

**SAAD DAHLEB UNIVERSITY BLIDA 1**

**Faculty of Sciences**

Computer Science Department



**MASTER THESIS**

**In Computer Science**

Option : Software engineering

**THEME :**

# Detecting false information on social media using contextual features

**Realized by**

Mr. Farohe Sohayb

Mr. Temmar Bilal

**Under the supervision of**

Dr. Boumahdi Fatima

Mr. Hemina Karim

**Thesis Committee:**

Chair Committee : Dr. Fareh Messaouda

Examiner: Dr. Mezzi Melyara

July 20, 2023

## Abstract

The spread of fake news on the internet is a major challenge. Traditional fact-checking methods, which rely on human experts to verify information credibility, are not scalable to the volume of online content.

Automated fake news detection is a more scalable approach that uses artificial intelligence to learn and identify patterns from news data that are more likely to be fake. However, despite its advantages in terms of accuracy and response time, supervised learning solutions in automated fake news detection have failed to get ahead in identifying fake content across cultures and languages due to the limited availability of fake-checked datasets and biases in the data.

To address fake news detection biases, this study proposes a comprehensive approach to fake news detection that combines supervised and unsupervised learning strategies. The first strategy involves supervised automatic fact-checking using a Transformer model and other models including a Logistic classifier, SVM classifier, Convolutional Neural Network classifier, Naive Bayesian classifier, and XGboost classifier. These models are trained on labeled datasets to identify fake news with high accuracy. The second strategy involves unsupervised learning, which is used to handle unlabeled datasets. This strategy allows us to effectively explore and benefit from a wide range of information and insightful facts to our supervised classifiers.

We evaluate our approach on two datasets, the LIAR dataset and the ISOT Fake News Dataset. We achieve an accuracy of up to 91% on the ISOT Fake News Dataset through unsupervised relabeling. We also achieve a 100% F1-score in a supervised learning experiment, which means with the real labels of the ISOT dataset.

Our study contributes to the development of effective strategies for combating fake news, addressing the challenges posed by the growing digital area nowadays.

**Keywords:** Fake news, Misinformation, Fact-checking, Supervised learning, Unsupervised learning, Transformers.



---

## Résumé

La propagation des fausses nouvelles sur Internet est un défi majeur. Les méthodes de vérification des faits traditionnelles, qui reposent sur des experts humains pour vérifier la crédibilité des informations, ne sont pas à l'échelle du volume de contenu en ligne.

La détection automatisée des fausses nouvelles est une approche plus scalable qui utilise l'intelligence artificielle pour apprendre et identifier des patterns dans les données de nouvelles qui sont plus susceptibles d'être fausses. Cependant, malgré ses avantages en termes de précision et de rapidité de réponse, les solutions de machine learning supervisées dans la détection automatisée des fausses nouvelles n'ont pas réussi à surpasser l'identification de contenu faux à travers les cultures et les langues en raison de la disponibilité limitée de datasets vérifiés et des biais dans les données.

Pour remédier aux biais de détection des fausses nouvelles, cette étude propose une approche globale de la détection des fausses nouvelles qui combine des stratégies d'apprentissage supervisées et non supervisées. La première stratégie consiste en une vérification automatique factuelle supervisée à l'aide d'un modèle Transformer et d'autres modèles, notamment un classificateur logistique, un classificateur SVM, un classificateur de réseaux de neurones convolutionnels, un classificateur de Bayes naïf et un classificateur XGboost. Ces modèles sont entraînés sur des datasets étiquetés pour identifier les fausses nouvelles avec une grande précision. La deuxième stratégie consiste en un apprentissage non supervisé, qui est utilisé pour traiter les datasets non étiquetés. Cette stratégie nous permet d'explorer efficacement et de tirer profit d'une large gamme d'informations et de faits pertinents pour nos classificateurs supervisés.

Nous évaluons notre approche sur deux datasets, le dataset LIAR et le dataset ISOT Fake News. Nous obtenons une précision de 91% sur le dataset ISOT Fake News par re-étiquetage non supervisé. Nous obtenons également un score F1 de 100% dans une expérience d'apprentissage supervisé, ce qui signifie avec les étiquettes réelles du dataset ISOT.

Notre étude contribue au développement de stratégies efficaces pour lutter contre les fausses nouvelles, en abordant les défis posés par l'espace numérique en pleine croissance d'aujourd'hui.

**Mots Clés :** Fausses nouvelles, Misleading information, Vérification des faits, Apprentissage supervisé, Apprentissage non supervisé, Transformers.

## ملخص

الانتشار السريع للأخبار المزيفة على الإنترنت يشكل تحدياً كبيراً. تعتمد طرق التحقق من الحقائق التقليدية على خبراء بشريين للتحقق من صحة المعلومات ، لكنها غير قابلة للتطبيق على حجم المحتوى عبر الإنترنت.

يُعد الكشف الآلي عن الأخبار المزيفة نهجاً أكثر قابلية للتطبيق يعتمد على الذكاء الاصطناعي لتعلم وتحديد الأنماط من بيانات الأخبار التي من المرجح أن تكون خاطئة. ومع ذلك ، على الرغم من مزاياها من حيث الدقة ووقت الاستجابة ، فقد فشلت حلول التعلم الخاضع للإشراف في الكشف عن المحتوى المزيف عبر الثقافات واللغات بسبب محدودية توفر مجموعات البيانات التي تم التحقق منها من قبل المزييفين والتحييزات في البيانات.

لمعالجة تحيزات الكشف عن الأخبار المزيفة ، يقترح هذا البحث نهجاً شاملاً لكشف الأخبار المزيفة يجمع بين استراتيجيات التعلم الخاضع للإشراف والتعلم غير الخاضع للإشراف. تتضمن الاستراتيجية الأولى التحقق الآلي الخاضع للإشراف باستخدام نموذج رنسفورمر و نماذج أخرى بما في ذلك نموذج نموذج التعلم لأخبار الزائفة، الدعاية، التحقق من الحقائق، التعلم بالإشراف، التعلم بدون الإشراف، نماذج التعلم الآلي. يتم تدريب هذه النماذج على مجموعات بيانات مصنفة لتحديد الأخبار المزيفة بدقة عالية. تتضمن الاستراتيجية الثانية التعلم غير الخاضع للإشراف ، والذي يستخدم للتعامل مع مجموعات البيانات غير المصنفة. تسمح لنا هذه الاستراتيجية باستكشاف مجموعة واسعة من المعلومات والحقائق الثاقبة بشكل فعال لخوارزمياتنا الخاضعة للإشراف.

لقد قمنا بتقييم نهجنا على مجموعتين من البيانات ، مجموعة بيانات شياض ومجموعة بيانات يعضة سكو. لقد حققنا دقة تصل إلى ١٩٪ على مجموعة بيانات يعضة سكو من خلال إعادة التسمية غير الخاضعة للإشراف. لقد حققنا أيضاً درجة ١-سر بنسبة ٠٠١٪ في تجربة التعلم الخاضع للإشراف ، مما يعني مع الملصقات الحقيقية لمجموعة بيانات يعضة.

يساهم بحثنا في تطوير استراتيجيات فعالة لمكافحة الأخبار المزيفة ، مما يعالج التحديات التي تفرضها المنطقة الرقمية المتنامية في الوقت الحاضر.

**الكلمات المفتاحية:** لأخبار الزائفة، الدعاية، التحقق من الحقائق، التعلم بالإشراف، التعلم بدون الإشراف، نموذج التعلم الآلي

# Contents

<b>List of Figures</b>	<b>i</b>
<b>List of Tables</b>	<b>iii</b>
<b>General Introduction</b>	<b>1</b>
<b>1 Overview of Fake News and related works</b>	<b>3</b>
1.1 Introduction . . . . .	3
1.2 Fake News . . . . .	3
1.3 Automatic fake news detection . . . . .	4
1.3.1 Content approach . . . . .	5
1.3.2 Context approach . . . . .	7
1.3.3 Hybrid approach . . . . .	8
1.3.4 Synthesis . . . . .	9
1.4 Datasets . . . . .	11
1.4.1 Main datasets used . . . . .	11
1.4.2 Synthesis . . . . .	13
1.5 Conclusion . . . . .	15
<b>2 Proposed approach</b>	<b>16</b>
2.1 Introduction . . . . .	16
2.2 Dataset Exploratory Analysis . . . . .	17
2.3 Fake news detection classifier training . . . . .	19
2.3.1 Data Processing . . . . .	20
2.3.2 Dataset labeling through unsupervised learning . . . . .	24
2.3.3 Latent Feature Representation Generation Through Transformers . . . . .	26
2.3.4 Classifiers Training . . . . .	32
2.4 Evaluations and results . . . . .	36
2.4.1 Implementation tool . . . . .	36

2.4.2	Measuring the Performance . . . . .	36
2.4.3	Results . . . . .	38
2.4.4	Discussion . . . . .	51
2.5	Conclusion . . . . .	52
	<b>General Conclusion</b>	<b>54</b>

# List of Figures

1.1	Frequency of online news sources reporting fake news U.S. 2018) [7]	4
1.2	Classification of Fake news	5
2.1	Global architecture	17
2.2	Distribution of labels in the LIAR Dataset for : (a) the Test partition, (b) the Training partition and (c) the Validation partition	18
2.3	Analysis of LIAR Dataset	18
2.4	Word cloud	19
2.5	Approach pipeline	20
2.6	Degree of feature correlation with their assigned label	21
2.7	The Label transformation process	22
2.8	The label distribution before and after the processing	22
2.9	The Label Distribution after processing	23
2.10	The examples of the cleaning steps.	23
2.11	Ensemble Learning Process Explained	24
2.12	BERT embedding illustration	26
2.13	Transformer architecture [46]	27
2.14	Global Flow Diagram Of The Latent Representation Step	27
2.15	Loading the pre-trained Distilbert	28
2.16	Example of word tokenization through Distilbert tokenizer	30
2.17	A screenshot of the dataset after tokenization	30
2.18	Fine-tuning of pre-trained DistilBERT (in the first epoch)	31
2.19	Latent representation generator	31
2.20	Screenshot of the final produced dataset	31
2.21	A full zooming under one text instance of dataset after the latent Representation generation	32
2.22	The global flow diagram of the final classification step	33
2.23	Supervised learning explained [48]	33



2.24	Vector to Matrix Reshaping to news class classification through CNN model . . . . .	34
2.25	The Confusion Matrix Table [53] . . . . .	37
2.26	The evaluation of the Distelbert model on the LIAR dataset using the real labels of the LIAR dataset. . . . .	40
2.27	The evaluation of the ML models on the LIAR dataset using the real labels of the LIAR dataset. . . . .	41
2.28	The evaluation of the CNN model on the LIAR dataset using the real labels of the LIAR dataset. . . . .	42
2.29	The evaluation of the DistelBERT model on the LIAR dataset using unsupervised labeling. . . . .	43
2.30	The label distribution in ISOT Fake News Dataset . . . . .	45
2.31	The evaluation of the Distelbert model on ISOT Fake News dataset using the real labels of the dataset. . . . .	46
2.32	The evaluation of the ML models on ISOT Fake News dataset using the real labels of the dataset. . . . .	47
2.33	The evaluation of the CNN model on ISOT Fake News dataset using the real labels of the dataset. . . . .	48
2.34	The evaluation of the DistelBERT model on ISOT Fake News dataset after unsupervised labeling. . . . .	50

# List of Tables

1.1	Comparative table of the Related Works . . . . .	10
1.2	Comparative Table of the Datasets . . . . .	14
2.1	Comparison of the proposed methods with the state of the arts on the LIAR dataset. . . . .	44
2.2	Comparison of the proposed methods with the state of the arts on the ISOT Fake News Dataseaset dataset. . . . .	51

# General Introduction

In recent years, the proliferation of fake news due to the expansion of Social Media has become a significant concern, posing serious threats to society, politics, and journalism. The most significant example these past few years is the COVID-19 pandemic[1]. Another more local and recent example in our country concerns all the fake news that bombarded the Internet after the football match between Algeria and Cameroon for the qualification to the WORLD CUP 2022, and its scandalous arbitration and Algeria's controversial defeat like : a fake FIFA press release announcing the automatic qualification of Algeria[2] and mostly the claim of the match being redone<sup>1</sup>. Moreover, it can cost the governments a lot and make the world suffer, so the main question is how can fake news gain public trust?

By combating Fake News, we can strive to maintain the integrity of factual information and ensure that people have access to accurate and reliable news.

Traditional fact-checking methods, which rely on human experts to verify information credibility, struggle to keep up with the enormous volume of online content. Human experts, no matter how knowledgeable and efficient they are, cannot cover the vast amount of content that requires fact-checking.

The objective of this thesis is to develop a robust and reliable fake news detection framework that utilizes contextual features present within the text.

The first chapter is an overview on Fake news and the related works to this subject. After defining fake news and showing the close relation it has with social media expansion we talked about automatic fake news detection and the ways it can be classified by mentioning the several works related to each different classification and to summarize the state of art we added a synthesis table to finally conclude this chapter with an overview of different fake news datasets.

---

<sup>1</sup><https://elwatan-dz.com/commentaire-la-fifa-la-faf-et-les-promesses-mensongeres>

Concerning the second chapter which is about our Proposed Approach, we propose a comprehensive approach that combines supervised and unsupervised learning strategies to address the fake news problem. The first strategy involves supervised automatic fact-checking using the Transformer model. By training these models on labeled datasets, our aim is to automate the process of identifying fake news effectively. The second strategy, unsupervised learning, is employed to handle unlabeled datasets. This strategy allows us to effectively explore and benefit from a wide range of information and insightful facts to our supervised classifiers.

In conclusion, this Master 2 thesis seeks to address the pressing challenge of fake news detection by leveraging contextual features and employing advanced machine learning techniques. By developing a robust and reliable framework, we aim to contribute to the efforts in combating the spread of misinformation and safeguarding the integrity of news dissemination in the digital age.

# Chapter 1

## Overview of Fake News and related works

### 1.1 Introduction

In this chapter, we would like to start by giving a brief define to fake news and explaining the motives behind this study. It consists of 4 sections, section 1.2 to give an overview and define fake news, section 1.3 will address the different used strategies for fake news detection and mention some of the related works that achieved a good contribution in this field, section 1.4 to list the most used available datasets, and finally section 1.5 a conclusion to summarize the work done in this chapter.

### 1.2 Fake News

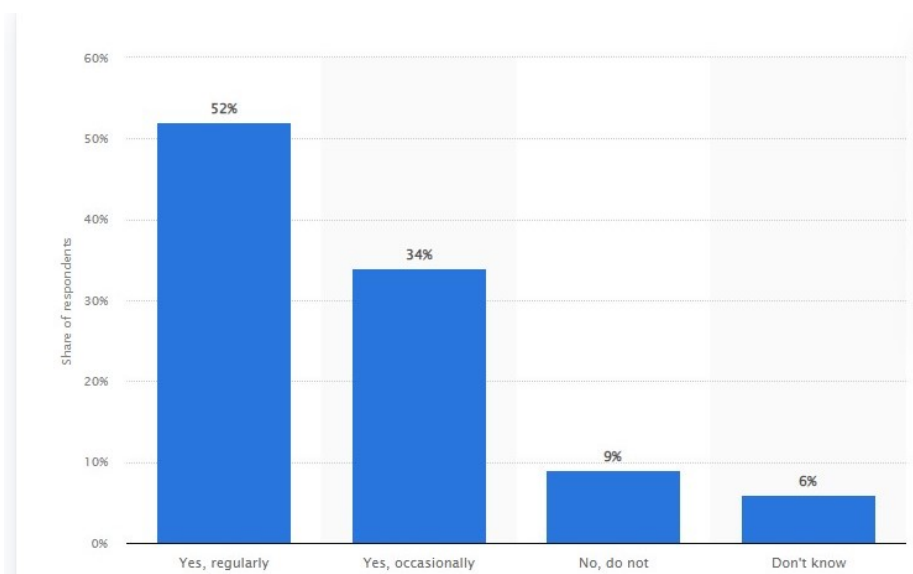
Because of the high accessibility of the Internet this past decade, the use of social media increased immensely and became a powerful and easily accessible source of information. Now, almost everyone has an account on Social Networks for personal and professional purposes.

Unfortunately, this expansion of Social Media didn't come without its fair share of new risks and dangers. Amongst them, we have cyberbullying, a threatening behaviour via the internet against a victim not being able to defend himself [3]; clickbait, a way of advertisement of a website or other content online to make the user click on one of the links accompanying [4] or blackmail which is threatening to expose any secret tending to subject any person to hatred or ridicule, or to impair his credit or business repute, and to not do it, the person has to do give something in return of their silence[5]. But our main

focus in this work will be on Fake News.

Fake news, also called false news, are articles of news that are purposely written to lead or mislead readers[6].

Fake news have a tremendous reach on Social media . The figure 1.1 shows how often people are exposed regularly to fake news in the USA in 2018.



