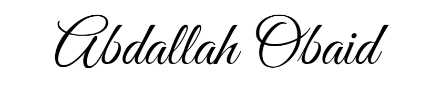
Tuesday, February 25, 2025

**News article classification**





Written by Abdullah Obaid

Under Supervised Dr Bassam AlKasasbeh

[21110383@htu.edu.jo](mailto:21110383@htu.edu.jo) / [abdallah.obaid@outlook.com](mailto:abdallah.obaid@outlook.com)

[Bassam.AlKasasbeh@htu.edu.jo](mailto:Bassam.AlKasasbeh@htu.edu.jo)

1 Al Hussein Technical University, Amman 11831, Jordan

Contents

[1.Introduction 3](#_Toc188873983)

[*My research questions is: -* 4](#_Toc188873984)

[*My objectives are: -* 4](#_Toc188873985)

[*Sections: -* 4](#_Toc188873986)

[2.Related Work 5](#_Toc188873987)

[*Papers :-* 5](#_Toc188873988)

[*Gaps: -* 10](#_Toc188873989)

[3.Data collection and Description 11](#_Toc188873990)

[3.1 Primary Data 11](#_Toc188873991)

[3.2 Secondary Data 12](#_Toc188873992)

[4.Research Approach and Methodologies 13](#_Toc188873993)

[4.1 Onion Model 14](#_Toc188873994)

[4.1.1 Philosophy 14](#_Toc188873995)

[4.1.2 Theory Development Approach 15](#_Toc188873996)

[4.1.3 Methodological Choice 16](#_Toc188873997)

[4.1.4 Research Strategy 17](#_Toc188873998)

[4.1.5 Time Horizons 18](#_Toc188873999)

[4.1.6 Techniques and Procedures 18](#_Toc188874000)

[4.2 Research Methodology 23](#_Toc188874001)

[5.Results and Discussion 28](#_Toc188874002)

[*5.1 Interviews: -* 28](#_Toc188874003)

[*5.2 Python code* 36](#_Toc188874004)

[*5.3 Were the questions answered after the analysis?* 43](#_Toc188874005)

[6.Conclusion and Recommendations 46](#_Toc188874006)

[6.1 Conclusion: - 46](#_Toc188874007)

[6.2 From the findings: - 46](#_Toc188874008)

[6.3 Recommendations: - 47](#_Toc188874009)

[6.4 Future Work 49](#_Toc188874010)

[7.Reflections 53](#_Toc188874011)

[7.1 Selected Research Methodology 53](#_Toc188874012)

[7.2 Alternative Research Methodologies 54](#_Toc188874013)

[7.3 Recommended Actions and Future Considerations 54](#_Toc188874014)

[7.4 Recommended Methodology 56](#_Toc188874015)

[8.References 58](#_Toc188874016)

بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

**Abstract:-**

In this world, there is a rapid development in digital content, and this development in the field of news texts needs intensive study and highlighting, especially with the development of artificial intelligence tools and the presence of some techniques that can be relied upon, whether they are techniques pre-trained on some words or traditional methods that are made from scratch, such as using TF-IDF. I made a comparison between these models, which used BERT-trained models and also TF-IDF models to make a comparison between the classification of news texts, which used BBC data. Many methodologies were adopted to conduct the study, which used the same qualitative and quantitative data collection techniques and analyzed them to conduct a comprehensive study of what was found in this data. All these processes come to answer one of the two questions that are raised in this paper, which are: How do news institutions and organizations use classification based on artificial intelligence? Also, can models that use TF-IDF excel in the classification process over pre-trained models such as BERT and compare them in the same data, which reveals that one of the best models obtained was logistic clustering and SVM In addition, I have made general recommendations for researchers who will continue this journey and some recommendations for news organizations that will use AI models in their organizations.

**Keywords:-** News text classification , NLP , AI in news organizations

# 1.Introduction

W

e are now in 2025 and modern technologies are developing very rapidly. With the development of these technologies, our data has developed, so data has become widely spread, especially in news sites and social media. Therefore, with this tremendous growth in digital content, classifying text data has become very important in this era, especially in the media field and the spread of news, because every minute we have new news in this world and it often spreads at a tremendous speed to reach all the inhabitants of this planet. Article classification aims to assign specific categories to all articles based on their content. This content may be, for example, in politics, sports, technology, etc., and classification can be used for a variety of purposes, including: -

1. **Recommendations: -** Customizing users' favorite digital content based on some articles they have previously read
2. **News filtering: -** Determining a person's priorities and displaying news that matches their interests and ideas.
3. **Improving people's experience on news sites: -** Improving access to news and reducing the spread of news that does not interest the person.
4. **Directing content to specific people: -** Ensuring that people get information that interests them.
5. **Reducing time Necessary for searching for news: -** Through the process of filtering and classifying news, the time required to search for news that interests the individual is reduced.

and so on, as there are many benefits and uses that can be used in classifying content.

