model

3.1. Machine Learning models

The framework of machine learning model is visualized in Figure 1 and the steps included in building these classifiers are explained below:

3.1.1. Pre-processing

Pre-processing is the preliminary step in building learning models and it involves cleaning and transforming the raw text data to a suitable format required for subsequent processing. Usually, text data contains noise in the form of: user mentions, hashtags, punctuation, digits, and hyperlinks, and eliminating this irrelevant information makes the data less complex and improves the performance of the classifier. Hence, in this work, this irrelevant information and stopwords are removed during pre-processing. Further, English and Spanish stopwords available at NLTK library[1] are used as references to remove English and Spanish stopwords respectively from the given dataset.

3.1.2. Feature Extraction

The role of feature extraction is to extract relevant features from the given data to train the learning models. TF-IDF of char n-grams is used to represent English and Spanish text. Char n-grams are sequences of n consecutive characters in a word and char n-grams in the range (1, 5) are obtained from the text and converted to TF-IDF vectors using TfidfVectorizer[2].

[1] https://www.nltk.org/search.html?q=stopwords
[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Table 3
Hyperparameters and their values used in machine learning models

Model	Hyperparameters	Values
MNB	alpha	1.0
	fit_prior	True
	class_prior	None
LSVC	C	1.0
	class_weight	balanced
	max_iter	10000
	random_state	123
LR	max_iter	1000
RF	n_estimators	100
	random_state	42

3.1.3. Model Construction

The performance of the learning model relies on the features and the classifier used to carry out classification. This work utilizes machine learning classifiers (MNB, LR, RF, and ensemble of LSVC, LR, and RF with majority voting), to distinguish between conspiracy and critical texts in English and Spanish language. A brief description of the machine learning classifiers is given below:

- **MNB** - is a probability-based classifier suitable for text classification with discrete characteristics like word frequency counts [13].
- **LinearSVC** - used from the Scikit-learn library[3] attempts to maximize the distance between classified samples by finding a hyperplane.
- **LR** - is used to predict the probability of certain classes based on dependent variables and is suitable for binary classification task. Further, regularisation approaches in LR classifiers are useful for reducing overfitting in high dimensional space [14].
- **RF** - is one of the supervised learning algorithms which is flexible and can be adapted easily to different situations but it is necessary to build a minimum number of trees in order to classify the data [15].
- **Ensemble learning** - is a strategy for building a new classifier from several heterogeneous base classifiers taking benefit of the strength of one classifier to overcome the weakness of another classifier to get better performance for the classification task [16]. In this work, three machine learning classifiers (LSVC, LR, and RF) are ensembled with hard voting to distinguish between critical and conspiracy texts.

The hyperparameters and their values used in the machine learning models are shown in Table 3.

3.2. Transfer Learning

The technique of transfer learning within the broader field of machine learning utilizes knowledge gained from one task to improve the performance of another related task. This is realized with the help of pretrained transformer models which are trained on large unlabeled data and are widely accessible and applicable to various tasks. The pretrained transformer models are fine-tuned with the given dataset to fit the models to a particular task or domain. The framework of the proposed transfer learning model is shown in Figure 2.

The text is pre-processed to clean and transform the raw text into a consistent format by converting numeric information to corresponding words, and removing URLs, user mentions, hash tags, and special characters. Preprocessing is applied to the sentences of the given text to retain the sentence structure in the text and the preprocessed text is used to fine-tune the transformer models. A brief description of the transformer models used in this study to fine-tune are given below:

[3]https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

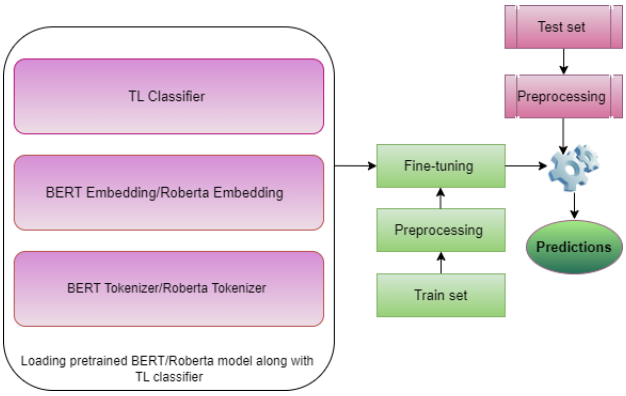

Figure 2: Framework of the proposed transfer learning model

- **BERT_base**[4] is a conceptually simple and empirically powerful pretrained language model using a Masked Language Modeling (MLM) objective trained on Toronto Book Corpus and Wikipedia and exclusively used for tasks involving English texts.
- **English_BERT**[5] - is a bilingual Legal BERT model trained with 2,000 Dutch and 6,000 English legal documents amounting to 12 GB legal text from various areas belonging to legal domain such as legislation and court cases. This domain specific BERT has resulted in improved performance compared to using standard BERT models for legal tasks.
- **CT_BERT_v2**[6] - is a BERT-large-uncased model pretrained on a corpus of messages from Twitter about COVID-19. This model is identical to covid-twitter-bert but trained on more data (40.7M sentences and 633M tokens) resulting in higher performances for many downstream applications.
- **EN_RoBERTa**[7] - is a large multi-lingual language model trained on 12.5TB of filtered CommonCrawl data. Based on Facebook's RoBERTa model, this model is fine-tuned with the conll2003 dataset in English.
- **ES_BERT**[8] - BETO: Spanish BERT is trained on the Spanish edition of Wikipedia, the OPUS Project, and Spanish books and news articles, and is exclusively used for tasks involving Spanish texts.
- **Distil_SpanBERT**[9] - is a distilled version of SpanishBERT, trained on Spanish text sources, and is also exclusively used for tasks involving Spanish texts, but is optimized for efficiency and speed.
- **Spanish_BERT**[10] - is a sentence-transformers model which maps sentences and paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search.
- **ES_RoBERTa**[11] - a variant of RoBERTa is a BERT-based model, specifically tailored for the Spanish language. It is trained on large Spanish text corpora to understand and generate contextually relevant representations of words and sentences.

We employed the above mentioned BERT variants from the Hugging Face library. The hyperparameter and their values used in the above transfer learning models are shown in Table 4. These BERT variants are fine-tuned with the pre-processed Train set and is used to train transformer classifier (ClassificationModel) to distinguish between conspiracy and critical texts in English and Spanish language.

[4]https://huggingface.co/google-bert/bert-base-uncased
[5]https://huggingface.co/Gerwin/legal-bert-dutch-english
[6]https://huggingface.co/digitalepidemiologylab/covid-twitter-bert-v2
[7]https://huggingface.co/FacebookAI/xlm-roberta-large-finetuned-conll03-english
[8]https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[9]https://huggingface.co/dccuchile/distilbert-base-spanish-uncased
[10]https://huggingface.co/hiiamsid/sentence_similarity_spanish_es
[11]https://huggingface.co/bertin-project/bertin-roberta-base-spanish

Table 4
Hyperparameter and their values used in transfer learning models

Model	Hyperparameter	Value
BERT_base	architectures	BertForMaskedLM
	attention_probs_dropout_prob	0.1
	hidden_size	768
	intermediate_size	3072
	max_position_embeddings	512
	model_type	bert
	vocab_size	30522
English_BERT	architectures	BertModel
	hidden_size	768
	max_position_embeddings	512
	model_type	bert
	hidden_size	768
	num_hidden_layers	12
	vocab_size	105879
CT_BERT_v2	hidden_act	gelu
	hidden_size	1024
	intermediate_size	4096
	model_type	bert
	vocab_size	30522
EN_RoBERTa	architectures	XLMRobertaForTokenClassification
	hidden_act	gelu
	max_position_embeddings	514
	model_type	xlm-roberta
	vocab_size	250002
ES_BERT	architectures	BertForMaskedLM
	hidden_act	gelu
	hidden_size	768
	intermediate_size	3072
	position_embedding_type	absolute
	vocab_size	31002
Distil_SpanBERT	architectures	DistilBertForMaskedLM
	model_type	distilbert
	vocab_size	31002
Spanish_BERT	architectures	BertModel
	hidden_act	gelu
	model_type	bert
	position_embedding_type	absolute
	vocab_size	31002
ES_RoBERTa	architectures	RobertaForMaskedLM
	hidden_act	gelu
	hidden_size	768
	max_position_embeddings	514
	vocab_size	50262

4. Experiments and Results

The datasets provided by the organizers of the shared task consisted of only Train set and are highly imbalanced. Statistics of the datasets are as follows:

- **English dataset**: 2,621 samples belong to 'CRITICAL' class and 1,379 samples belong to 'CONSPIRACY' class.
- **Spanish dataset**: 2,538 samples belong to 'CRITICAL' class and 1,462 samples belong to 'CONSPIRACY' class.