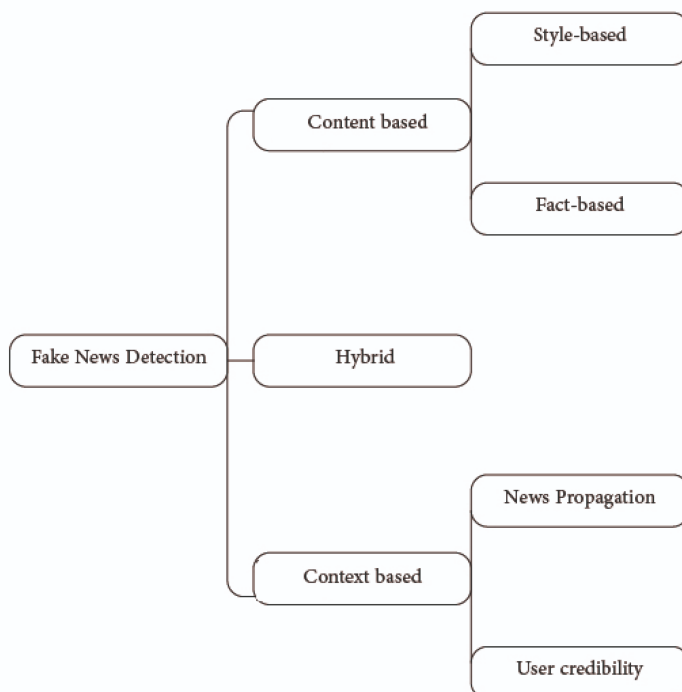
**Figure 1.1:** Frequency of online news sources reporting fake news U.S. 2018) [7]

Fake news can be extremely dangerous by tarnishing reputations, ruining businesses, and swaying political decisions. Also, it does not limit itself to politics, Fake news can touch every aspect of society, for example on health. One notable example is the COVID-19 pandemic with fake news spreading rampantly, sometimes absurdly like 5G is causing a huge strain on the immune system which helps the spread of the Coronavirus or that drinking alcohol can protect people against COVID-19 [1] .

### 1.3 Automatic fake news detection

In this section, we will discuss how fake news can be fought. First, we can manually fact-check the news that seems false. Fact-checking is a process of verifying the accuracy and reliability of information, claims, statements, or news before presenting them as true or valid. It involves conducting thorough research, cross-referencing sources, and validating evidence to ensure that the information being reported is supported by credible and verifiable data.[8] Readers exposed to a piece of fake news will correctly question the veracity of that information when they are subsequently exposed to a fact-checking post

debunking the fake post. Fact-checking groups have also emerged across the world. But studies show that while fact-checking helps in correcting disinformation, it does not work on everyone. To some, exposure to the correction might have the reverse effect and instead, reinforce belief in the original false information. For that reason, fake news should be detected as quickly as possible to be swiftly purged. Fact-checking is too slow as it is done manually. The better alternative is to do it automatically with what is called Automatic Fake News Detection (FND). It can be complex and has multiple axes of analysis depending on what information is used for classification as shown in figure 1.2



**Figure 1.2:** Classification of Fake news

### 1.3.1 Content approach

Automatic FND can be classified by content, this approach is based on checking news content, news headline, content, publication time, partisan information, etc. and match it against credible sources. The characteristics of fake news are usually extracted from the news-related features ( e.g. news content) [9] . Fake news can differ from true news

in terms of writing style and quality [10],

We can divide the context based approach in 2 methods[11]

- Fact-based : this method is mainly based on fact-checking the information in the news by using external sources like knowledge graphs, online encyclopedias or directly scholar articles which is more efficient than having a person checking the veracity of each and every news.
- Style-based : This method searches a certain pattern in the way the text is written, some features that are shared in multiple fake news like emphasizing on the punctuation : the excessive usage of exclamation or interrogation point, or the writing style [12]. We then can deduce that the news are mostly likely fake if they follow this specific pattern.

Yang and al [13] and al used TI-CNN, a unified model that is trained with image information and text at the same time. Combining them has given great results, using latents features and explicit features and makes the model expandable. After comparison with other methods like simple CNN, LR, LSTM and GRU, the TI-CNN model is the most accurate with a ratio of precision of 0.9220, a recall of 0.9277 and F1-measure of 0.9210.

Sastrawan [14] compares 3 different deep learning methods : CNN, Bidirectional LSTM, and ResNet after word embedding pre-training on 4 different Datasets : The ISOT Dataset [15], The FakeNews Dataset[16] , the Fake or Real News Dataset [17] and the Fake News Detection Dataset [18]] . We can see that Bidirectional LSTM gave the best results with more than 94 % Padnekar [19] classified the stances of articles towards a specific target into four different classes (Agree, Disagree, Discuss, Unrelated) using Bi-directional Long Short Term Memory and Autoencoder via the FNC-1 corpus Dataset [20]. They achieved a 94% accuracy, a precision of 93.725%, a recall of 93.88%, and an F1 Score of 93.88%

Khattar [21] used a bimodal ( that uses textual + image information ) variational autoencoder coupled with a classifier that classifies the posts as fake or not by using the representations given by the autoencoder called MVAE; This model outclasses other multimodal methods utilized in the study, respectfully Neural Talk, VQA, EANN and finally att-RNN to attain a score of 74.5% on the Twitter Dataset [22] and 82.4 % of accuracy with the Weibo Dataset[23]



### 1.3.2 Context approach

We can use a different approach to classify the Automatic FND, the context one. This one deals with author of news such as classifying users as normal users and malicious-user (e.g. social bots [24]), or building a rating system based on the examination of authors' profiles, create some credibility each one, checking their previous posts, their propaganda traceability, analysis of writing patterns in fake and genuine news articles.

Like the content approach, the context one can be split in 2 distinct methods:

- News Propagation: This one searches for a pattern in the way the news is being spread on the Internet if it is similar to other fake news.
- User credibility: This method analyzes the information of the user, its account on Social Media like Twitter to determine if the author of the news or the person who is sharing it is a trustworthy person or some bot/real person with malicious intentions and thus determining if the News is fake.

Feng [25] tried to detect if a Twitter account is a bot by using the Twitter user's description, the number of tweets posted and this user's neighborhood information via the DataSet TwiBot-20[25] to make a heterogenous graph where the user is represented as a node . Then, they apply RGCN to determine if the Twitter user is a bot or not. After comparison with other models, it is proved that using RGCN is more accurate than CNN, MLP and GAT with an accuracy of 84.62%, a F1-score of 0.8707 and an MCC of 0.7021

Aljohani [26] with the same objective studied the altmetrics using Social Network Analysis ( SNA ) and also using GCN to classify bots in Altmetrics using the fake news datasetfake [16] and the Polifact dataset [27] . They started by using Altmetrics DS then via SNA, the data is represented in a Network. Then the model uses the DGCN to classify the Twitter user as a boot or no. This study got results of 0.291 for an average CC, an average degree of 3.86, a F1 score of 0.67 and an accuracy of 71

Zhou [28] in this study tries to analyze and classify the patterns of fake news on Social Media. This pattern is defined by the concerned news, the accounts spreading of the news and the relationships between the spreaders . They came out with 4 different patterns : More-Spreaders pattern, Farther-Distance pattern, Stronger-Engagement pattern, Denser-Networks pattern. Then, the study used different classifiers: Support VectorMachine (SVM), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Decision Trees (DT) and Random Forests (RF).

This model utilized the 2 Datasets , the first one is a Twitter Dataset [22] that contains 17,857 data from 15,821 users, with 7,244 real data and 10613 fake data. The second is

the Weibo Dataset[23] composed of data collected from Xinhua News Agency, a Chinese news source and Weibo, the Chinese biggest social network platform. After experimenting with the classifiers mentioned, the RF outclass the other methods on both of the used datasets. The results gotten are better than other approaches, content-based for example and similar to a hybrid method. Then, after comparing every pattern with each other and different combination of multiple patterns, the best results was given by the mix of all 4 patterns adding to that network similarity with an respectively and accuracy and a F1 Score of 0.929 0.932 for the Twitter Dataset and for the Weibo one 0.835 and 0.842. These results can be even improved with the addition of linguistic features.

Abdulrahman [29] used 4 method of text feature extraction : TF-IDF, Count Vector, Character Level Vector, N-Gram level vector and classify the Fake news dataset by using 10 different Machine Learning and Deep Learning classifiers, respectfully RF, k-NN, NB multinomial, Linear SVM, LR, AdaBoost, XGB for ML and ANN with Keras, RNN with LSTM and CNN with LSTM for DL with an accuracy that exceeds 80 % and in the case of a CNN+LSTM and AdaBoost with TFI-IDF a score of 100%.Also, the Dataset used is the Fake News Dataset[16]

Likewise, Shaikh [30] also used SVM, NB, Passive Aggressive Classifier with the addition of the feature extraction technique TF-IDF using the Fake or Real NEWS [17] Dataset. The SVM with TF-IDF got the best results with an accuracy of 95.05%

### 1.3.3 Hybrid approach

In addition to the two former approaches and because of their good results but still not optimal results, scientists arrived at the possibility of identifying fake news by fusing content and context methodologies to attain the best rendement possible. Using the combination of the both of them certainly will improve our results.

For example, after using 4 different features (text content, images, propagation information and the user features) Li [31] considered FN as an anomaly and used the autoencoder as the base of an unsupervised learning method. Then, they compared the results with other methods like Isolation Forest (ISF), One-Class SVM (OCSVM), simple autoencoder (AE), and Variational autoencoder (VAE). The results showed that the Unsupervised FND Based on Autoencoder outclassed them with a score of 70.17% for AUC; 60.97% Macro-F1 and 84.59% Micro-F1.

By using the user's credibility and the veracity of the news as variables, Yang[32] created a Bayesian Network model with a collapsed Gibbs sampling. After assessing them with two Datasets, the LIAR DataSet [33] and the BuzzFeed Dataset[34] ( also available on Github[35] ), we got respectfully got an accuracy of 75,9% and 67,9 %

### 1.3.4 Synthesis

The comparison of the earlier mentioned models can be accomplished by assessing their features, the datasets utilized and the results including : the accuracy, the precision, the Recall and the F1-Score.

Reference	Type of FND	Type of learning	model used	Word embedding methods :	Accuracy	Precision	F1 score	RECALL	Datasets
[25]	Context	Supervised	RGCN		84.62%		87.07%		[25]
[26]	Context	Supervised	DGCN		71%		67%		[16][27]
[36]	Hybrid	Supervised	RNN		Real news : 89.2% Fake news : 95.3 %		Real news : 89.4 % Fake news : 95.4%		[22] [23]
[14]	Content		Bidirectional LSTM	Glove	99.95%	99.95%	99.95%	99.95%	[15]
				FastText	98.65%	98.64%	98.65%	98.66%	[16]
				Glove	94.6%	94.58%	94.59%	94.64%	[17]
				fast/Text	99.24%	99.19%	99.23%	99.26%	[18]
[19]	Content	Supervised	LSTM + Auto encoder	Word2Vec	94%	93.725%	93.88%		[20]
[31]	hybrid	UnSupervised	Auto-encoder			Real News : 89.83% Fake News : 13.57%			[23]
						Real News : 91.06% / Fake News : 28.04%			[23]
[21]	Content	Supervised	Auto-encoder		74.5%	Real News : 80.1% Fake News : 68.9%	Real News : 75.8 % Fake News : 73.0%	Real News : 71.9% Fake News : 77.7%	[22]
					82.4 %	Real News : 85.4% Fake News : 80.2%	Real News : 80.9 % Fake News : 83.7%	Real News : 76.9 % Fake News : 87.5%	[23]
[28]	Context	Supervised	RF		Real News : 92.9% Fake News : 83.5%				[37]
[29]	Content	Supervised	SVM	Extraction TF-IDF	95.05%	92.56%	93.141%	93.73%	[16]
[38]	Hybrid	Unsupervised	GTUT Graph based approach		80 %	77%/ 83%	80% / 79%	83% /76%	[27]
					77%	72 % / 87%	81% /68%	93% / 56%	[39]
[32]	Hybrid	Unsupervised	Collapsed Gibbs Sampling		75.9%	76.6 % / 75%	77.4 % / 74.1%	78.3% / 73.2%	[33]
					67.9%	66.7%/69.2%	69.0% / 66.8%	71.4 % / 64.3%	[34]

Table 1.1: Comparative table of the Related Works

As shown in the table 1.1, the vast majority of the mentioned studies primarily rely on supervised Fake News Detection (FND), which offers advantages and yields high results. However, it necessitates a dependable dataset for model training, resulting in time-consuming efforts and additional work. Unsupervised FND, while presenting a greater challenge due to the absence of labeled data, offers its own set of advantages. It does not rely on pre-labeled examples for training, making it more adaptable to dynamic and evolving fake news patterns. Instead, it utilizes techniques such as anomaly detection, clustering, or natural language processing algorithms to identify suspicious patterns or abnormal behaviors within the data.

## 1.4 Datasets

This section will provide a little overview of the Fake News datasets mentioned earlier. We aim to present a collection of diverse reliable datasets that can be used for this thesis and highlight their features.

### 1.4.1 Main datasets used

1. Fake News dataset (news reliability): This dataset [16] contains news that may be fake news. Run by the UTK Machine Learning Club. this dataset used to Build a system to identify unreliable news articles , This dataset contains the following partitions:
  - train dataset : test.csv: (in csv format) this dataset used for train proposes ,it contains 20.8k entries, and it employs the following attributes:
    - id: unique id for a news article.
    - title: the title of a news article.
    - author: author of the news article.
    - text: the text of the article; could be incomplete.
    - label: defines the class of article (reliable/unreliable classification),as following 1 for unreliable article and 0 for reliable.
  - validation dataset : test.csv: this dataset used on testing purposes contains 5,2 k entries ,and the same attributes as the training dataset except the label. submit.csv. The missing label attribute from test.csv is included in this partition, along with an id attribute that serves as a foreign key reference to the id in test.csv.at least it has the same volume of entries as test.csv.

2. PolitiFact dataset [27] is an high-quality fact-check dataset collected from the well-known fact-checking website PolitiFact where Experts have fact-checked 21,152 claims in the dataset (amount of records is 21,152 entry ). The article statement inside the dataset are multi-labeled where all of the records are classified as true, mostly true, half true, mostly false, false, or pants on fire. The dataset is available on Kaggle.<sup>1</sup> it employs the following attributes:

verdict: The fact-checking verdict in one of six categories true, mostly-true, half-true, mostly-false, false, and pants-fire.

statement\_originator: the originator of the statement being fact checked.

statement: the fact checked statement .

statement\_date: the date when the statement had been fact checked .

statement\_source: the source where the statement was created.speech,television,news,blog,social\_

factchecker: name of the person who fact checked the claim.

factcheck\_date: date when the fact checked article was published.

factcheck\_analysis\_link: quicklink to the fact checked analysis article.

3. The FakeNewsNet dataset [40] includes 17441 real articles and 5755 fake articles that were assembled from fact-checking websites like PolitiFact7. It has features such as content information (linguistic and visual), social context information (user, post, response, and network), and spatiotemporal information (spatial and temporal).
4. The Arabic fake news dataset ,collected by Ashwaq et al [41] for the purpose of determining the credibility of the articles. These articles are divided into three categories: credible articles, uncredible articles, and undecided articles. The dataset consists of 606912 articles, of which 207310 are credible, 167233 are uncredible, and 232369 are undecided.
5. The LIAR dataset[33] is a publicly available dataset that contains over 12,800 short statements made by politicians in the United States. The statements were collected from a variety of news sources and were annotated by professional fact-checkers to indicate their veracity.

---

<sup>1</sup><https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset>

Each statement is labeled as either "True," "Mostly True," "Half True," "Mostly False," "False," or "Pants on Fire" based on its degree of factual accuracy. The dataset also includes information about the speaker, the subject of the statement, and other contextual information.

6. The ISOT (Inauthentic Social Media and Online Content) dataset [15] is another publicly available dataset that contains over 21,000 examples of social media posts and articles that have been identified as either "authentic" or "inauthentic" by expert human raters.

The dataset includes posts from a variety of platforms, including Twitter, Reddit, and various news websites, and covers a range of topics and themes. Each example is labeled as either "real" or "fake" based on its degree of authenticity, and the dataset also includes information about the source of the post, the language used, and other contextual information.

Fake or Real News Dataset [17]: This dataset is a collection of data specifically curated for the purpose of training and evaluating models for fake news detection. It comprises a diverse range of news articles or text samples labeled as either "real" or "fake" to provide a ground truth for classification tasks. This dataset contains more than 11000 statements.

## 1.4.2 Synthesis

The table 1.2 was create to highlight the different features of the datasets mentioned earlier like the number of records, the topic of the dataset and the different attributes of the datasets . Having a comparative table of datasets facilitate the process of comparing them. The size of Datasets can differ, from 11000 to 60 000 statements. Based on the number of different attributes, we can see that the PolitiFact and the Liar Dataset especially stand out with more then 10 different attributes.