Human language is characterized by being ambiguous in origin, and there are some words that are in the same sentence, the meaning of which is implicit and ambiguous, and there are also some sentences that depend on the context or grammar in their sentence, so researchers and scholars in the field of language and the field of artificial intelligence have emerged in this world and have brought us its essence, which is natural language processing (NLP) and other methodologies that can help us make our own classification. Natural language processing has emerged as the cornerstone of understanding and processing human language, and over the years, we have reached this advanced development of machine learning models and models specialized in deep learning, which has led to bridging the gap between the human understanding system (human language) and automated and electronic systems, and after years of research and development, little by little, only where in the beginning the classification relied on primitive or relatively simple techniques such as Rule-based systems and statistical learning These methods were effective in the past for small-scale problems that do not require multiple operations until there were other problems facing them, such as the difficulty of expanding and generalizing them to all applications or using them in a large and highly diverse data set. Machine learning appeared to us as a transitional stage for the process of machine learning and text classification, as it provided us with large models capable of learning from data without specific rules that oblige them to do so. We started first with machine learning, which was a real revolution and a rocket launch to what came after it, which was deep learning and learning from electronic neural networks such as ANN's, CNN's, RNN's, and also learning from transformers and using them to produce some tasks related to producing and processing texts such as BERT and GPT. We have reached what we are today in terms of great development in electronic classification processes. In this research, I will try to address some methods and techniques in the field of NLP and provide good content in this paper in the field of text classification, as I will use data (BBC Articles Dataset with Extra Features) which consists of 2127 rows and taken from the BBC News Network We have five types of classifications which are either in the field of sports or in the field of business or technology or entertainment and finally in the field of politics

The production of this research has scientific and economic implications in the media and digital content industry and beyond, as it can enhance the accurate and effective classification of sentences, articles and news headlines from providing content tailored to people and improving the ability to quickly access information and improve the management of text data whether this is through individuals or through institutions. In addition, this research helps in the ongoing dialogue about what are the best practices and emerging trends in the world of data science, artificial intelligence and natural language processing.

### *My research questions is: -*

1. **How do News businesses and institutions use the trained models for their applications?**   
   **The paper (ERNIE and Multi-Feature Fusion for News Topic Classification)** discussed the advantages of **ERNIE,** but it did not explicitly say how it will be used in real life and did not focus on its deployment in business in general
2. **Can traditional methods such as TF-IDF with machine learning models outperform pre-trained models such as BERT for the BBC dataset?**

**The paper (ERNIE and Multi-Feature Fusion for News Topic Classification)** showed that pre-trained models such as **BERT and ERNIE** can outperform other models but without testing this in real life and directly on the same dataset.

### *My objectives are: -*

1. Obj1: Based on ("ERNIE and Multi-Feature Fusion for News Topic Classification"), we will compare the performance of a pre-trained model such as BERT with our performance to find out whether traditional methods are better than pre-trained models for this specific case.
2. 2- Obj2: To understand how the news organizations can use the AI and NLP on their companies.

### *Sections: -*

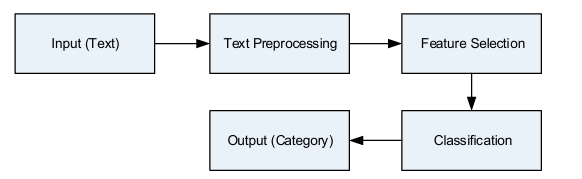
This research consists of eight sections, which are:

1. Introduction
2. Related work
3. Data collection and description
4. Research approach at methodology.
5. Results and discussions
6. Conclusion and recommendations
7. Reflections
8. References

# 2.Related Work

### *Papers :-*

In this first reference that I have reviewed, this study focused on the process of classifying news in Indonesian using artificial intelligence tools. The dataset included about five thousand records, which were taken from the website cnnindonesia.com and were divided into several categories such as technology, sports, and health. There were pre-processing steps such as removing unwanted words, lemmatization, and TF-IDF was used to select the data that would enter the model. The results showed that Multinomial Naive Bayes achieved a percentage of 98.4%. It was also fast, taking only 0.702 seconds. Compared to other models, it was the best, although SVM was good, but it took a very long time to process the data. This research focused on the importance of pre-processing and also how to choose features when you have data that is text. The image below (which was taken from the same paper) shows us the methodology that their project followed, as they started with the data, which was of the text type, then moved to Pre-processing process, then selecting features, then selecting the appropriate model for them, and finally we have the result, The paper does not talk about whether the data is balanced or not or if the article was classified based on whether the article was long or short. Also, the study lacked the use of the originally learned embedding techniques. They only used TFIDF.[1]



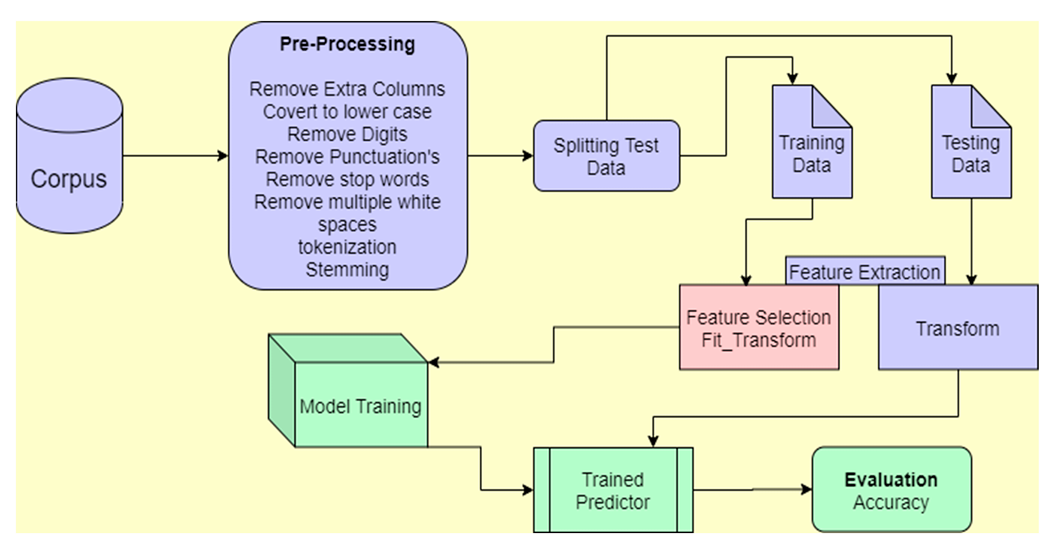
In this second reference that I studied, it was talking about classifying news articles using machine learning, where it used data that included approximately 210,000 records that were taken from the Kagen website, and this data was collected between 2000 2012 to 2022. Several algorithms were used, and their results were studied. Text mining techniques were also used, as well as some techniques that worked to convert unstructured data into structured data, such as TF-IDF. The percentages that were achieved in their experiments were as follows: Random Forest achieved the highest accuracy at 91.94%, followed by K-Neighbor Classifier (91.27%), Gaussian Naive Bayes (88.25%), and Decision Tree (79.19%) , In the image below (taken from the same paper), the methodology that they followed appears, where they first selected the data, then prepared it, and also worked on converting the texts into vectors, such as the vector in the TF-IDF, then they By choosing the features that interest them and finally putting them on their models, One of the things that we lack in this paper is that it consists of news headlines only, which means that they are very short, meaning that the model has learned on short sentences, and this leads to a decrease in the model’s performance when applied to a large data set, and the text is rich in words and also longer. This study relied mainly on using TFIDF only without using any other word embedding.[2] A diagram of a data processing process