Table 5
Performances of the proposed machine learning and transfer learning models on English language Validation set

Models		Precision	Recall	Macro F1 Score
Machine Learning	MNB	0.84	0.51	0.42
	LSVC	0.86	0.87	**0.86**
	LR	0.82	0.79	0.80
	RF	0.84	0.80	0.81
	Ensemble	0.84	0.82	0.83
Transfer Learning	BERT_base	0.83	0.82	0.82
	English_BERT	0.92	0.89	**0.90**
	CT_BERT_v2	0.80	0.80	0.80
	EN_RoBERTa	0.34	0.50	0.40

Table 6
Performances of the proposed machine learning and transfer learning models on Spanish language Validation set

Models		Precision	Recall	Macro F1 Score
Machine Learning	MNB	0.82	0.50	0.40
	LSVC	0.83	0.83	0.83
	LR	0.83	0.77	0.79
	RF	0.81	0.75	0.77
	Ensemble	0.83	0.79	0.81
Transfer Learning	ES_BERT	0.92	0.90	**0.91**
	Distil_ESBERT	0.69	0.70	0.69
	Spanish_BERT	0.90	0.87	**0.88**
	ES_RoBERTa	0.78	0.74	0.75

As the datasets consists of only Train sets, 33% of the Train sets at random are considered as Validation sets to evaluate the performances of the proposed models for both the languages and the remaining as the Train sets. Experiments were carried out by training various machine learning models using TF-IDF of char n-grams in the range (1, 5) and by fine-tuning various BERT variants mentioned above, to distinguish between conspiracy and critical texts in English and Spanish. The performances of the proposed models were evaluated on the Validation set based on macro F1 score and the performances are shown in Tables 5 and 6 for English and Spanish datasets respectively.

The results shown in Tables 5 and 6 illustrate that transfer learning models have performed better than machine learning models. As the shared task participants were allowed to submit the predictions of only two models on the Test sets, we fine-tuned English_BERT and EN_RoBERTa with the complete English Train set and Spanish _BERT and ES_RoBERTa with the complete Spanish Train set and the predictions of these models on English and Spanish Test sets submitted to the organizers were evaluated based on MCC scores. MCC is a metric used to evaluate the quality of binary classifications especially with imbalanced datasets. The MCC value ranges from -1 to +1, where +1 indicates perfect prediction, 0 indicates no better than random prediction, and -1 indicates total disagreement between prediction and observation. MCC provides a balanced and comprehensive measure of model's performance, considering all types of classification errors. MCC scores are calculated using the following formula:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where:

- TP (True Positives) - number of correct positive predictions,
- TN (True Negatives) - number of correct negative predictions,
- FP (False Positives) - number of incorrect positive predictions,

Table 7
Performances of English_BERT and Spanish_BERT models on Test set in English and Spanish - models submitted to the shared task

Language	Model	MCC	MACRO F1 SCORE	F1 CONSPIRACY	F1 CRITICAL	POSITION
English	**English_BERT**	0.7162	0.8538	0.7994	0.9082	61
Spanish	**Spanish_BERT**	0.6293	0.8060	0.7363	0.8756	36

- FN (False Negatives) - number of incorrect negative predictions.

Among the two transfer learning models submitted to the shared task, English_BERT and Spanish_BERT models obtained MCC scores of 0.7162 and 0.6293 for English and Spanish texts securing 61^{st} and 36^{th} position respectively. The performances of these models is shown in the Table 7. The low performances of fine-tuned English_BERT and Spanish_BERT models could be attributed to the following reasons:

- The given datasets are highly imbalanced with approximately 2/3 of the total samples belonging to 'CRITICAL' class and 1/3 belonging to 'CONSPIRACY' class in both the languages. The highly imbalanced data will significantly impact the models' performance with a bias towards majority class.
- English_BERT designed for both Dutch and English may potentially reduce its effectiveness for purely English tasks. Further, domain-specific models like English_BERT might not generalize well outside their specialized contexts.
- Spanish_BERT model might be more suited for sentence similarity tasks rather than classification.