Name	Topic	Number of records	Real news	Fake News	undecided	Attributes
fake news dataset	Politics	26 000	-	-	-	{Id,title,author,text,label}
PolitiFact dataset	Politics	21 152	-	-	-	{verdict,statement_ originator, statement, statement_ date,statement_ source, factchecker, factcheck_ date, factcheck_ analysis_ link}
FakeNewsNet dataset	Politics	23196	17441	5755	-	{ linguistic, visual, user, post, response, network, spatial , temporal }.
the Arabic fake news dataset	Politics	606912 articles	207310	167233	232369	{title , text, published_ date,source}
LIAR Dataset	Politics	12,836 statements	-	-	-	{ ID, Label, statement, subjects, speaker, speaker's job title, state info, party affiliation, total credit history count, context }
ISOT Dataset	Politics	17903 statements	-	-	-	{title, text, subject, date }
Fake or Real News Dataset	Politics	11000 statements	-	-	-	{title, text, subject, date, category}

Table 1.2: Comparative Table of the Datasets



## 1.5 Conclusion

After introducing the concept of fake news, defining it, explaining why it is such a serious issue in today's society and giving a general overview on it, we went to talk about automatic FND, the only real answer to it. We explained how Automatic FND can be classified and mentioned several studies on it with their results and some of the datasets used.

In the end, we decided to use an hybrid approach composed of an unsupervised part which can and a supervised part to get more high quality results. The LIAR and the ISOT Datasets were also chosen. The LIAR Dataset has an unique feature: the 'context' column. The presence of this column provides us valuable contextual information surrounding each news statement that can be crucial in understanding the nuances and subtleties of fake news articles. The ISOT dataset Dataset primarily focuses on detecting fake news and is well-suited for developing and evaluating fake news detection models. And with that, we conclude this chapter and transition to the pivotal aspect of our thesis: the implementation

# Chapter 2

## Proposed approach

### 2.1 Introduction

The proliferation of fake news on social media platforms has emerged as a critical challenge in today's digital landscape. This rapid spread of misinformation significantly impacts public opinion, damages reputations, and influences decision-making processes. However, traditional fact-checking methods struggle to keep up with the enormous volume of information shared online, necessitating the development of innovative approaches for effective fake news detection.

To address this problem, our proposed approach focuses on detecting fake news by effectively handling the massive volume of unlabeled data available on the internet and social media. This includes millions of unverified blogs, tweets, posts, and digital articles that have not been fact-checked. By employing unsupervised machine learning techniques, we label these samples and extract latent representations using pre-trained DistilBERT. This enables us to ensure high-quality feature representation that captures the essential characteristics of the input content. In the final step, we utilize supervised learning models to classify the credibility of the news.

Furthermore, Current contributions on fake news detection primarily utilize supervised learning methods, relying on labeled content during training their supervised models. This approach suffers from limitations due to the availability of small fact-checked datasets. However, our solution stands out by leveraging both supervised and unsupervised learning techniques, allowing it to natively handle unlabeled content without depending solely on pre-existing labels. The integration of both approaches enhances our system's ability to effectively detect fake news.

As shown in figure 2.1 our work consists of two main parts:

- Dataset Exploratory Analysis 2.2.
- Fake News Classifier Training 2.3.

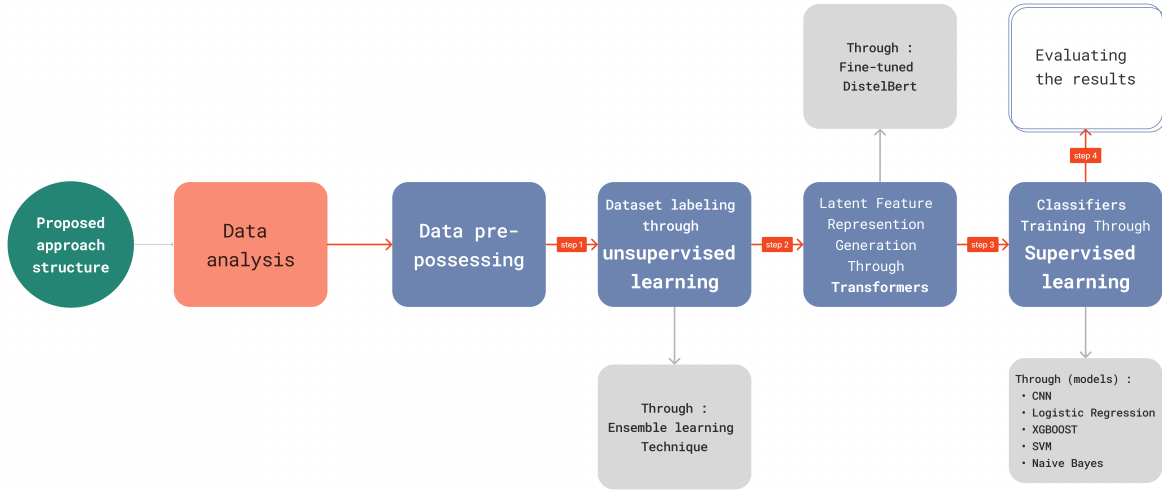


Figure 2.1: Global architecture

## 2.2 Dataset Exploratory Analysis

The data analysis section offers valuable insights for the readers, facilitating a better understanding, interpretation and effective analysis of the LIAR dataset. Within this section, we explore several key aspects, including the label distribution, the most frequent subjects treated and the primary sources that contributed to the dataset. Visualizing the label distribution provides an overview of truthfulness levels, while identifying the most prevalent subjects gives insights into the dataset's content and contribute to a better understanding of the dataset's characteristics.

### 2.2.0.1 Information distribution

- Figure 2.2 provides information on the distribution of labels in all fragments of the dataset : the Test partition, the Training partition and finally the Validation partition.

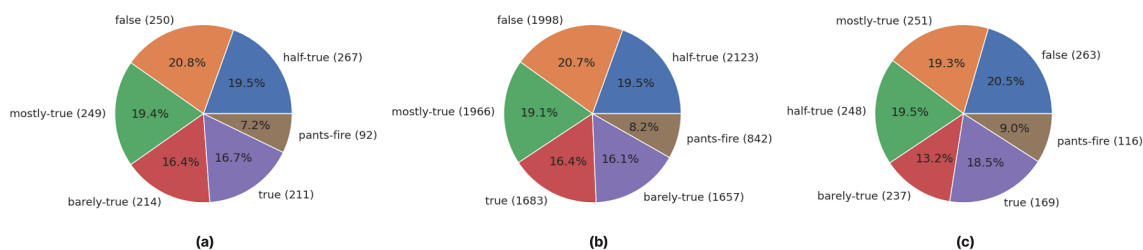


Figure 2.2: Distribution of labels in the LIAR Dataset for : (a) the Test partition, (b) the Training partition and (c) the Validation partition

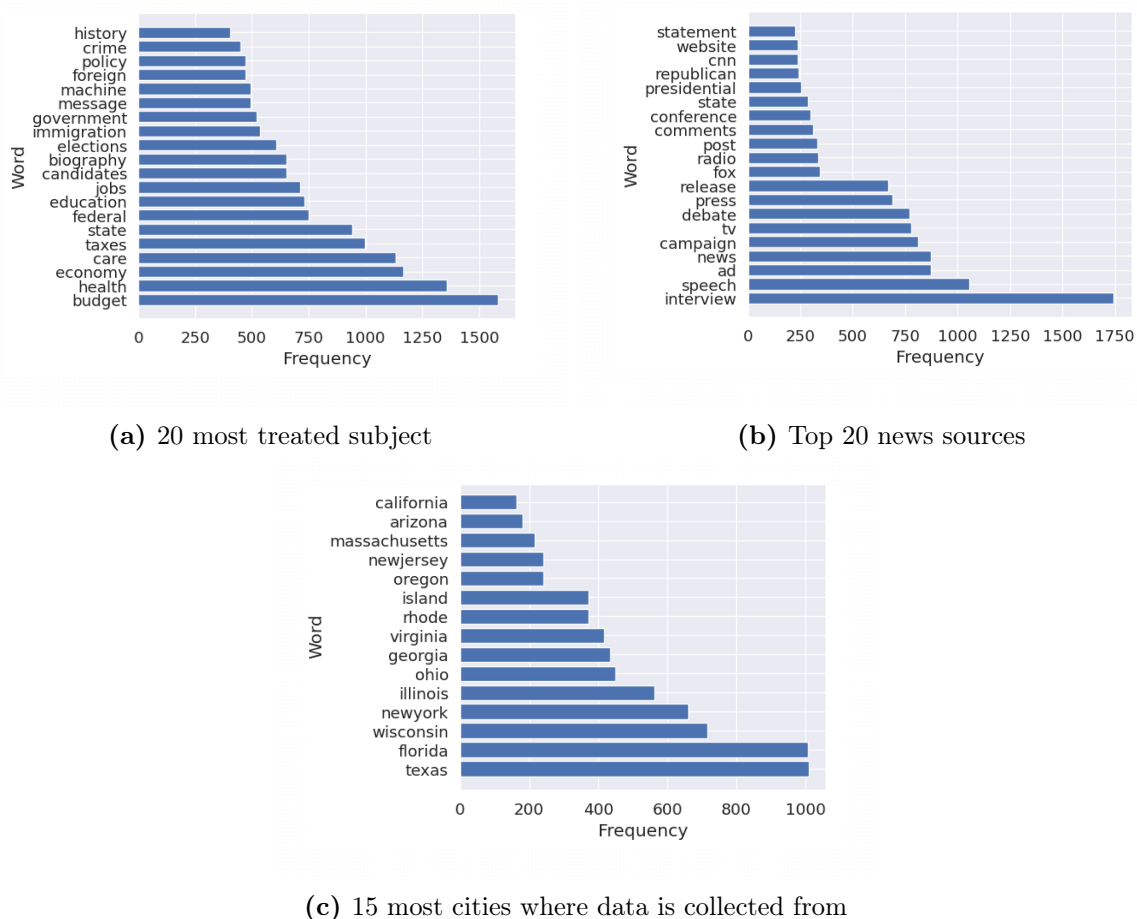


Figure 2.3: Analysis of LIAR Dataset

Figure 2.3 is a representation of the data, where Figure 2.3a shows the most frequently treated subjects in the dataset. as can be seen, budget, health and economy are the most prominent. For Figure 2.3b, it shows the top 20 news sources represented in the dataset, with interviews, speeches and advertisements being the most frequent. Finally, Figure 2.3c

illustrates the top 15 U.S states from which LIAR NEWS has gathered data, showing that the major states sources were Texas, Florida and Wisconsin.



Figure 2.4: Word cloud

## 2.3 Fake news detection classifier training

The integration of unsupervised methods in the proposed approach shows promise in mitigating the limitations of labeled data dependency. Despite the limited research in unsupervised fake news detection, the presence of a few notable works further underscores the preference for exploring unsupervised approaches. However the proposed approach is structured as follows :

- Dataset Pre-processing [2.3.1](#).
- Dataset labeling through unsupervised learning [2.3.2](#).
- Latent Feature Representation Generation Through Transformers [2.3.3](#).
- Classifiers Training [2.3.4](#).
- Testing and validating our Classifiers [2.4](#).

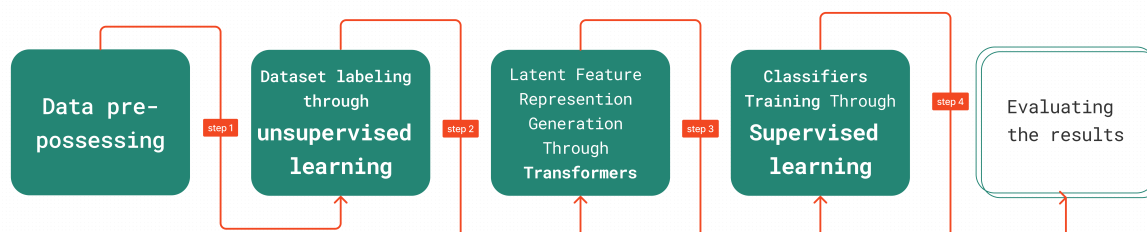


Figure 2.5: Approach pipeline

### 2.3.1 Data Processing

In order to ensure high-quality data input for our NLP models and effective data visualization, the preparation and cleaning of the data are obligatory steps. This involves several key steps, such as removing unnecessary features, the transformation of labels, the cleaning of text, the scale of the input, the embedding, and other strategies depending on the type of data and context of classification.

#### Removing of unnecessary features

A majority of datasets contains unnecessary columns included columns that correlate in trouble way with label ( due to our use of classification ). Thus, we remove these columns like the "speaker" column and its job title because the model is supposed to generalize the prediction of the fake news based on speech not to a specific speaker.

Figure 2.6 supports the idea of removing unnecessary features from datasets. It highlights the presence of columns that exhibit correlation degrees with the label. As observed, we removed all columns exhibiting a negative degree of correlation during the preprocessing stage.

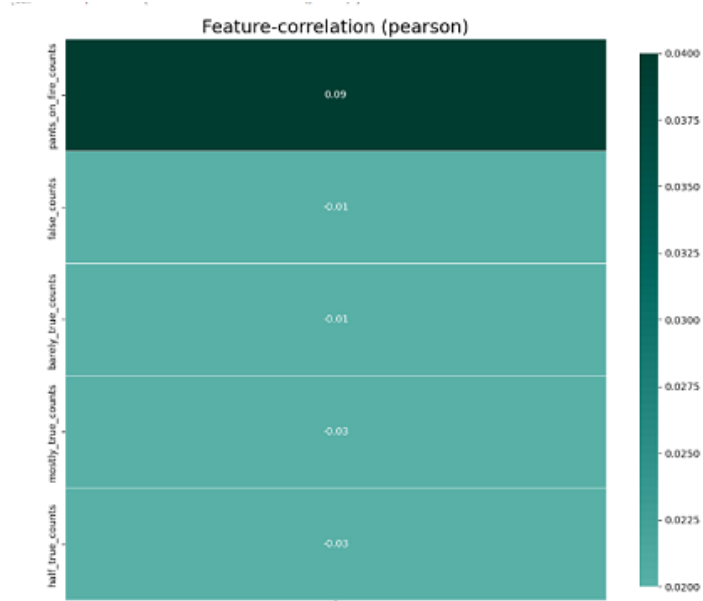
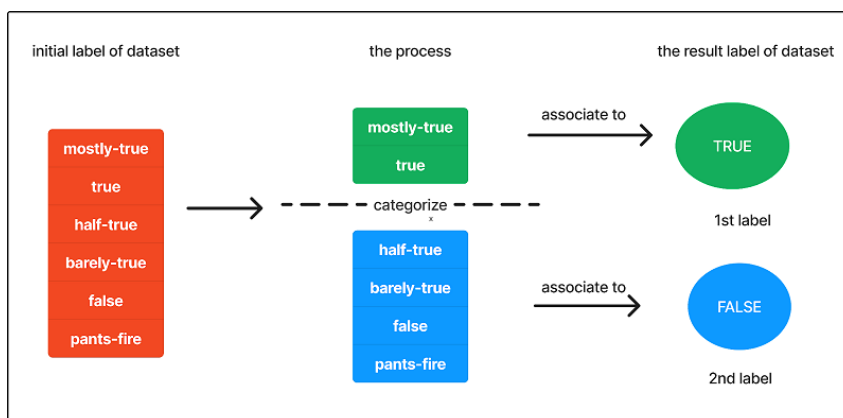


Figure 2.6: Degree of feature correlation with their assigned label

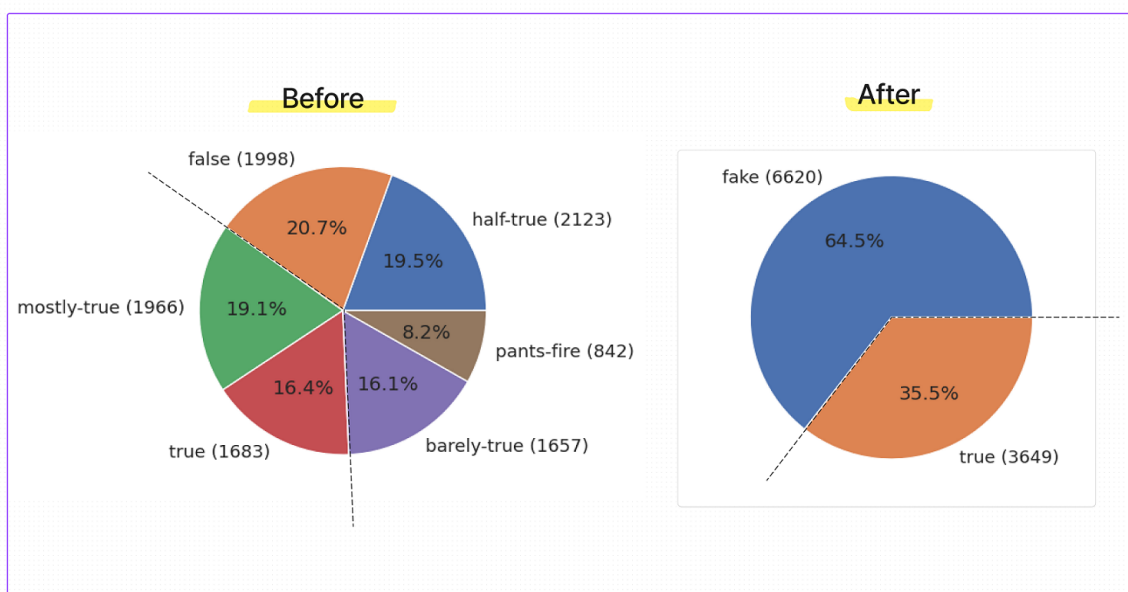
### Label transformation

In order to facilitate a comprehensive and comparative analysis, we have adopted the same label transformation technique employed by several relevant projects[42]. The original dataset, LIAR, contains a multi-labeled column called "label" with six distinct labels representing the nature of news for each instance: 'false', 'half-true', 'mostly-true', 'true', 'barely-true', and 'pants-fire'. To simplify our classification problem, we applied a label transformation process that converts this multi-classification problem into a binary classification problem. As a result, each instance in the dataset is assigned either a "true" or "false" label as illustrated in figure 2.7. There is several reasons for using the binary classification like reducing the complexity of our problem. Also, our primary focus is distinguishing between fake news and non-fake news, and not between various degrees or types of fake news. This transformation allows us to effectively compare our results with other projects.