Description automatically generated

In these three references that I studied, it focuses on the challenges of classifying news texts with a study of the length of the text and also its feature extraction. Also, a new model was revealed, called DCLSTM-MLP, which consists of a process of merging CNN, LSTM, and MLP. This new model came to work on learning the relationships specific to the text more and more accurately, as the accuracy rate came as follows: DCLSTM-MLP: 94.82%, MLP: 88.76%, Text-CNN: 92.46%, Text-LSTM: 92.35% and CNN-MLP: 93.68%, and it obtained the highest percentage among them all, What I noticed in this paper is that firstly the DCLSTM-MLP model can be complex and computationally expensive and requires very high resources to run. Also, this paper has limited evaluation metrics as the focus was on Accuracy and recall and other critical matters such as F1-score were not mentioned.[3]

In this fourth reference that I studied, this paper focuses on the process of improving the classification process for Bengali news by processing unbalanced data and also taking advantage of some modern techniques in the field of artificial intelligence such as BERT and SMOT, where the data was made balanced by using one of the techniques called Random Under-Sampling (RUS) and also using SMOTE, where after preparing the data, many models were used, whether they were regular models or deep learning models, and they were as follows (Logistic Regression, Decision Tree, Stochastic Gradient Descent, ANN, CNN, and BERT), where BERT achieved the highest result obtained, which was 99.04% on the balanced data set, and it obtained a percentage of 72.23% on the unbalanced data set, Some notes on this paper are the size of the data, as there is difficulty in finding many Bengali texts and putting them in the dataset even after using some techniques such as SMOT, RUS, this may affect our model. Also, using the BERT model requires very high computational resources. Finally, the ability to interpret, as techniques such as LIME AND SHAP were used to interpret the model’s predictions, but this increased the complexity of searching in the sentence itself. [4]

In this fifth reference that I studied, it provides us with the classification of news on the Internet using artificial intelligence, but with a focus on computational efficiency and reducing the complexity of each model. I also suggest in this study using the SVM model by improving it by changing some hyperparameters, because SVM is not used much in the process of classifying texts. Therefore, in this paper, a comparison was made between it and some other models such as SGD, RF, LR, KNN, NB, and the only thing that the superiority here came to SVM with a percentage that reached 85.16%, but it was the slowest of the existing methods that were used. The research also used some techniques such as removing noise and removing unwanted words, as well as stemming, and tokenization, The image below (which was taken from the same paper) shows us the methodology that their project followed , Where they started with me taking the data and making the Data pre-processing, then they worked on dividing the data into two parts, one for testing and one for training, then they made feature selection for the data that is specific to The train then worked on training the model and then they did their own evaluation, One of the notes on this paper is the need for high computational efficiency due to their use of SVM and also the lack of use of modern techniques in the process of converting words into vectors, using only TFIDF..[5]



In this sixth reference that I have reviewed, they used machine learning to supervise in order to classify the type of articles into some specific types such as politics, sports, and entertainment. A data set of 75,000 articles was used and collected from the HuffPost website. Pre-processing was applied, which included some steps such as removing unnecessary words and also why the encoder. Then the data was separated into two parts with a ratio of 30/70. Cross-validation was also applied to ensure the guarantee of strong evaluation. Some classifiers were tested, and the percentages were as follows: NB got a percentage estimated at 93%. In second place came LR with a percentage of 81%. In third place came SVM with a percentage of 76%. In last place came KNN with a percentage estimated at 72%, The image below (which was taken from the same paper) shows us the News article classification process that their project followed.

They first started by collecting data through a website, then they converted it to News Documents, then they downloaded the data after their own process, then they divided it and converted it to vectors for each article they had, then they trained their model and predicted the news class, and finally they did an evaluation of their model, One of the problems I saw in this paper is the incorrect distribution of the data. Although there are 75,000 articles, the distribution is not equal for some categories, such as politics, which is very common in the data. Also, one of the strange things is the lack of using specific word embeddings, or perhaps they did not mention them explicitly in the article. They only said that when they pre-processed the data, they used the NLTK library. Finally, they did not use all the metrics, such as F1-score.[6]