Further, differences in pre-training data, fine-tuning processes, and the values of hyperparameters used in the models could also contribute to the disparity in the performances of the proposed models.

5. Conclusion and Future Work

In this paper, we - team MUCS, describe the models submitted to Subtask-1: 'Distinguishing between critical and conspiracy texts' of the shared task "Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives" at 'PAN 2024', to distinguishing between critical and conspiracy texts in English and Spanish. Experiments are carried out with TF-IDF of char n-grams in the range (1, 5) to train several machine learning classifiers and several BERT variants are fine-tuned with the English and Spanish Train sets in transfer learning models, to distinguish the given unlabeled English and Spanish text as 'CONSPIRACY' or 'CRITICAL'. As the shared task participants were allowed to submit the predictions of only two models on the Test set, we fine-tuned English_BERT and EN_RoBERTa with the complete English Train set and Spanish_BERT and ES_RoBERTa with the complete Spanish Train set and the predictions of these models on English and Spanish Test sets were submitted to the organizers for evaluation. Among these two models, English_BERT and Spanish_BERT obtained MCC scores of 0.7162 and 0.6293 for English and Spanish texts securing 61^{st} and 36^{th} position respectively. As the given datasets are imbalanced, suitable text augmentation techniques followed by efficient text representation methods and context-aware models to distinguish between the two forms of oppositional thinking will be explored further.

References

[1] H. Sharp, Oppositional Ideas, Not Dichotomous Thinking: Reply to Rorty, in: Political Theory, Sage Publications Sage CA: Los Angeles, CA, 2010, pp. 142–147.

[2] A. Giachanou, B. Ghanem, P. Rosso, Detection of Conspiracy Propagators using Psycho-Linguistic Characteristics, in: Journal of Information Science, SAGE Publications Sage UK: London, England, 2023, pp. 3–17.

[3] K. M. Douglas, R. M. Sutton, What are Conspiracy Theories? A Definitional Approach to their Correlates, Consequences, and Communication, in: Annual review of psychology, volume 74, Annual Reviews, 2023, pp. 271–298.

[4] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification - Condensed Lab Overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, 2024.

[5] J. Moffitt, C. King, K. M. Carley, Hunting Conspiracy Theories during the COVID-19 Pandemic, Social Media+ Society 7 (2021) 20563051211043212.

[6] S. G. Thornton, R. M. Romano, Beyond Oppositional Thinking: Radical Respect., in: Philosophical Studies in Education, volume 38, ERIC, 2007, pp. 199–209.

[7] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024, pp. 3–10.

[8] M. Moosleitner, B. Murauer, On the Performance of Different Text Classification Strategies on Conspiracy Classification in Social Media., in: MediaEval, 2021.

[9] K. Tahat, A. Mansoori, D. N. Tahat, M. Habes, R. Alfaisal, S. Khadragy, S. A. Salloum, Detecting Fake News during the COVID-19 Pandemic: A SEM-ML Approach, in: Comput. Integr. Manuf. Syst, 2022, pp. 1554–1571.

[10] Y. Peskine, P. Papotti, R. Troncy, Detection of COVID-19-Related Conpiracy Theories in Tweets using Transformer-Based Models and Node Embedding Techniques, in: MediaEval 2022, Multimedia Evaluation Workshop, 12-13 January 2023, Bergen, Norway, 2023.

[11] A. Recordare, G. Cola, T. Fagni, M. Tesconi, Unveiling Online Conspiracy Theorists: a Text-Based Approach and Characterization, in: arXiv preprint arXiv:2405.12566, 2024.

[12] M. Ghasemizade, J. Onaolapo, Developing a Hierarchical Model for Unraveling Conspiracy Theories, in: EPJ Data Science, Springer Berlin Heidelberg, 2024, p. 31.

[13] P. Harjule, A. Gurjar, H. Seth, P. Thakur, Text Classification on Twitter Data, in: 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), IEEE, 2020, pp. 160–164.

[14] J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors), in: The annals of statistics, Institute of Mathematical Statistics, 2000, pp. 337–407.

[15] M. Huljanah, Z. Rustam, S. Utama, T. Siswantining, Feature Selection using Random Forest Classifier for Predicting Prostate Cancer, in: IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2019, p. 052031.

[16] M. Li, P. Xiao, J. Zhang, Text Classification based on Ensemble Extreme Learning Machine, in: arXiv preprint arXiv:1805.06525, 2018.