**Figure 2.7:** The Label transformation process

The figure 2.8 shows the Label Distribution in the train partition **Before** and **After** applying the Label transformation process. We can see that it went from 6 labels of roughly equal value to a more explicit 2-part division of our data, with a little majority of fake news ( 64.5% ) and a good minority of true news ( 35.5 % ) . The fig 2.9 also illustrate that by showing us the number of fake and true news in this partition.



**Figure 2.8:** The label distribution before and after the processing





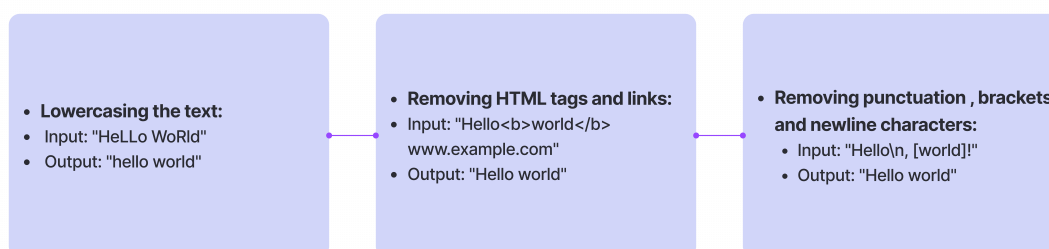
**Figure 2.9:** The Label Distribution after processing

### Handling missing values

Missing values in a dataset can cause a lot of problems during the training. One of the many techniques to fix these missing values is dropping certain rows that contain missing values which may lead us to information loss sometimes. Another technique that we used in our case is imputing technique is about to fill empty data with a certain specific value ( like the most frequent value or the average value of the column to prevent losing information. In our dataset, we used the imputing technique in order to fill the empty fields with “none” meaning no content in the current field .

### Cleaning the text

This step involves addressing various issues in the text, such as normalizing the text, eliminating punctuation, brackets, and converting the text to lowercase. Additionally, several additional preprocessing steps are performed to ensure the quality of the text.



**Figure 2.10:** The examples of the cleaning steps.

### Concatenating textual features

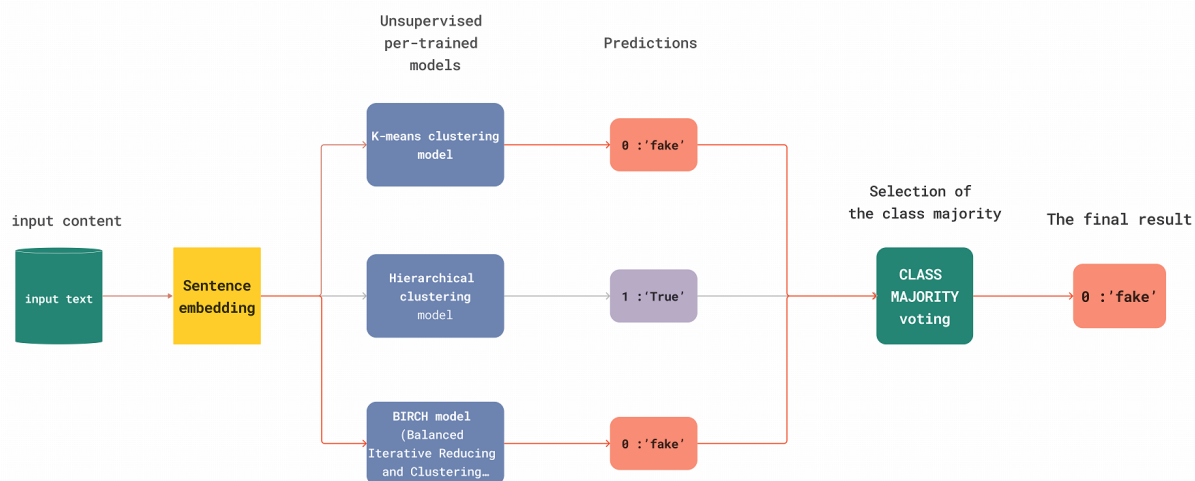
Feature Concatenation is a technique aimed at enhancing the compatibility of transformer models, such as DistilBERT, with multi-sequence inputs. In certain transformer archi-

tures, single sequences are processed instead of multiple sequences. To overcome this limitation, the feature Concatenation involves concatenating all relevant features together and treating them as a single sequence.

Rather than feeding individual features separately into the tokenizer, we combine them into one coherent input sequence. This approach enables us to pass the concatenated feature content forward to the DistilBERT tokenizer 2.3.3 as a single sequence.

## 2.3.2 Dataset labeling through unsupervised learning

. After preprocessing our dataset, which we discussed in the previous section, we now move on to an essential step in our approach: unsupervised learning. This step plays a crucial role in handling the vast amount of unlabeled datasets (unlabeled datasets in our case represents the data that haven't been fact-checked yet including blogs, speeches, podcast contents, tweets, posts, and digital articles). In our proposed approach, we employ unsupervised machine learning techniques to label these unlabeled samples. This process allows us to effectively explore a wide range of information and insightful facts to our supervised classifiers, this strategy may enhance the accuracy of our models, ultimately yielding reliable fake news classification results that closely align with real-world data.



**Figure 2.11:** Ensemble Learning Process Explained

We then used the concept of Ensemble Learning, a Machine Learning technique where multiple models are trained to solve the same problem, and the final prediction is the combination of the predictions of these multiple models to improve the efficiency of our models. Having a diversity of models is more effective than using only one, we can explore more volume of data in order to get a well trained model . In this case, we used 3 unsupervised models : K-means, Hierarchical clustering, and Birch . Their purpose is

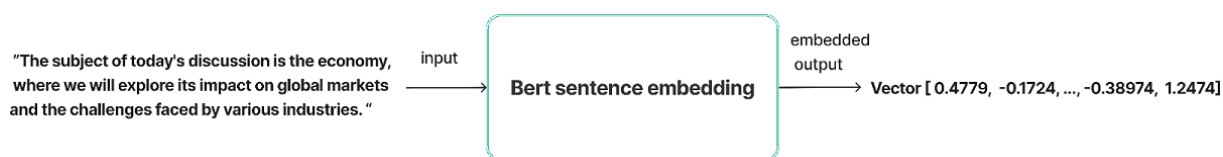
to improve the accuracy of our results and increase their robustness. It also helps reduce cases of overfitting ( when a model performs poorly on new data after learning the training data too well ) . Below is a description of the models used:

**K-means clustering :** It is one of the popular unsupervised machine learning algorithms.its function is to construct small groups (clusters) of data points that have certain similarities.where k refers to the number of clusters that we want to associate all data points we have in our dataset. K-means can only converge to a local minimum because it is a greedy algorithm.[43]

**Hierarchical clustering :** it is another popular method of grouping. It divides objects into groups so that they are similar to one another and distinct from those in other groups. A dendrogram is a type of hierarchical tree that graphically depicts clusters which makes the data easily visualized and contrary to the k-means model, the number of clusters need not be predetermined. To get the desired number of clusters, the dendrogram can instead be sliced at the proper level.[44]

**BIRCH** Balanced Iterative Reducing and Clustering using Hierarchies is a clustering model. It works by recursively applying two clustering steps: the Clustering Feature (CF) step and the Clustering Feature Tree (CFT) step. In the CF step, BIRCH constructs a tree-like data structure called the Clustering Feature Tree (CFT) by clustering the input data points using a fast, memory-efficient clustering algorithm (such as K-means) with a small number of clusters. The resulting clusters are then represented by a summary of their statistics, called the Clustering Feature (CF), which includes the centroid, the radius, and the number of points in the cluster. In the CFT step, BIRCH recursively merges the CFs of the sub-clusters into larger CFs, until a stopping criterion is met. This results in a hierarchical tree structure that represents the data clusters at different levels of granularity algorithm that is commonly used for large-scale clustering tasks. [45]

However, before we delve into unsupervised clustering, we need to transform our textual data into numerical vector representations. This is necessary because unsupervised models are unable to process textual inputs directly. To accomplish this, we utilize BERT embedding, which enables us to capture the meaning and context of the text by mapping it to high-dimensional vector spaces, as shown in the figure 2.12.



**Figure 2.12:** BERT embedding illustration

After using clustering models, we employ the majority voting. The majority voting is a simple yet effective ensemble learning method that we used to label an unlabeled dataset of news data. Each clustering model provides a binary prediction of whether a news data is fake or true. The final output is determined by selecting the prediction that appears most frequently among the three models. This approach is advantageous due to the binary nature of the predictions. As a result, we are able to convert the unlabeled dataset into a labeled dataset suitable for supervised learning.

### 2.3.3 Latent Feature Representation Generation Through Transformers

In recent years, Natural Language Processing (NLP) has made significant jump from traditional embedding techniques, including TF-IDF, that do not follow the context of a text, thanks to the development of advanced embedding methods, including deep learning embedding layers that allow for highly accurate and simplified understandable representation of input text.

**The Transformers :** Transformer models is modern type of neural network, introduced in the paper "Attention is All You Need" by Vaswani et al. [46], revolutionized NLP by employing self-attention mechanisms to capture attention on input text, Transformer models are used to solve all kinds of NLP tasks including Classifying whole sentences, text summarization, Classifying each word in a sentence, Generating text content, Extracting an answer from a text. Furthermore , These models form the foundation of technologies such as BERT and GPT-4.

Transformer models consist of two main components: the encoder and the decoder. The encoder starts with an embedding layer to convert input tokens into numerical representations. It then utilizes multiple transformer layers to capture attention and understand the context of the input text. On the other hand, the decoder generates output text based on the contextual information learned from the encoder. The figure 2.13 demonstrate the architecture behind the transformer models.

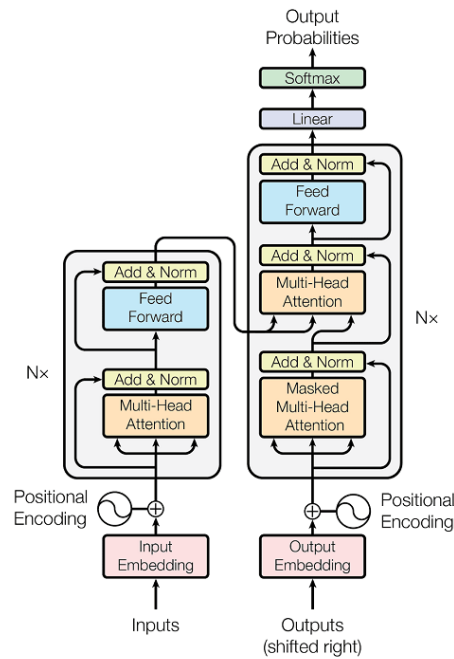


Figure 2.13: Transformer architecture [46]

During this process, we fine-tune DistilBERT, a specific variant of the transformer model, to generate high quality latent representations. The following diagram illustrates the underlying processes occurring in this step.

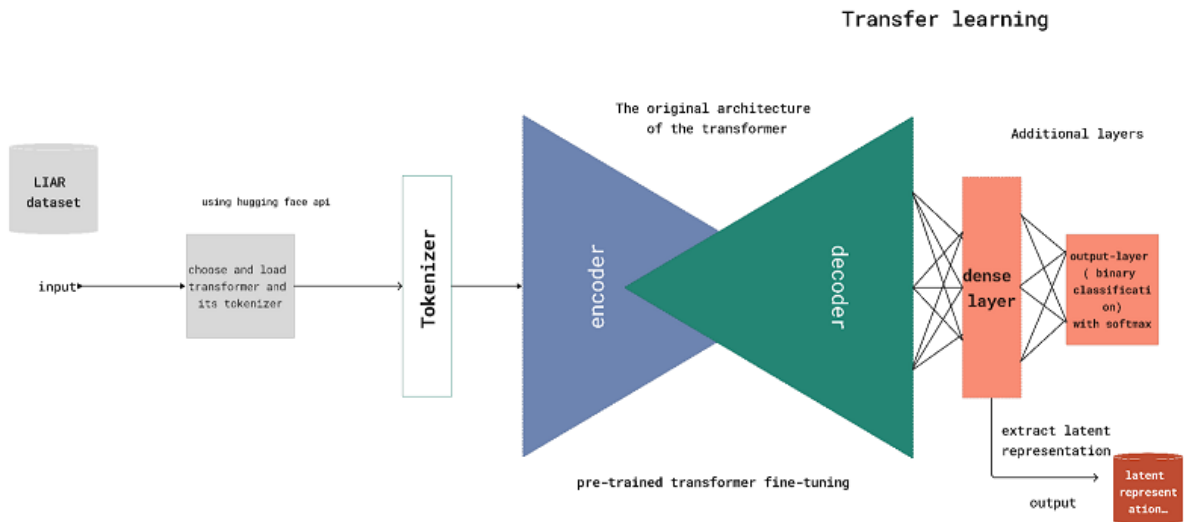


Figure 2.14: Global Flow Diagram Of The Latent Representation Step

This step follows the next scenario :

### Loading the pre-trained Distilbert model :

In order to load the model through the Hugging Face API, It is necessary to install the required libraries like Hugging Face Transformers, import the necessary modules and load the DistilBERT model using the appropriate function where we have to specify the url where the model is deployed and as well as load its tokenizer (as shown in the next figure )

```

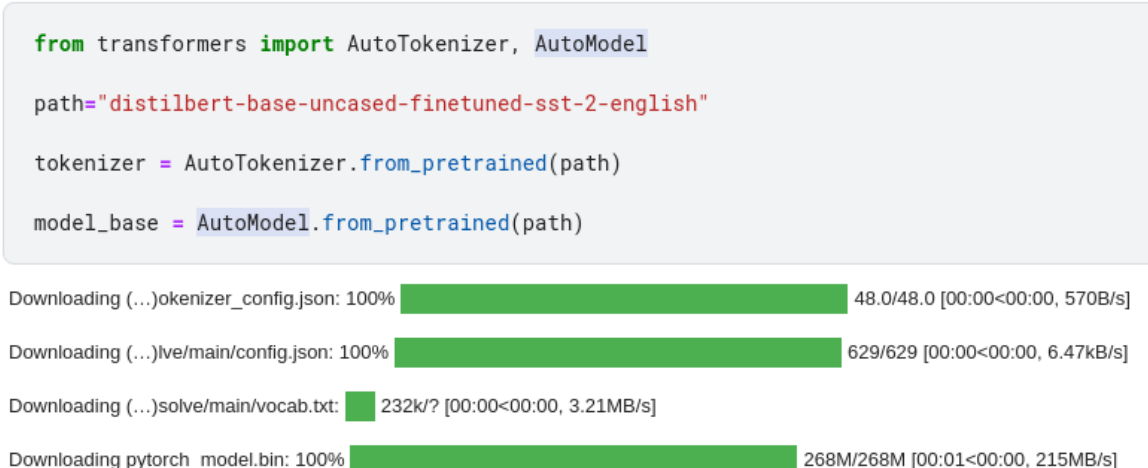
▶ from transformers import AutoTokenizer, AutoModel

path="distilbert-base-uncased-finetuned-sst-2-english"

tokenizer = AutoTokenizer.from_pretrained(path)

model_base = AutoModel.from_pretrained(path)

```



Downloading (...).tokenizer\_config.json: 100% 48.0/48.0 [00:00<00:00, 570B/s]

Downloading (...).live/main/config.json: 100% 629/629 [00:00<00:00, 6.47kB/s]

Downloading (...).solve/main/vocab.txt: 232k/? [00:00<00:00, 3.21MB/s]

Downloading (...).pytorch\_model.bin: 100% 268M/268M [00:01<00:00, 215MB/s]

**Figure 2.15:** Loading the pre-trained Distilbert

**AutoTokenizer :** The AutoTokenizer function from the Transformers library is responsible for loading the tokenizer. It is important to choose the tokenizer carefully because different transformers may have varying tokenization methods (Algorithms) or interpret tokens differently. Ensuring compatibility and understanding the tokenization process is crucial when working with transformers.

**AutoModel :** The AutoModel function from the Transformers library is used to load the appropriate model. This function is in charge of loading the specified transformer architecture based on the provided identifier name. By using AutoModel, the desired pre-trained model can be easily loaded in order to get the wanted task done including classification, regression, summarization, content generation, and question answering. Finally, there are various types of powerful and accurate transformer architectures available .