A diagram of a process

Description automatically generated

In this seventh source, the research paper talks about something called ERNIE, which is known as (Enhanced Representation through Knowledge Integration) with a multi-feature fusion, where most of the time the methods face some problems in classifying texts. ERNIE has the power to overcome some of the problems that occur during the process of classifying texts, such as the inability to capture some of the meanings in the text itself. ERNIE is a pre-trained language model that overcomes these problems by working to hide some words (like the mask on the BERT) and dealing with them in specific ways. It works to extract the main features in texts, contexts and important sentences and also provides attention to it, which includes a greater and broader understanding of the data. This model was tested on a wide range of data such as BBC News, where it achieved an accuracy rate of up to 97.47%, and all the metric results were significantly high. As for another set of data, the accuracy rate was 98.31%. It is also worth noting that the multi-feature fusion technique greatly helped this model classify topics. News by merging some diverse features that are related to each other in a unified framework, this work emphasizes the importance of taking advantage of modern methods such as ERNIE along with feature merging technology to ensure high accuracy in classification , ERNIE makes huge improvements in text classification processes, but it has several drawbacks, or we can consider them as challenges. First, it requires very large data for training, high quality and free of errors. It also requires high computational resources, as it requires much higher computational challenges than machine learning models, which makes it less useful in people’s hands.[7]

We will classify news as to whether it is fake or not. We have not strayed too far from the topic, as the topic is still the process of classifying news articles.

In this eighth reference that I have reviewed, NLP was applied to classify fake news and articles spread on the Internet. A dataset called LIAR-PLUS was used, which contains 14,787 classified phrases. Some techniques were used to help classify whether the information is fake or not, such as sentiment analysis, capturing news from fake sources, and also analyzing word frequencies and context, etc. The steps dedicated to pre-processing were also applied, such as removing some symbols and removing unwanted words. Several types of models were used, such as SVM, LR and NB. However, the SVM model obtained the highest percentage, which was 92%, and the other metrics were all superior in SVM over the rest of the models.[8]

A diagram of a model

Description automatically generatedIn this ninth reference that I have reviewed in this paper, a study was conducted on fake news related to the Corona Covid-19 disease using models for machine and deep learning, where the data, which consisted of real and fake news, was processed using text processing and also placed on specialized models such as Naive Bayes, Random Forest, Logistic Regression, CNN, LSTM, and Bi-LSTM, where CNN and Bi-LSTM achieved the highest accuracy of 97%. At the end of the research, the risks of spreading fake news in the health field are emphasized, especially during the Corona pandemic, and they explained the effectiveness of their models and the effectiveness of artificial intelligence in the process of detecting incorrect or false texts.

In the image behind (taken from the same paper), the methodology that they followed.

They first started by taking the data related to fake news, which was primary data, meaning they collected them, then started with the steps of pre-processing the data, then cleaning the text, then converting it to a vector to enter the model, then testing several models on it and coming out with the best result, then working on a model for prediction and producing the best result, then the data for the test comes and a prediction is made on it, and in the end we have the sentiment of fake news of test set[9]

### *Gaps: -*

Simplified applications of using artificial intelligence in the real world and in commercial uses in particular, where a problem came in the papers in that they lacked broad applications in practical life for the use of artificial intelligence and the use of classification in business, as these papers focused on improving the accuracy of classification and giving us new discoveries and studies in this principle, but they lacked practical implementation such as the common uses in companies and organizations for artificial intelligence in the process of classifying news, comparing different models, as these papers lacked finding a clear comparison on the same data in different papers and under the same circumstances, which left us with unanswered questions about some of the processes that may occur on the same data in different research and different papers. It was also noted that there are generally no reasons that tell us why this news was classified into a specific category, which left a clear gap that hinders understanding the classification of this model for texts.

# 3.Data collection and Description

Define the primary and secondary data and describe the differences between them.

**Primary data: -** It is a type of data that is collected by the person himself for the first time and is the information and data that needs direct answers. This method is considered expensive as analyses are used using large and technical resources. Data is collected through tests, observations, questionnaires, etc. [10]

**Secondary data: -** It is the data that has been made, implemented and collected by other researchers who have searched for it to reach an answer to their questions and it was not collected by the questioner himself. It can be in several forms such as statistics that are made by organizations and government agencies such as population censuses, books, newspaper articles whether paper or on websites, reports, etc. This method is simple and inexpensive in terms of research and material, but its main drawback is that it may not be very accurate or may not meet the main purpose of the research that is used later.[10]

3.1 Primary Data

The way I collected the primary data was through interviews, where I conducted a personal interview with two people who are experts in the field of natural language processing and text classification. Their questions were open-ended questions and one to two closed questions. The questions were simple and easy to understand for the people I asked, and all their answers will be analyzed.

The search and finding process was good and not complicated, given that the people are in the place where I study. The process of contacting them and booking an appointment with them was smooth and simple. Choosing the questions and making them simple and easy was difficult, so that I could obtain accurate information in the field of news classification.

I chose my primary data process to be interviews because it gives me greater knowledge and allows me to conduct a dialogue and discussion with the people I had been talking to. These interviews were effective and gave us the main idea from them and conveyed the required information in a better and greater way, especially since most of the answers were not finished, which helped us understand the biggest topics. Which were raised when conducting interviews with them and this experience was unique because it was the first time I had done this type of information gathering in this way. Also these interviews were very effective in the research and answered questions that were raised in this research

***Merits:***

1. ***Good delivery of information: -*** The interviews allowed me to conduct a detailed and comprehensive collection of data, which conveyed ideas in a greater and better way.
2. ***Clarifying the questions: -*** There was a question that was not clearly understood, so the interviews made it possible for me to clarify the question in a good way, which gave a better answer and a more accurate answer to the question.
3. ***Continuous modification: -*** When going through an interview, there may be a change in the questions that I can ask, and this is what actually happened, as I changed one of the questions during the interview that I conducted with them.

***Limits:***

1. ***Few interviewers: -*** Because of conducting interviews, the target segment (the people we are interviewing) may be small, which makes it difficult for us to generalize the results.
2. ***Relying on quality: -*** In The interview method depends on the person asking the questions. Since it is one of the few interviews I have done interviews, I have faced a problem with interview skills, how to ask questions, and also how to listen well to them.
3. ***Difficulty of analysis: -*** This method is more difficult to analyze than other methods, for example (surveys). It does not give us charts immediately, but people must analyze the answers in detail and more accurately.

3.2 Secondary Data

I used the BBC classification data (this is the link for the data that I had picked from Kaggle <https://www.kaggle.com/datasets/jacopoferretti/bbc-articles-dataset> (it is similar data that paper number 7 used (**ERNIE and Multi-Feature Fusion for News Topic Classification**) [7] )), It contain from *7 columns and 2127 rows* , the data is related for a BBC News , each text classified by the content for their class like (sport , politic ……etc.)

***The columns***

**Text: -** It is the full text of the news article

**Label: -** It is the category that the news article classify based on

**no\_sentences: -** It is the number of sentences on the article

**Flesch Reading Ease Score: -** It is a score shows how the text is easy to read

**Dale-Chall Readability Score:-** It is another measure that shows how the words ore difficult to read on the article

**text\_rank\_summary & lsa\_summary: -** They are summaries created by some algorithms.

***Justify my chose: -***

The purpose of using this dataset is because of the great similarity in the content of this data and the content of one of the papers that I read, which was titled (ERNIE and Multi-Feature Fusion for News Topic Classification), where they used one of the models called ERNIE and data like this was tested on it and in order to make a comparison between the results that I will come out with and the results they obtained using this data, I found that this is the closest data that can be obtained in addition to the great convergence in the number of rows, in addition to the fact that this data is very suitable for conducting research in the field of natural language processing due to the large size of the texts in it and the presence of many features that can be used in the search process and improving the results and coming out with the best result in the automatic classification process and the process of summarizing texts. This data also contained several features that can help researchers in the field of natural language processing by using it and giving them results that may be good and using some modern methods in text summarization processes such as using TextRank and also using LSA to make a summary For the text, which can help in the process of research, investigation and comparison between different summarization techniques if the researcher needs it

Using this data set is very useful because it provides us with readings that have been given to us to study and analyze texts, such as the number of sentences in the text and the ease and difficulty of reading sentences and words in each article, as this saves the researcher the analysis process related to text analysis, through which we can see the data better and more accurately as well, which makes it useful for research applications in various fields in the world of natural language processing

***The Merits and Limits***

***Merits: -***

1. **Different features:-** This data provides some features that can help researchers in some operations such as providing the degree of difficulty and ease of words in each text and also providing some summaries that were created automatically, which helps to focus on evaluating the model without the need to conduct extensive data analysis.
2. **Balanced data: -** Most of the data categories are equal, which means that there is no category that dominates all categories, which helps to increase the accuracy of the data.
3. **The data is ready: -** Because it is of the secondary data type, I do not need to collect it myself, as it is pre-collected.
4. **Diversity of texts: -** There are several categories into which the texts have been divided, which helps to cover many of our news articles. This means that the data is very useful for natural language processing applications to a large extent.

***Limits: -***

1. **Lack of metrics in the data: -** The data depends on three types of metrics such as no\_sentences, Dale-Chall Readability and Flesch Reading Ease Score, and these may not constitute a comprehensive and complete analysis. For texts
2. **Potential bias: -** Because the data was collected from only one source, which is the BBC, this (possibly) leads to bias towards news articles issued by the BBC, which can limit the possibility of machine learning on the data
3. **Automated summaries: -** If the researcher needs summaries, they may not be as accurate as what humans summarize

# 4.Research Approach and Methodologies

In this section, I will talk in detail about the research methodology that I used in this study, which included collecting primary data and then how to analyze secondary data. The process of collecting primary data was done by conducting interviews with experts in the field of natural language processing. These interviews were aimed at collecting insights from specialists in the field of artificial intelligence and natural language processing and benefiting from their experiences in this field, in addition to asking them some questions that aimed to reach an answer to them in this research, such as future improvements, benefits and challenges that media institutions may face in using artificial intelligence and the classification processes of the texts used by them.

Then, the secondary data was analyzed, as a data set collected from the BBC news website was used to evaluate the performance of our classification models. Some pre-processing techniques were applied, transforming the text form and training the models, which were models divided into two parts: Machine Learning Models and also pre-trained models such as BERT. In the end, we used the metrics that we measured the performance of the models with and checked whether there was overfitting in the data or not. Combining the research objectives collected from the interviews (which were the primary data) with the data that was analyzed (the primary data that was used from the BBC) and then drawing a diagram that shows how these two approaches work.

## 4.1 Onion Model

It is a model that was discovered and presented by Saunders, Lewis and Thornhill, these researchers, in one of their books. The aim of this model was to help students in their academic stages to write a thesis in an organized and clear methodological manner. It has several layers, which are: *Philosophy, Theory Development Approach, Methodological Choice, Research Strategy, Time Horizons and Techniques and Procedures* [12]