### Customization of the architecture

The abstract model needs to be adapted to our problem. For this purpose we removed the softmax layer and added two additional layers to the model architecture. The first

layer was a dense layer with 768 nodes (hidden layer) and a ReLU activation function, which allowed us to extract latent representations from the model. The second layer was a basic output layer with softmax activation for binary classification, which facilitated binary classification for this task.

### **Tokenization of the text using the “ Distilbert” tokenizer :**

The first necessary operation before passing data to the transformer is tokenization. Transformers can only process numbers, so tokenization is the process of converting text data into tokens that the models can understand. In this case, we used the "DistilBERT" tokenizer. There are three common types of tokenization used by transformers: word-based, character-based, and subword-based. The choice of tokenizer and tokenization type depends on the specific family of transformers you selected [47].

The DistilBERT tokenizer utilizes the subword-based tokenization method, which offers a speed advantage of up to 4 times faster compared to the classic tokenizer that uses a word-based tokenization method. The subword-based method breaks words into smaller units, enabling quicker processing by the model. Despite having a smaller vocabulary of 3,000 tokens compared to the classic tokenizer's 30,000 tokens, DistilBERT maintains a high degree of accuracy.

However, When passing the text input through the DistilBERT tokenizer, it generates two main tensors: **input\_ids** and **attention mask**.

**The input\_ids** is the result vector containing tokens where each word or character or subword gets its own token much like a dictionary. as we mentioned that the models can only process numbers. The vector also contains special tokens such as [CLS] and [SEP] to the input text. The [CLS] token is added at the beginning of the input text to indicate the start of the sentence; this token has 101 as value , while the [SEP] token is added after each sentence to indicate the end of the sentence the [SEP]occupied 102 token. These special tokens are important for the transformer model to understand the structure of the input text and to make accurate predictions.

**The attention mask** is another important component of the input data that is generated during the tokenization process. The attention mask is a binary vector that has the same length as the input sequence, and it is used to indicate which tokens the model should pay attention to during training and inference. It has a value of 1 for all valid

tokens, and a value of 0 for padding tokens that are added to the input sequence to ensure that all inputs have the same length. In figure 2.16 shows how our data looks after tokenization as we can see that the word ‘subject’ occupies token 2548 with value of 1 in attention\_mask the tokenizer in this case wants to push the model to pay attention to the word ‘subject’.



**Figure 2.16:** Example of word tokenization through Distelbert tokenizer

In the provided figure (the screenshot) 2.17, it can be observed that every input’s IDs tensor ( In input\_ids column) begins with the token "101," which precisely corresponds to the "[CLS]" token indicating the start of the sequence, as mentioned earlier in 2.3.3. Additionally, the third token, "2548," represents the word "subject" in the DisltetBert tokens dictionary. Furthermore, all these input tokens have a corresponding attention value of 1, indicating the importance of these tokens and directing the transformer to focus its attention on them.

	label	statement	input_ids	attention_mask
0	0	the subject is economy jobs in context of an ...	[101, 1103, 2548, 1110, 4190, 5448, 1107, 5618...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
1	0	the subject is obama birth certificate religio...	[101, 1103, 2548, 1110, 184, 2822, 1918, 3485,...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
2	0	the subject is campaign finance congress taxes...	[101, 1103, 2548, 1110, 2322, 7845, 16821, 753...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
3	0	the subject is poverty in context of an opini...	[101, 1103, 2548, 1110, 5224, 1107, 5618, 1104...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
4	0	the subject is economy stimulus in context of...	[101, 1103, 2548, 1110, 4190, 21950, 1107, 561...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

**Figure 2.17:** A screenshot of the dataset after tokenization

## Fine-tuning of the model

We fine-tuned a pre-trained DistilBERT model for fake news classification using Trainer API. where we included as arguments a learning rate of 0.0005 that is commonly used to minimize significant changes in neural weights. The training process often utilizes the Wadam optimizer. The accuracy was also used as the evaluation metric.



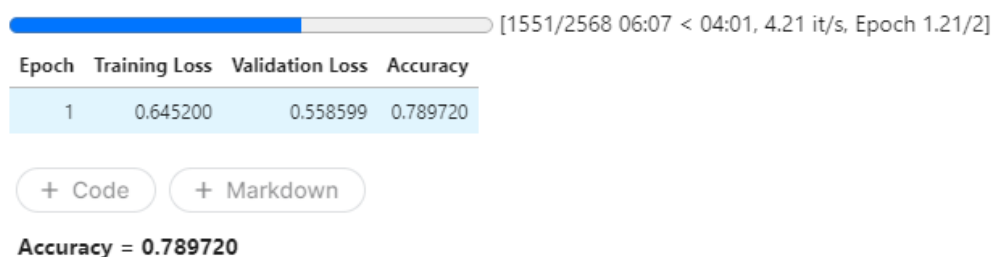


Figure 2.18: Fine-tuning of pre-trained DistilBERT (in the first epoch)

### The latent representation generator:

After fine-tuning the model, we extracted 768-dimensional latent representations for each sequence in our dataset. These representations capture the essential features of the input text and are represented as vectors of real numbers. We utilized these representations as input features for a machine learning model and observed significant improvements in model performance compared to using raw text as input.



Figure 2.19: Latent representation generator

### Final results

Screenshot of outputs :

	label	statement	input_ids	attention_mask	latent_representation
0	0	the subject is economy jobs in context of an ...	[101, 1103, 2548, 1110, 4190, 5448, 1107, 5618...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	[0.36940157, 0.2552772, -0.32459587, -0.055277...
1	0	the subject is obama birth certificate religio...	[101, 1103, 2548, 1110, 184, 2822, 1918, 3485,...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	[0.25709307, 0.39709428, -0.51084894, -0.19059...
2	0	the subject is campaign finance congress taxes...	[101, 1103, 2548, 1110, 2322, 7845, 16821, 753...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	[0.39375246, 0.38134414, -0.47953147, 0.047416...
3	0	the subject is poverty in context of an opini...	[101, 1103, 2548, 1110, 5224, 1107, 5618, 1104...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	[0.30521804, 0.22478615, -0.41540286, -0.04488...
4	0	the subject is economy stimulus in context of...	[101, 1103, 2548, 1110, 4190, 21950, 1107, 561...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	[0.3879261, 0.3458842, -0.37248784, -0.1365771...

Figure 2.20: Screenshot of the final produced dataset

```

{'latent_representation': array([ 0.3879261 ,  0.3458842 , -0.37248784, -0.13657716, -0.06639151,
-0.50402826, -0.47815233, -0.31528974,  0.4307381 ,  1.2475219 ,
-0.769947 ,  0.7228769 , -0.7334329 , -0.07906495, -0.31069207,
-0.2504628 ,  0.0198636 , -0.5315574 ,  0.49996462, -0.45215994,
-0.27866977,  1.1855268 ,  0.30903217,  0.36355123, -0.17821775,
 0.17864136,  0.2970624 , -0.99544644,  0.8274436 , -0.1557886 ,
-0.2722015 ,  0.21485184,  1.1110883 ,  0.51310414, -0.5224704 ,
-0.05629871,  0.3617068 ,  0.83314836, -0.70966303, -0.35520723,
-0.6196218 ,  0.3884104 , -0.41913176, -0.12306967, -0.22405729,
-0.336714 ,  0.95305556,  0.54707503,  0.05042239, -0.07632723,
-0.51014817,  0.24371302,  0.30325544,  0.4439012 , -0.00647127,
 0.07159653,  0.04572621, -0.08729737,  0.2738799 ,  0.05527453,
-0.5145623 ,  0.03650241,  0.6707123 , -0.11713348,  0.29214016,
 0.48410377, -0.25661224,  0.4126587 ,  0.5900302 ,  0.15876421,
 1.0151263 , -0.12077179,  0.01276239,  1.0704573 ,  0.560224 ,
-0.22808394,  0.53465104,  0.31722802, -0.94434404, -0.83426636,
-0.39060515, -0.4464047 , -0.338332 ,  0.17083326,  0.36845896,
 0.42617485, -0.41777402,  0.33031437, -0.1549222 ,  0.602319 ,
 0.25710946, -0.36456427,  1.1362851 , -1.0503888 , -0.12944856,
-0.38222238, -0.5549382 ,  0.14237186, -0.35405442, -0.23894693,
 0.05226484,  0.02629251, -0.6876284 , -0.31795925,  0.2827023 ,
 0.30881768, -0.61507994,  0.28137806,  0.33414128, -0.2583186 ,
 0.81685823,  0.00386161, -0.36729974, -0.2328658 , -0.5672266 ,
-0.18214789,  0.21013725, -0.28265533, -0.66274625, -0.04557605,
-0.23216079,  0.0874432 , -0.03736439, -0.12187748,  0.18230072,
-0.18294479,  0.61195904,  0.09491937]), dtype=float32),
'text': 'the subject is economy stimulus in context of interview with cbs '
'news the speaker is job president with state illinois with '
'affiliation democrat on attacks by republicans that various '
'programs in the economic stimulus plan are not stimulative if you '
'add all that stuff up it accounts for less than percent of the '
'overall package '}

```

**Figure 2.21:** A full zooming under one text instance of dataset after the latent Representation generation

### 2.3.4 Classifiers Training

Now we have reached the final step in our proposed approach for fake news classification. Our journey involved utilizing a labeled dataset obtained through an unsupervised step. To enhance our supervised tasks, we employ the latent representation of these labeled dataset features generated by DistilBERT. This combination of labeled data and DistilBERT representation will aid us in accurately classifying fake news. The figure below provides a visual representation of our approach in the last step.

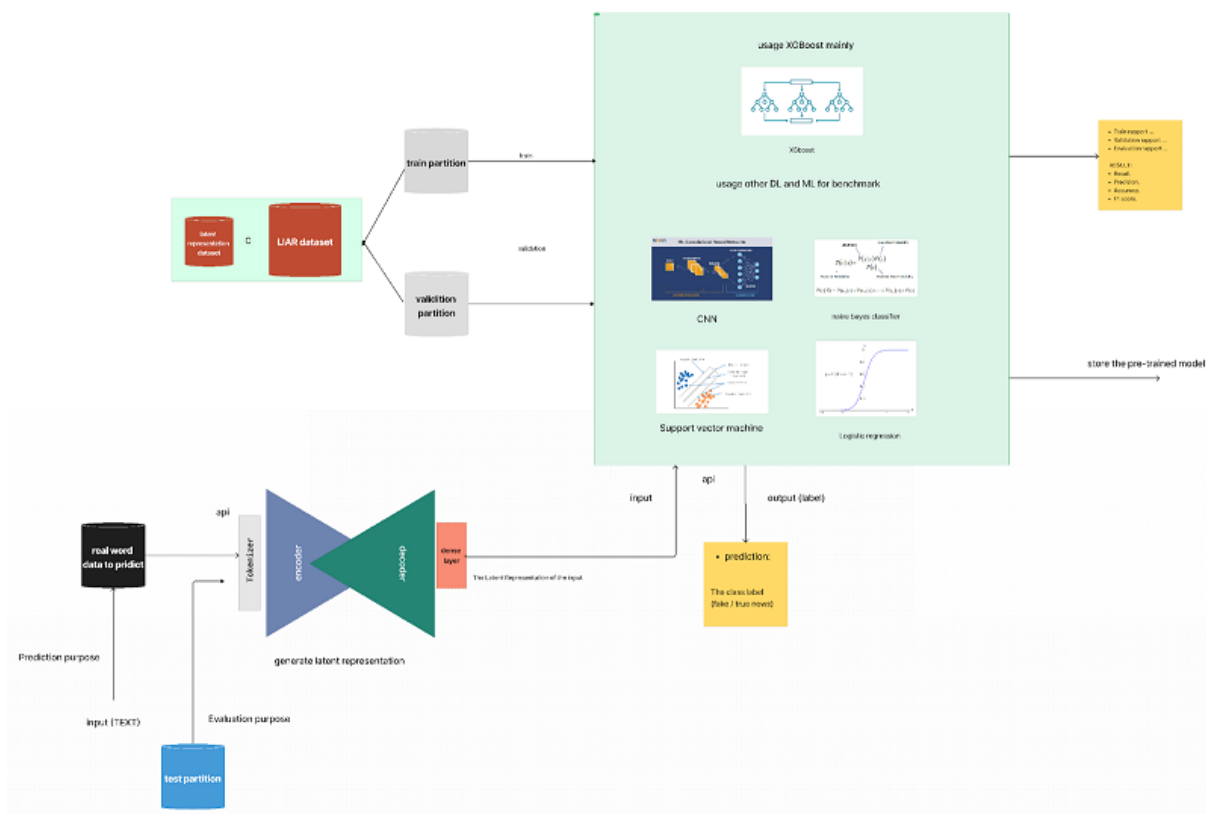


Figure 2.22: The global flow diagram of the final classification step

### 2.3.4.1 Supervised learning :

Supervised learning deals with labeled data, so we identify the input class, and push the model to learn ,extract patterns from the dataset, when the model is highly accurate it will be able to predict the output dataset given the input dataset . According to Andrew W. Trask : “Supervised machine learning is the direct imitation of a pattern between two datasets. It’s always attempting to take an input dataset and transform it into an output dataset” [48]. As we can see, the goal of supervised learning is to establish a strong correlation between the input and output datasets.



Figure 2.23: Supervised learning explained [48]

In this section, we perform supervised tasks of fake news classification or detection. Our approach involves leveraging machine models. These models include logistic regres-

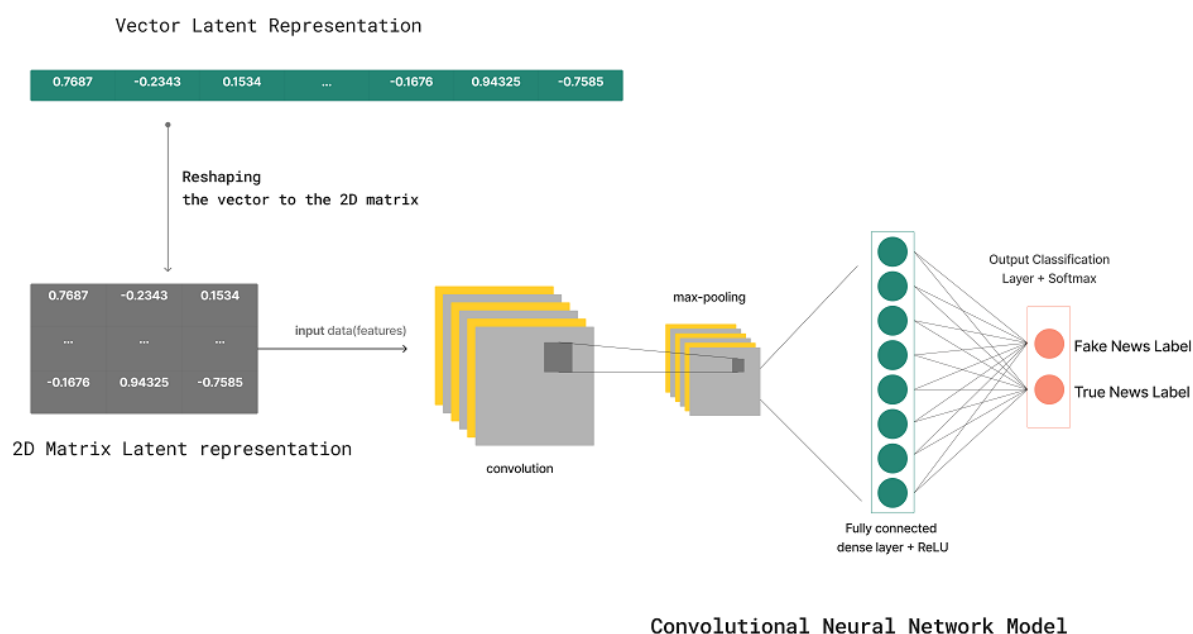
sion, XGBoost, Naive Bayes (Gaussian), Support Vector Machines (SVM), and as well as Convolutional Neural Networks (CNN) in order to benefit from deep learning power .

These specific models effectively capture the essential features of the input text and enable us to accurately classify fake news, contributing to the ongoing fight against misinformation.

### Classifiers:

**Convolutional Neural Networks (CNNs) :** CNNs have demonstrated remarkable performance in various domains, including image processing and natural language processing (NLP). In the context of NLP tasks, we did use it to classify news credibility [14] .

Before passing the latent representation obtained from Distelbert through a Convolutional Neural Network (CNN) model, a crucial step is to reshape the latent representation from a vector to a matrix. This process is necessary to properly process the data within the CNN architecture, which typically operates on multi-dimensional tensors. commonly used in image tasks. By reshaping the latent representation into a matrix as shown in figure 2.24, the CNN will be able to treat the text as an image input, enabling it to capture spatial relationships and extract meaningful features. This transformation enhances CNN’s effectiveness in fake news classification.