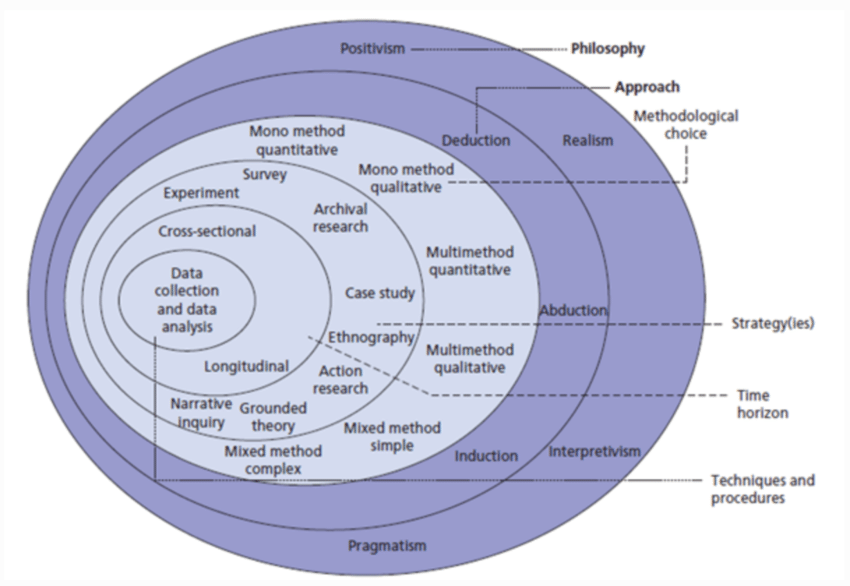


Figure 1: The structure of the Onion Research Model [13].

### 4.1.1 Philosophy

This layer tests the reality of the assumptions that the research begins to study, as this layer forms the cornerstone in shaping the research method in the research conducted by the researcher, as this layer directly affects the form and content of the research as a whole. This layer studies the assumption methodologies that the researcher performs, whether through analysis or through social experiments or points of view or interactions and human experiences. Some of the methodologies that are in this layer are:

1. **Positivism: -**It is the process of identifying some laws and patterns on which confirmatory tests and analysis of topics are conducted. This method is based on looking at the different assumptions from all points of view, while also focusing on observing some phenomena that can be seen with the eye and some measurable evidence. OR it focuses on topics that are realistic (seeing facts and facts) and extracting them so that it is not possible for two people to disagree on this fact. For example, water consists of two oxygen atoms and one atom of hydrogen. This is a fact that two people cannot disagree with. This (positivism) states that knowledge can only be acquired when conducting research. Experimentalism, which uses the process of measurement and observation, and emphasizes the establishment of some laws and the identification of general patterns through tests and analyses, with a focus on some phenomena that can be observed and measured completely.
2. **Interpretivism: -**It is the subjective nature of research, which refers to the methodology of assumption, meaning that through human experiences and the work of some cultural contexts and human social interactions, we have some experiences and some interpretations.
3. **Realism: -**It is the research theory that focuses on the fact that the research focus is independent of human capabilities, whether the human principle agrees on it or not, whether the human mind can understand it or not, it is a reality greater than human capabilities.
4. **Pragmatism: -**It is more between positivism and interpretivism, and this assumption looks at different points of view in the processes of conducting experiments and analysis processes

For my research (News text classification), I used ***Positivism*** because it was the most suitable for my research, because I am dealing with a set of data (BBC news dataset) and this data includes several ready-made metrics such as the number of sentences, word difficulty, sentence ease, etc. The main goal of my research is to compare the performance of the ERNIE model with models that will use traditional TF-IDF.

In my research, as I mentioned, I used Positivism because this philosophy emphasizes observation and measurement, which is what their study based on evaluating model’s needs. If I want to make comparisons between different methods, I must use measurable measures to make these comparisons and to make sure which models are better in relation to the measures used. This philosophy is suitable for me. It is possible to use several measures that can help us confirm which models are better, such as Confusion matrix, which can help us extract some measures that can be used to infer whether the model is efficient for this task or not.

I also relied on ***Pragmatism*** to discover how institutions and companies can use text classification using artificial intelligence and natural language processing.

### 4.1.2 Theory Development Approach

This layer contains the strategies that will work on developing, testing and understanding theories in research, whether it is starting with collecting data, looking at theories, or collecting and studying observations. This layer contains three research strategies, which are:

1. ***Deduction: -*** is the process of deducing and understanding theories or hypotheses, and then the process of testing all hypotheses is done by collecting observations and data on them to reach a specific goal about one of the specific topics.
2. ***Induction: -*** It is the process of observation whereby a person notices something and wants to apply it in a broader form and builds theories in order to generalize these observations or patterns that he observed. In other words, it begins with observation and ends with generating and establishing theories about the specific topic.
3. ***Abduction: -*** The focus here is on developing theories and hypotheses about some specific observations that were observed and ends with prediction processes.

In my research, I used ***Deduction*** because it is the best approach that can be followed, and the reason is that we already have a pre-defined hypothesis, which was that ERNIE performance and would it be better to use TF-IDF with machine learning Models are better or not, and we can also test the results that we will get through measurable measures and results, such as the measures determined by the confusion matrix.

### 4.1.3 Methodological Choice

This layer is concerned with the methods, how and forms of data that will be collected and the methods by which it will be analyzed later.

1. **Mono method quantitative: -** It is a method of using only one technique to collect quantitative results through experiments
2. **Mono method qualitative: -** It is a method of using only one technique to collect qualitative results through experiments
3. **Multi method quantitative: -** It is a method of using several multiple techniques to collect quantitative results through experiments
4. **Multi method qualitative: -** It is a method of using several techniques to collect qualitative results through experiments
5. **Mixed method simple: -** It is a method of using qualitative and quantitative techniques to better solve the research problem and to collect results through experiments
6. **Mixed method complex: -** It is the use of qualitative and quantitative techniques to better solve the research problem with the use of complex methods to collect data, analyze and extract information better

In this research, I used a ***mixed method simple***approach, which combines a quantitative and qualitative method in order to conduct a comprehensive analysis of news text classification using natural language processing.

I used The ***quantitative*** to evaluate the performance of the machine learning model that I used and also used some numerical statistical measures such as accuracy, which helps provide clear and measurable measures of the performance of this model with the data that I studied, which also helps in comparing between the machine learning that I used and ERNIE.

I also used the ***qualitative*** by conducting some personal interviews with people specialized in natural language processing who are very familiar with how to classify texts. These interviews came with the aim of understanding how companies supported by artificial intelligence can use the process of classifying news texts and helping them with it and discovering the benefits and impact of artificial intelligence in it and also identifying some of the challenges that may hinder this process.

### 4.1.4 Research Strategy

This layer is concerned with the path that will be taken to collect and analyze data.

1. ***Experiment:*** It is the process of manipulating and changing variables, I hope tests and seeing the results
2. ***Survey:*** It is the process of collecting a set of data from people through one of the techniques such as questionnaires, which helps in the process of analyzing people's opinions using them
3. ***Archival Research:*** It is the process of analyzing data and verifying the data found in old records.
4. ***Case Study:*** It is the process of studying a case or problem within the aspects related to this problem.
5. ***Ethnography:*** It is the process of observing and recording data for people's lives or people's culture, and this is done through close monitoring of them
6. ***Action Research:*** It involves using strategies, planning and cooperation between people in order to solve one of the practical problems that you face
7. ***Grounded Theory:*** It is the process of creating and developing some theories derived from data in general (whether historical or modern)
8. ***Narrative inquiry:*** It is the process of discovering problems, identifying them and solving them if possible, by discovering them through stories and experiences from people Different

In my research, I used a combination of archival research and case study to find answers to the questions I asked.

***Archival Research***

I first studied some data that helped me understand how to classify texts and news articles based on the speech in them, which helped the research to develop more in my knowledge in this aspect. I also used data issued by the BBC, which is data that has been made available to the public. This data will help us evaluate our model and train it on this data.

***Case Study***

I conducted some interviews with specialists in the field of natural language processing to discover how news classification based on texts will be used in companies and how companies and institutions will adopt these models that help in classification. These interviews will provide us with a strategic understanding based on real-world benefit from the model provided to them and how it will help companies classify their news After combining these techniques, this matter enhanced my research and enhanced my knowledge, as the combination of these two techniques led to enhancing the research and answering the questions better

### 4.1.5 Time Horizons

This stage focuses on the time period that the research takes, and here and about.

1. ***Cross-sectional:*** It is the process of collecting, analyzing and studying data within a specific time frame such as interviews and surveys
2. ***Longitudinal:*** It is the process of collecting, analyzing and studying data within a long- and extended-time frame such as an experiment

In my research, I used the ***Cross-sectional approach*** because this study focuses on analyzing data and studying ideas within a specific time frame and not extended over a long period of time, as I will analyze the BBC data set on Python code, which is non-continuous data, and I also conducted some interviews with specialists in the field of natural language processing, so I followed this time frame

### 4.1.6 Techniques and Procedures

In this section, I will explain the techniques, steps, and stages that I took in order to collect, analyze, and interpret data and information in my research.

1. ***Data collection through interviews: -*** I conducted semi-structured interviews (some questions were open-ended, and others were closed-ended) with experts in the field of artificial intelligence, especially in the field of natural language processing. These interviews came to **achieve** the goals that I wanted to answer some questions related to his research, such as how to use artificial intelligence to classify news, what are the benefits that may occur when using it, the expected challenges when using it, where will artificial intelligence reach in the future in companies, and use cases. The questions were semi-structured questions, meaning that most of the questions were open and others were closed. The number of questions was **11 questions**. These questions were carefully formulated so that I could extract detailed insights from them. As I mentioned, the questions came to the heart of the subject, which is the use of artificial intelligence in news organizations. The interviews allowed me the flexibility to ask questions and clarify questions that were not widely understood by people before answering. I **also recorded all the interviews** after taking (*Approval from the people I interviewed*) so that I could listen to it again and also so that I would not lose or forget any question or answer that I had asked during the interview, which also helped me to analyze the interview correctly and not lose any information, but one of the problems that I faced was the limited number of people I interviewed.
2. Secondary data analysis: The code (of Python type) that I used and implemented was of great help in analyzing the data previously collected from real data issued by the BBC news agency, where the goal was to analyze using libraries in Python and then train and evaluate the performance of all the models that we have, whether they were pre-trained or regular machine learning models. The process of analysis, processing and training was as follows:
3. 1- **Pre-processing the text: -** Here the process of preparing the data and removing impurities and defects from it takes place so that this data is correct and useful to us and does not negatively affect the model and it came as follows
4. *Removing punctuation marks*
5. *Converting words to lowercase letters*
6. *Returning words to their origin*
7. *Removing stop words*
8. *Separating the sentence and cutting it into words*

Thus, we have completed the process of preparing the text before we perform the feature engineering process and converting its form, and in this way all the data comes in a good and regular form and does not affect the performance of the model

1. **Feature engineering: -** Where the use of TF-IDF came in order to convert and change the form of the data into a digital form so that it can enter our models (this comes because the model does not receive words, but only receives numbers, so this feature came in order to convert the text into a digital matrix that can enter the model)
2. **Data division: -** The data set was also divided in a way that ensures a large training process, with the selection of data for training, and the percentage came to 80% For training and 20% for testing.
3. **Cross validation:-** This technique was tested to know whether this data set has overfitting or not (meaning that when training the data, we are given a very high result in the training data and bad in the test data) so because of the high accuracy that the models get, I used cross validation and this came to make sure that the data does not have overfitting and thus it turned out that the data is sound and has no problems
4. **Exploratory data analysis (EDA):-** Where I analyzed this data and extracted some visuals that could help us understand the nature of this data and I also made visuals before and after I did the pre-text processing to know what are the most frequently repeated words in each class before and after the processing process
5. **Training and testing: -** I trained this data on several models and also tested them. These models were the following.
6. *Regular machine learning using TF-IDF (logistic regression, random forest, SVM)*
7. *Pre-trained the model and it was BERT.*
8. **Metrics: -** In order to compare the performance of the models, we had to use some metrics on which we determined whether these models were efficient or not, and I used precision recall f1-score accuracy