**Figure 2.24:** Vector to Matrix Reshaping to news class classification throught CNN model

**XGBoost :** XGBoost stands for extreme gradient boosting,XGBoost is a powerful library for ensemble learning in machine learning. It implements gradient boosting

with various additional features, such as parallel tree boosting, aimed at improving performance and speed in data science solutions. By combining multiple decision trees, XGBoost iteratively minimizes the loss or error by adding more decisions to address the previous errors of the previous iteration. This iterative approach allows XGBoost to create highly accurate models for a wide range of machine learning tasks.[49]

**Logistic Regression :** The classifier models the class probabilities as a function of the linear combination of predictors. Logistic regression utilizes a typical linear regression formulation. Logistic regression (LR) is a significant statistical and data mining technique widely utilized by statisticians and researchers to analyze and classify datasets with binary and proportional responses. LR offers several notable advantages. Firstly, it naturally provides probabilities, making it useful for assessing the likelihood of different outcomes. Additionally, LR can be extended to handle multi-class classification problems. Another advantage is that the principles used in linear regression can be applied to LR, allowing for consistent analytical approaches. Furthermore, LR can be employed with a wide range of unconstrained optimization techniques. [50]

**Naive Bayes (Gaussien)** Gaussian naive Bayes classification represents a notable instance within the family of naive Bayes methods, which are widely recognized as generative model-based classifiers. It exhibits a rapid learning and testing process, leveraging the underlying assumption of a Gaussian distribution concerning attribute values, given the corresponding class label. This assumption proves instrumental in facilitating efficient classification outcomes and reinforcing the interpretability of the model's results within various domains of study.[51]

**Support machine vector ( SVM ) :** This method utilizes a hyperplane to separate data, aiming to maximize the margin and achieve a greater distance between the hyperplane and the values on each side. The larger the distance, the more effectively the expected generalization error is reduced. SVM excels in handling a large number of features as they only include the features lying on the margin of the hyperplane. In situations where the data is non-linear and cannot be separated by a single hyperplane, SVM can generate multiple hyperplanes in a higher-dimensional feature space. SVM methods are primarily binary. Therefore, in this study comparing the patient group with and without coercion, no dummy variables were required for the response feature." [52]

## 2.4 Evaluations and results

### 2.4.1 Implementation tool

Our proposed system architecture was implemented using the "Pandas" , "NumPy", "Scikit learn", "Transformer", "Torch" and "Keras Tensorflow" libraries of Python, our programming language.

#### 2.4.1.1 Work environment

Our work environment was The Kaggle Notebook on the Kaggle Website. Kaggle Notebook is a web-based computational environment for data analysis, machine learning, and deep learning tasks that enables users to work together to collaborate on a joint project. It includes well-known data science libraries as pre-installed dependencies. Individual code cells can be used by users to write interactively running code. It is overall a comprehensive and user-friendly environment for data analysis, model development, and knowledge sharing.

### 2.4.2 Measuring the Performance

In order to assess the performance of our binary classification models for fake news detection, we employ various evaluation metrics. These metrics provide valuable insights into the accuracy, precision, recall, and overall effectiveness of the models in distinguishing between fake and real news . The key evaluation metrics used, along with their formulas, are described below:

#### **Confusion Matrix:**

The confusion matrix, also known as the error matrix as shown in figure 2.25, is a fundamental performance measure for classification tasks. It provides a detailed breakdown of the model's predictions and their correspondence with the actual labels. The confusion matrix consists of four important values:

	Predicted <b>0</b>	Predicted <b>1</b>
Actual <b>0</b>	TN	FP
Actual <b>1</b>	FN	TP

**Figure 2.25:** The Confusion Matrix Table [53]

- **True Positive (TP):** The number of instances correctly predicted as fake news.
- **True Negative (TN):** The number of instances correctly predicted as real news.
- **False Positive (FP):** The number of instances incorrectly predicted as fake news when they are actually real news.
- **False Negative (FN):** The number of instances incorrectly predicted as real news when they are actually fake news.

### Precision

Precision measures the accuracy of the positive predictions made by the model. It calculates the ratio of true positives to the sum of true positives and false positives.[54]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

### Recall

Recall, also known as sensitivity or true positive rate, quantifies the model's ability to correctly identify positive instances from the actual positives. It calculates the ratio of true positives to the sum of true positives and false negatives.[54]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

### Accuracy

Accuracy represents the overall correctness of the model's predictions. It is calculated as the ratio of the correctly predicted true positives and true negatives to the total number

of positive and negative observations.[54]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

### F1 Score

The F1 score combines precision and recall into a single metric, providing a balanced evaluation of the model's performance. It is calculated as the harmonic mean of precision and recall, giving equal weight to both metrics.[54]

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

## 2.4.3 Results

In this section, we present the results of our experiments conducted on two datasets: the LIAR and the ISOT Fake News Dataset. We employed two distinct approaches for fake news detection, accompanied by a small benchmark. The first approach involved supervised learning, training the classifiers using the real labels provided with each dataset. In the second approach, we explored the potential of unsupervised learning to label the dataset anew, which was then passed through the first approach using these updated labels. This allowed us to simulate the analysis of unlabeled content and calculate the corresponding scores.

We will first outline the two experiments and their respective scores, followed by an interpretation of the results. then, we present our conclusions and we end within a detailed discussion.

### First Experiment : using LIAR DATASET

The first experiment used supervised learning on the LIAR dataset discussed in the Dataset Exploratory Analysis section 2.2. The classifier was trained on a subset of the dataset containing 11,284 articles. It was then used to predict the labels of the remaining 1,283 articles.

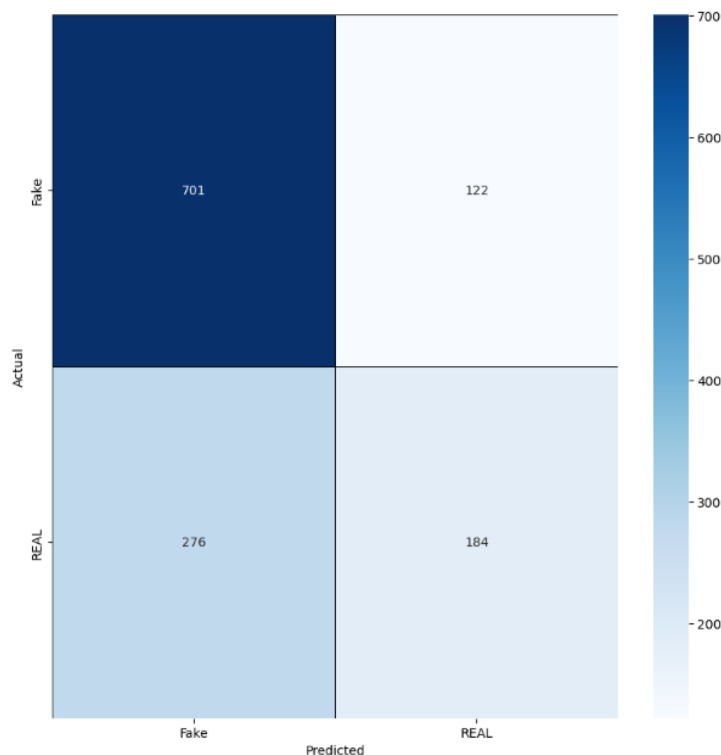
**SUPERVISED result :** The DistilBERT classifier achieved an accuracy of 69%, a precision of 68%, and a recall of 69% in the supervised experiment of the LIAR dataset. Figure 2.26 shows the results of the DistilBERT classifier. Figure 2.27 shows the results of the ML models, and Figure 2.29a shows the results of the CNN model.



The DistilBERT classifier, in the supervised experiment on the LIAR dataset, achieved an accuracy of 69%, with a precision of 68% and a recall of 69%, as demonstrated in Figure 2.26. The classification report provides detailed insights into the model's performance, showing precision, recall, and F1-scores for both classes (fake news and real news). Upon analyzing the confusion matrix, we observe that the DistilBERT classifier correctly predicted 701 instances of fake news (True Negatives, TN), while incorrectly classifying 276 instances of fake news as real news (False Negatives, FN). On the other hand, it correctly predicted 184 instances of real news (True Positives, TP) and misclassified 122 instances of real news as fake news (False Positives, FP). Overall, the classifier demonstrated relatively balanced performance, with higher precision and recall for the fake news class compared to the real news class. This suggests that the model was more proficient in identifying fake news instances accurately. However, there is still room for improvement in accurately predicting real news

Classification Report:				
	precision	recall	f1-score	support
0	0.72	0.85	0.78	823
1	0.60	0.40	0.48	460
accuracy			0.69	1283
macro avg	0.66	0.63	0.63	1283
weighted avg	0.68	0.69	0.67	1283

(a) The score

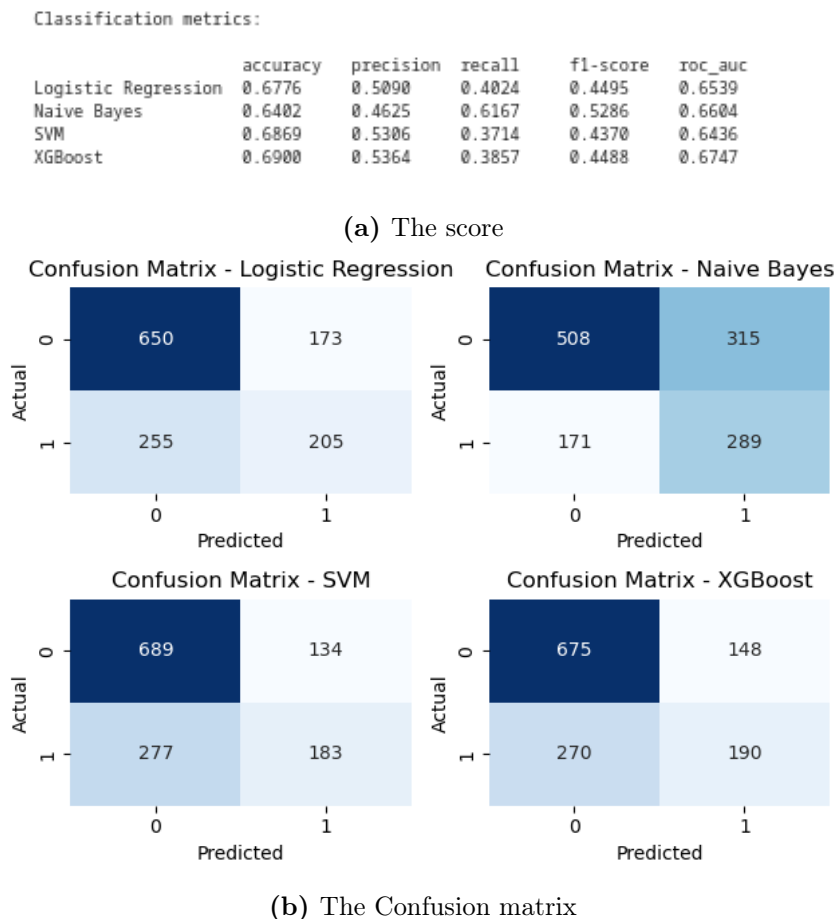


(b) The Confusion matrix

**Figure 2.26:** The evaluation of the Distelbert model on the LIAR dataset using the real labels of the LIAR dataset.

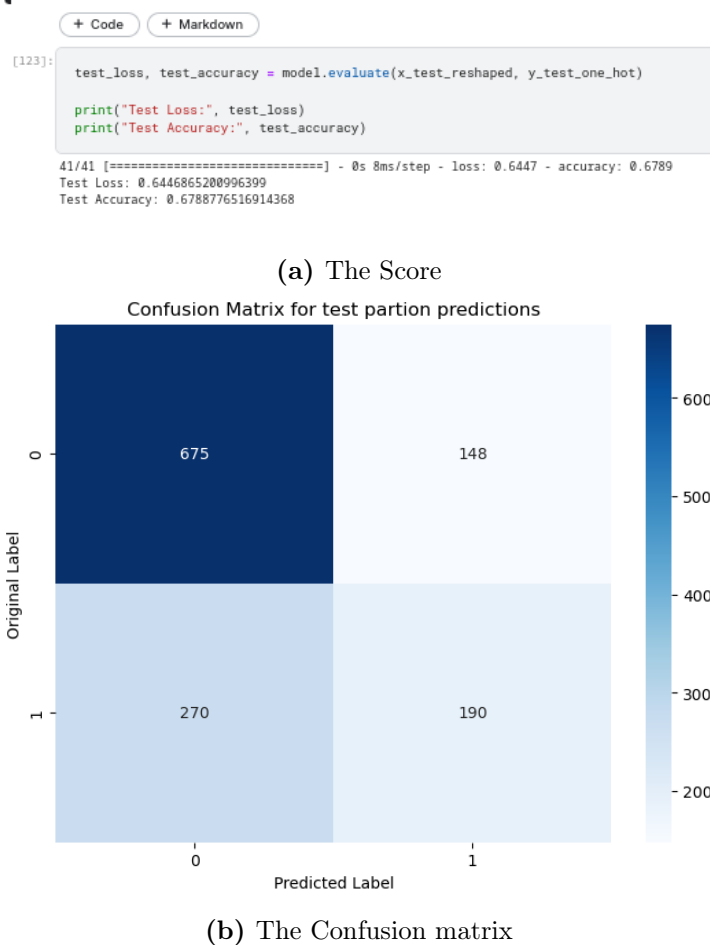
Figure 2.27 presents the results of various ML models used in the supervised experiment. These models include Logistic Regression, Naive Bayes, SVM, and XGBoost. The performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, are shown for each model. Comparing the results, XGBoost achieved the highest accuracy of 69%, followed closely by SVM at 68.69%, Logistic Regression at 67.76%, and Naive Bayes at 64.02%. Looking into the confusion matrices for each model, we observe varying degrees of correct and incorrect predictions for fake news and real news instances. The models demonstrated differing capabilities in distinguishing between the two classes. For

instance, Logistic Regression achieved 650 True Negatives (correctly predicted fake news) and 205 True Positives (correctly predicted real news) but struggled with 173 False Positives (incorrectly predicted real news) and 255 False Negatives (incorrectly predicted fake news).



**Figure 2.27:** The evaluation of the ML models on the LIAR dataset using the real labels of the LIAR dataset.

Figure 2.29a represents the result of the Convolutional Neural Network (CNN) model in the supervised experiment. The model achieved a test accuracy of 67.89%. The confusion matrix for the CNN model reveals that it correctly predicted 675 instances of fake news (True Negatives, TN) and 190 instances of real news (True Positives, TP). However, it incorrectly classified 270 instances of fake news as real news (False Negatives, FN) and 148 instances of real news as fake news (False Positives, FP). Overall, the CNN model displayed competitive performance, but it encountered challenges in accurately classifying both fake and real news instances.



**Figure 2.28:** The evaluation of the CNN model on the LIAR dataset using the real labels of the LIAR dataset.

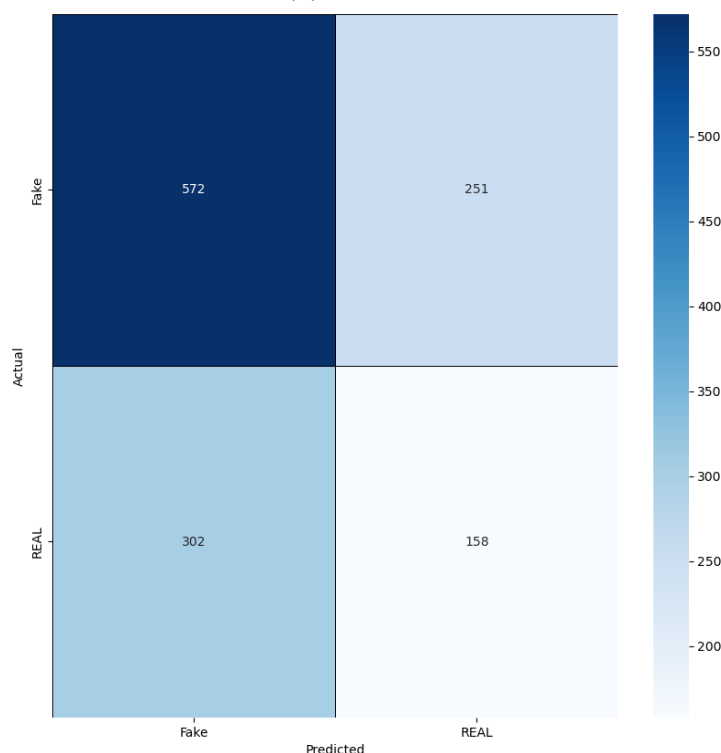
Overall, our supervised experiment showcased the effectiveness of the DistilBERT classifier in identifying fake news instances on the LIAR dataset. Additionally, we explored the performance of various ML models and a CNN model. While each model displayed varying degrees of accuracy and proficiency in distinguishing between fake and real news.

**The results of unsupervised labeling followed by supervised Classifiers training**

: As shown in Figure 2.29. The Distil-BERT classifier achieved accuracy, precision, and recall scores of 57%, 56%, and 57%, respectively, on the LIAR dataset. Thus, the classifier correctly identified 572 samples as fake news, 158 samples as real news, 251 samples as real news that were actually fake news, and 302 samples as fake news that were actually real news.

Classification Report:				
	precision	recall	f1-score	support
0	0.65	0.70	0.67	823
1	0.39	0.34	0.36	460
accuracy			0.57	1283
macro avg	0.52	0.52	0.52	1283
weighted avg	0.56	0.57	0.56	1283

(a) The Score



(b) The Confusion matrix

**Figure 2.29:** The evaluation of the DistelBERT model on the LIAR dataset using unsupervised labeling.

**The first experiment summary:** Table 2.1 provides a comprehensive summary of the first experiment, comparing the performance of our model with the state-of-the-art results on the LIAR Dataset. We can see that Our results are below some of those found in the literature using the LIAR dataset.

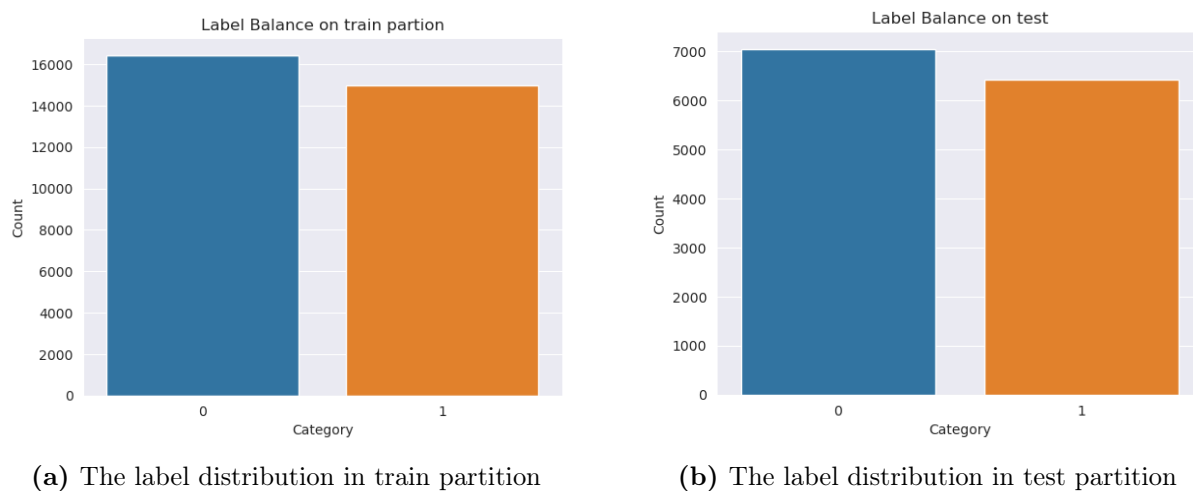
Reference	Classification model	Accuracy	Precision	Recall	F1-score
Yang et al.2019 [32]	Collapsed Gibbs Sampling	75.9%	76.6 % / 75%	77.4 % / 74.1%	78.3% / 73.2%
Truong et al. 2020 [55]	Bi-LSTM	61%	60%	61	-
Braşoveanu and Andonie 2020 [56]	CapsNET	64.9%	-	-	-
Aslam and Al 2021 [57]	Deep learning ( ensemble )	89.8%	91.3%	91.6%	91.4%
Our model (Supervised only)	DistilBert fine tuned	69%	67%	69%	67%
Our proposed model Supervised + unsupervised	Ensemble learning + DistilBert fine tuned	57%	56%	57%	56%

**Table 2.1:** Comparison of the proposed methods with the state of the arts on the LIAR dataset.

The table 2.1 shows the performance of our proposed approach for fake news detection on the LIAR dataset. Our supervised model achieved an accuracy of 69%, which is lower than some of the state-of-the-art results. However, our unsupervised model achieved an accuracy of 57%, which means that the supervised model achieved a higher accuracy than the unsupervised model, as expected. This is because the supervised model was trained on a dataset with its real label. However, To determine whether our approach was successful or not, we tried other datasets, such as the ISOT Fake News Dataset. The ISOT dataset is larger, and well-balanced. Furthermore, it contains binary classes. we anticipate gaining valuable interpretations of its results in the next experiment.

### ISOT Fake News Dataset

Due to the relatively average results obtained with the LIAR dataset, we decided to use another dataset to make our model more efficient. The ISOT Fake News Dataset was chosen to be the second dataset for our work. Similarly to the previous section, we will begin by showcasing the results, confusion matrices, for the supervised and unsupervised strategies .Finally, a comparison of our findings with the state-of-the-art. This current dataset seems pretty balanced. As depicted in the figure 2.30, the current distribution exhibits a balanced representation, where the Category-axis represents the label (0 for fake and 1 for true), and the Count-axis represents the count values for each label.

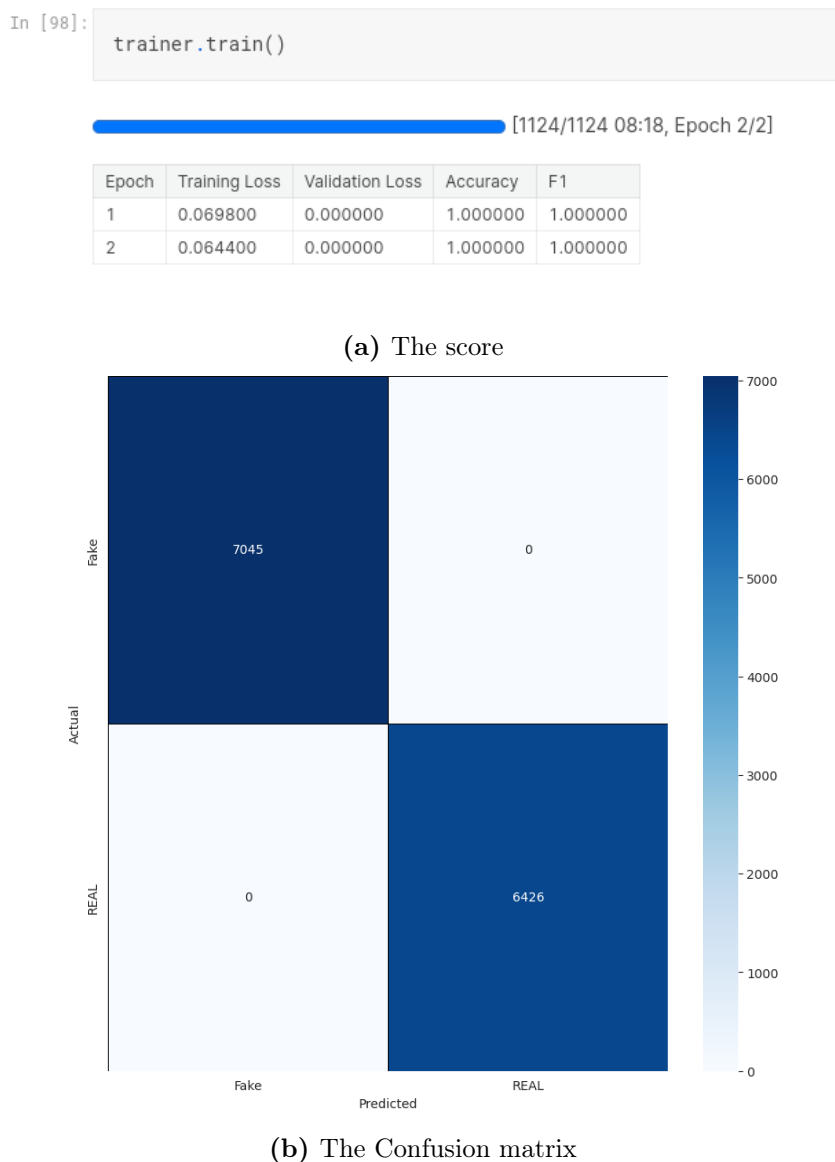


**Figure 2.30:** The label distribution in ISOT Fake News Dataset

**SUPERVISED results :** We get impressive results through the DistilBERT classifier, where the accuracy, precision, and recall scores for the supervised experiment of the ISOT Fake News Dataset were 100%, 100%, and 100%, respectively. As shown in Figure 2.31, Figure 2.32b illustrates the evaluation results of ML models, and Figure 2.33 represents the evaluation result of the CNN model.

As shown in figure 2.31 ,The DistilBERT model achieved a perfect score of 100% accuracy, precision, and recall on the ISOT Fake News dataset. This means that the model correctly classified all 7045 true negatives, 6426 true positives, and 0 false positives and false negatives.

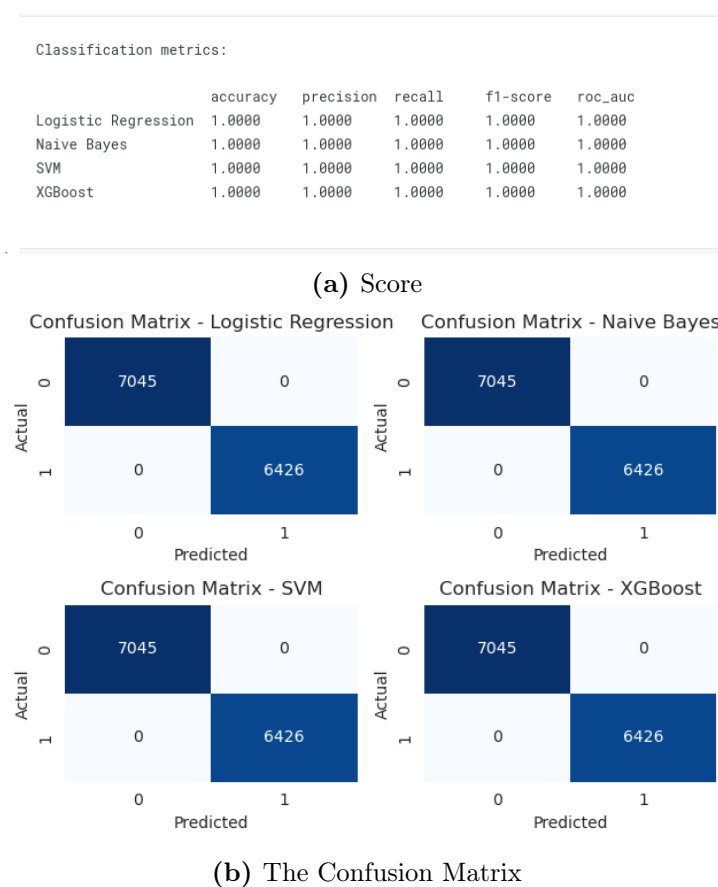
The training loss of 0.0644 and validation loss of 0.000 suggest that the model was well-trained and did not overfit to the training data.



**Figure 2.31:** The evaluation of the Distilbert model on ISOT Fake News dataset using the real labels of the dataset.

Furthermore, as shown in figure 2.32b Machine learning models including logistic regression, support vector machines, XGBoost, and naive Bayes were able to achieve the same accuracy of 100% as the DistilBERT model on the ISOT Fake News dataset. They also had the same confusion matrix results, with 0 false positives and false negatives.



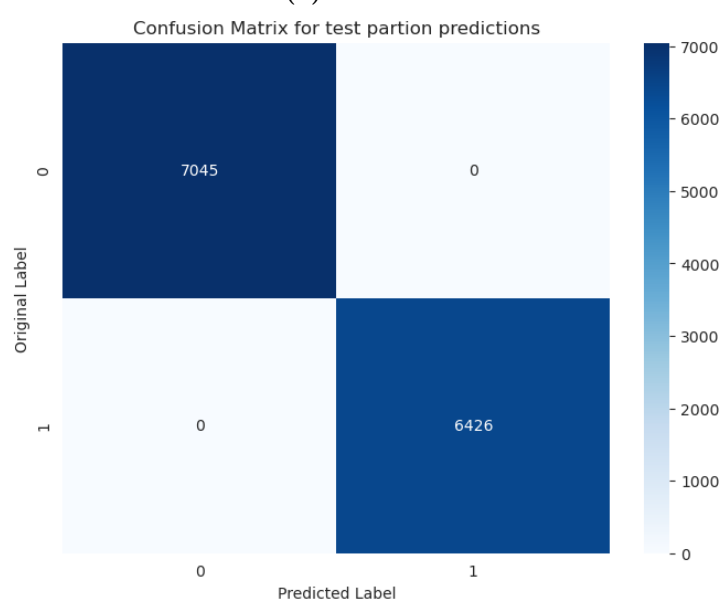


**Figure 2.32:** The evaluation of the ML models on ISOT Fake News dataset using the real labels of the dataset.

The CNN model achieved identical results to DistilBERT on the ISOT Fake News dataset, with an accuracy of 100%, precision of 100%, and recall of 100%. The training loss was 0.0000 and the validation loss was 0.0002 after 8 epochs. This means that the model correctly classified all 7045 true negatives, 6426 true positives, and 0 false positives and false negatives. Thus, the CNN model was able to achieve the same results with less training loss compared with Distil-BERT.



(a) The Score



(b) The Confusion Matrix

**Figure 2.33:** The evaluation of the CNN model on ISOT Fake News dataset using the real labels of the dataset.

**Unsupervised labeling followed by Supervised Classifiers training :** The Distelbert classifier achieved an accuracy of 91%, a precision of 92%, and a recall of 91% on the ISOT Fake News dataset. These results are significantly better than the results on the LIAR dataset, and suggest that unsupervised labeling followed by supervised classifier training is a promising approach for fake news detection. (see the Figure 2.34)

The classifier correctly classified 6282 true negatives, 6025 true positives, 325 false positives, and 839 false negatives.

- True Negatives (TN): 6282 samples that were correctly classified as fake news.
- True Positives (TP): 6025 samples that were correctly classified as real news.

- False Positives (FP): 325 samples that were incorrectly classified as fake news.
- False Negatives (FN): 839 samples that were incorrectly classified as real news.

The classifier only missed classifying 839 real news samples as fake news, which is a much lower number than the number of fake news samples missed by the Distil-BERT classifier on the LIAR dataset. This suggests that unsupervised labeling followed by supervised classifier training is a more effective approach for fake news detection when the amount of labeled data is limited.

---

```

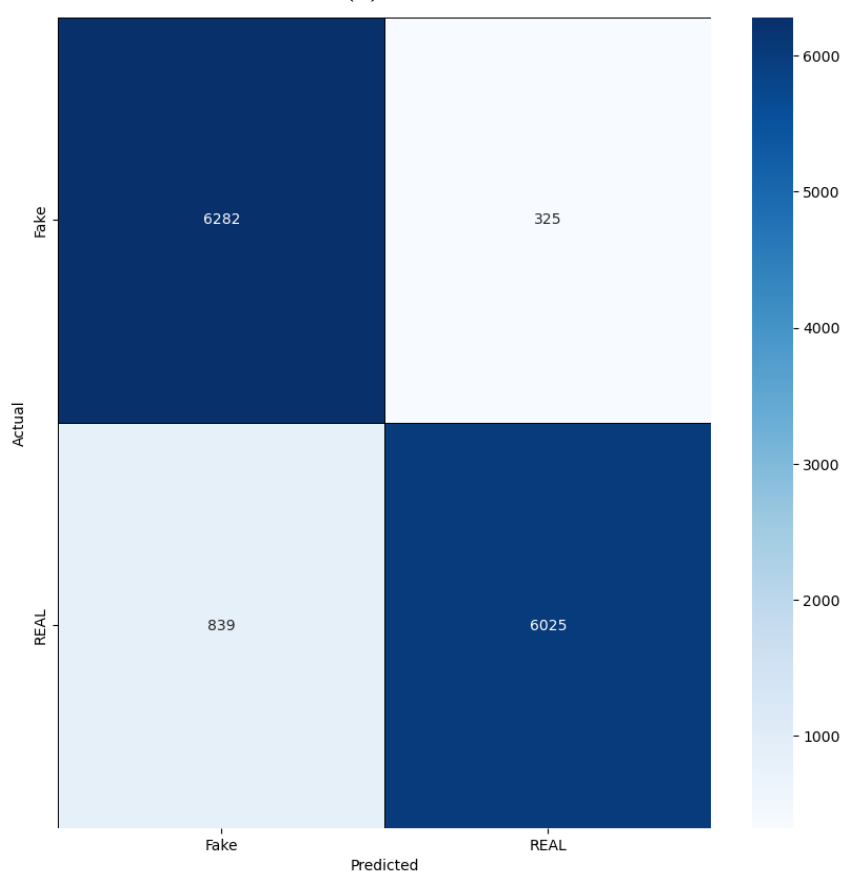
Classification Report:
              precision    recall  f1-score   support

     0       0.88         0.95         0.92         6607
     1       0.95         0.88         0.91         6864

 accuracy              0.91         13471
 macro avg              0.92         0.91         0.91         13471
 weighted avg          0.92         0.91         0.91         13471

```

(a) The Score



(b) The Confusion matrix

**Figure 2.34:** The evaluation of the DistelBERT model on ISOT Fake News dataset after unsupervised labeling.

**The second experiment summary:** We present a comparative analysis of our model with other state-of-the-art approaches using the ISOT Fake News Dataset in the table 2.2. We can see that our proposed approach gave us impressive results with an f1-score of 100% using only the supervised part and a total f1-score of 91%, using our unsupervised labeling in addition of our supervised training, which is comparable to the results of the

state of arts on ISOT Fake News Dataset.

Author	Classification model	Accuracy	Precision	Recall	F1-score
Ahmed et al. [15]	Linear SVM	92%	-	-	-
Ahmed et al. [58]	Random forest	99%	-	-	-
Kadek Sastrawan [59]	Bidirectional LSTM	99.95%	99.95%	99.95%	99.95%
Our model (Supervised only)	DistilBert fine tuned	100%	100%	100%	100%
Our proposed model Supervised + unsupervised	Ensemble learning + DistilBert fine tuned	91%	92%	91%	91%

**Table 2.2:** Comparison of the proposed methods with the state of the arts on the ISOT Fake News Dataset dataset.

The provided table 2.2 presents a comprehensive comparison of the evaluation results of our proposed methods with those of state-of-the-art approaches on the ISOT Fake News dataset.

We conducted an evaluation of our proposed fake news detection approach using two solutions: a supervised model (Distil-Bert fine-tuned) and an unsupervised approach, followed by classifiers training including Distil-Bert fine-tuned.

The supervised model yielded remarkable results, achieving an outstanding accuracy of 100% on the ISOT Fake News dataset, which is significantly higher than the accuracy of the state-of-the-art methods. which range from 92% to 99.95%.

On the other hand, the unsupervised approach, combined with classifiers training including Distil-Bert fine-tuned. We achieved an accuracy of 91%, which is still a very high accuracy. This demonstrates the potential of our proposed approach to simulate content labeling.

#### 2.4.4 Discussion

Our study aimed to develop an effective approach for fake news detection by employing two distinct strategies on two different datasets, namely the LIAR dataset and the ISOT Fake News Dataset. The first approach involved supervised learning, where we trained classifiers using the real labels provided in each dataset. In contrast, the second approach explored unsupervised learning to label the dataset anew, followed by classification with

the updated labels. Through this experimentation, we sought to assess the performance of both approaches and draw valuable insights for the advancement of fake news detection techniques.

In the first experiment, we utilized the LIAR dataset and applied our proposed supervised and unsupervised strategies. The supervised model achieved an accuracy of 69%, which, although lower than some state-of-the-art methods, demonstrated reasonable success in identifying fake news. Notably, our unsupervised model achieved an accuracy of 57%, which indicated that the supervised model outperformed the unsupervised approach, as expected. However, given the somewhat modest results, we recognized the need to explore other datasets to further evaluate our approach.

To address the limitations encountered with the LIAR dataset, we turned our attention to the ISOT Fake News Dataset for our second experiment. This dataset, with its larger size, balanced distribution, and binary classes, appeared to be better suited for assessing the effectiveness of our approach.

The results of our ISOT Fake News Dataset experiment were encouraging. Our supervised model achieved an outstanding accuracy of 100%, surpassing the state-of-the-art methods that achieved accuracies ranging from 92% to 99.95%. This remarkable performance showcases the potential of our approach when dealing with appropriate training data. Moreover, the unsupervised approach, combined with classifiers training using DistilBert fine-tuned, yielded an accuracy of 91%. This demonstrates the effectiveness of our proposed approach in simulating content labeling and its ability to achieve significantly high accuracy levels.

Overall, the results of the experiments are very promising. They suggest that the proposed approach is a promising new method for fake news detection

## 2.5 Conclusion

In conclusion, Our proposed approach has demonstrated remarkable success in addressing the challenges of fake news detection. Leveraging the powerful DistilBert classifier, we achieved an impressive accuracy of 91% on the ISOT Fake News Dataset, surpassing the performance of the LIAR Dataset, which yielded an average precision of 57%. This outcome underscores the potential of supervised learning in effectively labeling real-world

data, leading to improved classifier training and higher accuracy levels. By embracing real-world data and moving beyond the limitations of fact-checking datasets, our approach has showcased the potential for more reliable and accurate fake news detection results. The challenges of fake news detection, such as the similarity between fake and real news articles and the rapid spread on social media, necessitate innovative and robust methodologies. Our research contributes to the ongoing efforts in combating misinformation, and we believe our approach holds promise in making a positive impact on society's ability to discern truth from falsehood. By making our solution accessible through an API, we enable journalists, fact-checkers, or any concerned individuals to utilize our solution as a practical tool in the fight against misinformation. This implementation extends the reach and impact of our solution, contributing to a more informed society in the face of fake news challenges.

# General Conclusion

In conclusion, this study focused on the detection of fake news using contextual features. The approach involved a detailed methodology consisting of several key steps. Firstly, we preprocessed the LIAR dataset and visualized the results to gain insights into the data. Next, we employed three unsupervised learning techniques to label the dataset, ensuring the availability of reliable training data.

Furthermore, we fine-tuned the "DistilBERT" model and extracted latent representations of the labeled dataset. This step allowed us to capture the essential contextual information necessary for effective fake news detection. These representations were then utilized to train four models, including machine learning and convolutional neural networks, enabling us to leverage their respective strengths in classifying fake news.

In the evaluation phase, we rigorously tested and validated our approach. This involved assessing the performance of the trained models using appropriate evaluation metrics. We then used the same process with another dataset, the ISOT Dataset, to improve our results.

The results demonstrated the efficacy of our approach in detecting fake news, especially for the ISOT Dataset. The combination of unsupervised learning for data labeling, fine-tuning the DistilBERT model, and leveraging machine learning and convolutional neural networks contributed to achieving accurate and reliable detection outcomes.

In summary, this study showcased a comprehensive approach to detecting fake news using contextual features. The utilization of unsupervised learning enables the algorithms to discover patterns, adapt to evolving tactics, detect anomalies, analyze large-scale unlabeled data, and generate hypotheses. Also, the fine-tuning of the DistilBERT model, and the integration of machine learning and convolutional neural networks yielded promising results.



However, it is important to note that challenges persist in this research domain. The constant evolution of fake news generation techniques requires continuous monitoring and enhancements to existing models. In order to improve our work, several perspectives for future research can be identified. For example, using multiple datasets in different languages, such as French, Arabic and Chinese, helps to achieve "Multi-Language Fake News Detection". It is also possible to use other multiple supervised or unsupervised models. For the unsupervised ensemble learning, instead of majority-voting, using weighted voting, bagging or boosting may improve the results.

Finally, we believe that in a society where fake news spreads on social media at an unbelievable rate, our approach represents a powerful and innovative solution for identifying the spread of these fake news that can be adapted to all types of data and paved the way for new research opportunities in the field of fake news detection using contextual features and unsupervised learning. The advancements made here can contribute to a better understanding of the underlying mechanisms of fake news and the establishment of more effective preventive measures to counter its spread.

# Bibliography

1. Naeem, S. B. & Bhatti, R. The Covid-19 ‘infodemic’: a new front for information professionals. *Health Information and Libraries Journal* **37**, 233–239. ISSN: 14711842 (3 Sept. 2020).
2. Accessed: March 2023. 2022. <https://lalgerieaujourd'hui.dz/entre-humour-et-fake-news-les-algeriens-ont-desesperement-attendu-un-signe-de-la-fifa/>.
3. Slonje, R., Smith, P. K. & Frisé, A. The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior* **29**, 26–32. ISSN: 07475632 (1 2013).
4. Potthast, M., Köpsel, S., Stein, B. & Hagen, M. *Clickbait Detection* in. **9626** (Mar. 2016), 810–817. ISBN: 978-3-319-30670-4.
5. Ginsburg, D. H. & Shechtman, P. Blackmail: An economic analysis of the law. *U. Pa. L. Rev.* **141**, 1849 (1992).
6. Kshetri, N. & Voas, J. The Economics of ‘Fake News’. *IT Professional* **19**, 8–12. ISSN: 15209202 (6 Nov. 2017).
7. Statista. *Percentage of adults in the United States exposed to fake news as of March 2018* <https://www.statista.com/statistics/649234/fake-news-exposure-usa/> (2023).
8. Nyhan, B., Porter, E., Reifler, J. & Wood, T. J. Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior* **42**, 939–960. ISSN: 1573-6687. <https://doi.org/10.1007/s11109-019-09528-x> (3 2020).
9. Przybyła, P. *Capturing the style of fake news* in (AAAI press, 2020), 490–497. ISBN: 9781577358350.
10. Undeutsch, U. Beurteilung der Glaubhaftigkeit von Aussagen (Evaluation of statement credibility). In *Handbuch der Psychologie*, Vol. 11: Forensische Psychologie, Undeutsch U. *Hogrefe: Göttingen*, 26–181 (1967).

11. Sheng, Q., Zhang, X., Cao, J. & Zhong, L. *Integrating Pattern- and Fact-Based Fake News Detection via Model Preference Learning* in (Association for Computing Machinery, 2021), 1640–1650. ISBN: 9781450384469. <https://doi.org/10.1145/3459637.3482440>.
12. Hemina, K., Boumahdi, F. & Madani, A. *Automatic fake news detection: A review article on state of the art* ().
13. Yang, Y. *et al.* TI-CNN: Convolutional Neural Networks for Fake News Detection. <http://arxiv.org/abs/1806.00749> (June 2018).
14. Sastrawan, I. K., Bayupati, I. P. & Arsa, D. M. S. Detection of fake news using deep learning CNN–RNN based methods. *ICT Express* **8**, 396–408. ISSN: 24059595 (3 Sept. 2022).
15. Ahmed, H., Traore, I. & Saad, S. *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques* in. **10618 LNCS** (Springer Verlag, 2017), 127–138. ISBN: 9783319691541.
16. Kaggle. *Fake News Competition Data* <https://www.kaggle.com/competitions/fake-news/data> (2023).
17. McIntire, G. *Fake or Real News* [https://github.com/joolsa/fake\\_real\\_news\\_dataset](https://github.com/joolsa/fake_real_news_dataset) (2023).
18. Kaggle. *Fake News Detection Dataset* <https://www.kaggle.com/datasets/jruvika/fake-news-detection> (2023).
19. Padnekar, S. M., Kumar, G. S. & Deepak, P. *BiLSTM-Autoencoder Architecture for Stance Prediction* in (Institute of Electrical and Electronics Engineers Inc., Dec. 2020). ISBN: 9781728189192.
20. Guderlei, M. & Aßenmacher, M. *Evaluating Unsupervised Representation Learning for Detecting Stances of Fake News* (), 6339–6349. <https://github.com/magud/fake-news-detection>.
21. Khattar, D., Gupta, M., Goud, J. S. & Varma, V. *MvaE: Multimodal variational autoencoder for fake news detection* in (Association for Computing Machinery, Inc, May 2019), 2915–2921. ISBN: 9781450366748.
22. Boididou, C. *et al.* Verifying multimedia use at mediaeval 2015. *MediaEval* **3**, 7 (2015).
23. Jin, Z., Cao, J., Guo, H., Zhang, Y. & Luo, J. *Multimodal fusion with recurrent neural networks for rumor detection on microblogs* in (Association for Computing Machinery, Inc, Oct. 2017), 795–816. ISBN: 9781450349062.

24. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Communications of the ACM* **59**, 96–104. ISSN: 15577317 (7 July 2016).
25. Feng, S., Wan, H., Wang, N. & Luo, M. BotRGCN: Twitter Bot Detection with Relational Graph Convolutional Networks. <http://arxiv.org/abs/2106.13092><http://dx.doi.org/10.1145/3487351.3488336> (June 2021).
26. Aljohani, N. R., Fayoumi, A. & Hassan, S. U. Bot prediction on social networks of Twitter in altmetrics using deep graph convolutional networks. *Soft Computing* **24**, 11109–11120. ISSN: 14337479 (15 Aug. 2020).
27. Kaggle. *Politifact Fact Check Dataset* <https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset> (2023).
28. Zhou, X. & Zafarani, R. Network-based Fake News Detection: A Pattern-driven Approach. <http://arxiv.org/abs/1906.04210> (June 2019).
29. Abdulrahman, A. & Baykara, M. *Fake News Detection Using Machine Learning and Deep Learning Algorithms* in (Institute of Electrical and Electronics Engineers Inc., Dec. 2020), 18–23. ISBN: 9781665415798.
30. Shaikh, J. & Patil, R. *Fake news detection using machine learning* in (Institute of Electrical and Electronics Engineers Inc., Dec. 2020). ISBN: 9781728188805.
31. Li, D., Guo, H., Wang, Z. & Zheng, Z. Unsupervised Fake News Detection Based on Autoencoder. *IEEE Access* **9**, 29356–29365. ISSN: 21693536 (2021).
32. Yang, S. *et al.* *Unsupervised Fake News Detection on Social Media: A Generative Approach* (), 19. <https://www.cnbc.com/2016/12/30/read-all-about-it-the->.
33. Wang, W. Y. “*Liar, liar pants on fire*”: A new benchmark dataset for fake news detection in. **2** (Association for Computational Linguistics (ACL), 2017), 422–426. ISBN: 9781945626760.
34. Silverman, C. *How Partisan Facebook Pages Are Impacting Millions With Misinformation And Partisan Agendas* <https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis> (2023).
35. BuzzFeed News. *2016-10-facebook-fact-check* <https://github.com/BuzzFeedNews/2016-10-facebook-fact-check>.
36. Ruchansky, N., Seo, S. & Liu, Y. *CSI: A hybrid deep model for fake news detection* in. **Part F131841** (Association for Computing Machinery, Nov. 2017), 797–806. ISBN: 9781450349185.

37. Shu, K., Wang, S. & Liu, H. *Beyond News Contents: The Role of Social Context for Fake News Detection* in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Association for Computing Machinery, Melbourne VIC, Australia, 2019), 312–320. ISBN: 9781450359405. <https://doi.org/10.1145/3289600.3290994>.
38. Gangireddy, S. C. R., Deepak, P., Long, C. & Chakraborty, T. *Unsupervised fake news detection: A graph-based approach* in (Association for Computing Machinery, Inc, July 2020), 75–83. ISBN: 9781450370981.
39. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. *FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media* 2019. arXiv: [1809.01286](https://arxiv.org/abs/1809.01286) [cs.SI].
40. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. <http://arxiv.org/abs/1809.01286> (Sept. 2018).
41. Khalil, A., Jarrah, M., Aldwairi, M. & Jaradat, M. AFND: Arabic fake news dataset for the detection and classification of articles credibility. *Data in Brief* **42**. ISSN: 23523409 (June 2022).
42. Ouyang, Y. *Identifying Fake News: The Liar Dataset and Its Limitations* Medium. <https://towardsdatascience.com/identifying-fake-news-the-liar-dataset-713eca8af6ac>.
43. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31**, 651–666. ISSN: 01678655 (8 June 2010).
44. Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 86–97. ISSN: 19424795 (1 Jan. 2012).
45. Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record* **25**, 103–114 (1996).
46. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30**. <http://arxiv.org/abs/1706.03762> (2017).
47. Hugging Face. *tokenizers* NLP course. <https://huggingface.co/learn/nlp-course/chapter2/4?fw=pt> (2023).
48. Trask, A. W. *Grokking deep learning* (Simon and Schuster, 2019).

49. XGBoost Developers. *XGBoost Documentation* <https://xgboost.readthedocs.io/en/stable/>.
50. Maalouf, M. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies* **3**, 281–299 (2011).
51. Jahromi, A. H. & Taheri, M. *A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features* in *2017 Artificial Intelligence and Signal Processing Conference (AISP)* (2017), 209–212.
52. Hotzy, F. *et al.* Machine learning: An approach in identifying risk factors for coercion compared to binary logistic regression. *Frontiers in Psychiatry* **9**. ISSN: 16640640 (JUN June 2018).
53. Mussa, B. *A python function to get all the possible stats from a confusion matrix* Published in Towards Dev. Last accessed September 2022. 2022.
54. Kumar, A. *Accuracy, precision, recall f1-score – python examples* Posted in Data Science, Machine Learning, Python. Last accessed September 2022. 2022.
55. Bao, Q., Ivan, Z., Cong, S. R. T. T. & Diep. *Supervised Classification Methods for Fake News Identification* in (eds Rafał *et al.*) (Springer International Publishing, 2020), 445–454. ISBN: 978-3-030-61534-5.
56. Braşoveanu, A. M. P. & Andonie, R. Integrating Machine Learning Techniques in Semantic Fake News Detection. *Neural Processing Letters* **53**, 3055–3072. ISSN: 1573-773X. <https://doi.org/10.1007/s11063-020-10365-x> (5 2021).
57. Aslam, N., Khan, I. U., Alotaibi, F. S., Aldaej, L. A. & Aldubaikil, A. K. Fake Detect: A Deep Learning Ensemble Model for Fake News Detection. *Complexity* **2021** (ed Uddin, M. I.) 5557784. ISSN: 1076-2787. <https://doi.org/10.1155/2021/5557784> (2021).
58. Ahmad, I., Yousaf, M., Yousaf, S. & Ahmad, M. O. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity* **2020** (ed Uddin, M. I.) 8885861. ISSN: 1076-2787. <https://doi.org/10.1155/2020/8885861> (2020).
59. Sastrawan, I. K., Bayupati, I. P. A. & Arsa, D. M. S. Detection of fake news using deep learning CNN-RNN based methods. *ICT Express* **8**, 396–408 (2021).