After doing all these steps of dividing and transforming and making sure whether the data was correct or not, then training and finally measuring, all of this came to make tests to know the effectiveness of the pre-trained models in front of the small self like this compared to the normal machine learning models that were used, and after training, the efficiency of each model was measured and its results were verified more

Challenges on The Preprocessing

The data preprocessing process ensures that the data is well prepared before entering it into the machine learning models. This process ensures that it is free of defects and suitable for training, but sometimes there may be some challenges during the text preprocessing process that may affect the performance of machine learning models, whether they are based on TF-IDF or BERT, such as

1. Problems in removing stop words:- Removing stop words is very important because stop words can cause noise in the data. Removing them can reduce this noise in the data. We saw in ***5.2 Python code*** A graph that shows us that in one of the existing classes, the most frequently repeated words are stop words, so not removing them can affect the model, but at the same time, removing them can also affect the accuracy of the news. When there is a text containing the word NOT, when removing it, the news may be completely different. This matter can lead to problems in the classification process as well
2. Errors Spelling and typing: - It is possible, because some news stations rely on human reliance in the review process, that some articles contain some unintended spelling or typing errors from their employees, or even that these words are informal. This may reduce the effectiveness of one of the pre-processing steps due to the difference in symbols in the words. This means reducing the accuracy of the models we have.

Note: - I tried to test a code that corrects the text, and this code was effective on simple sentences, but this code has a defect as it requires a very large computing power to enter the data, and this is also another problem.

1. Dividing incorrect words: - Sometimes there may be a division of some incorrect sub-words, and this is very important for both pre-trained and untrained machine learning models. Incorrect and incorrect division means losing the context of the sentence. For example, dividing the word NewYork into both New and York affects the meaning and distorts the text, which reduces the accuracy of the model.
2. Marks Punctuation: Punctuation can sometimes be very useful and can be bad for texts. There are some sentences that are important to have a comma. When removed, it can reduce the accuracy of the model. Also, for BERT, which depends on the context of the meaning, punctuation can help to better understand the meanings.
3. Text Normalization Issues:- Converting words can convert some words incorrectly, which can reduce our accuracy and also reduce the validity of the text.

Potential impact on model performance

Errors that can occur in the pre-processing of data have a very large impact on the final output that can be produced. Errors can reduce our accuracy in the model, so that the model itself cannot understand the contexts between sentences and distinguish words with each other, which leads to severing the relationships between sentences and not connecting them. For example, TF-IDF errors can lead to incorrect representation of data in it. Also, if BERT obtains data that has no connection between its sentences, this can affect The accuracy of the model itself, and because its most important feature is understanding the contexts of sentences and linking them together, it only achieves this feature by entering poor quality data into it. Therefore, paying attention to the process of pre-processing the data is important and enhances the performance of the model more and increases the accuracy that the model is expected to reach.

The combination of primary data and secondary data was very useful to make a comprehensive answer to our questions, as the interviews (primary data) provided realistic questions and answers on the ground, while the secondary data (on which the model was trained) allowed for conducting experiments and tests and proving the validity of the hypotheses and questions that I raised, and these methods worked together to address The questions we have

Why Positivism for Model Evaluation?

This approach is readable and measurable, which makes it suitable for analysis of machine learning models. This is because machine learning models rely on metrics to evaluate their work, such as (accuracy, precision, recall, F1-score). This approach ensures that we have a consistent comparison between all our classification tools, such as TF-IDF and BERT. I also used some experiments that rely on numerical statistics that can be measured and compared, such as using cross-validation and confusion matrices, which also have numerical properties, making them better comparable. Why not use Interpretivism? Because it emphasizes the subjective experiences of human interpretation and opinions instead of numerical results, and this affects our evaluation of the model while it needs objective verification, as we must use measurable tools and stay away from human points of view that evaluate the model based on a personal point of view.

Why Pragmatism for Industry?

This methodology was used because it gives me the ability to combine quantitative and qualitative insights, and because in this research I am resorting to understanding how media organizations can use artificial intelligence tools in their work, and I have conducted interviews with experts in the field of natural language processing and asked them about this topic.

Pragmatism helps to find solutions and apply them in the real world, unlike both Interpretivism, which focuses on subjective experiences, and Positivism, which focuses on measurement. They come in between. We can use data and statistics that benefit us, and we can also use interviews.

Because this research is interested in the possibility of organizations to use artificial intelligence tools in their work, the best way to answer this question came by conducting interviews with experts, which enabled us to extract knowledge from people who actually deal with artificial intelligence tools and natural language processing, and also extract insights about the challenges that companies may face when adopting artificial intelligence, as well as points of view on how to apply it.

These insights that came out of the interviews are used by pragmatism to solve problems in the real world instead of expressing Expert opinion and also not just measuring on models without their applications

Why is Interpretivism not used in industry? This is because it focuses only on people's experiences and their beliefs stemming from their ideas instead of solving practical problems. If I use Interpretivism, I will only collect information from experts about how to understand how to apply artificial intelligence in institutions from the point of view of experts only without understanding how to apply it practically and realistically for companies. The research aims to find a practical solution for why artificial intelligence is adopted in media institutions, not personal interpretations of the role of artificial intelligence in those institutions

Does The Mixed Methods Beneficial?

Of course, it is very useful for research purposes, as it can enhance the study. Collecting several methods of data analysis, as well as collecting people’s opinions, studying and analyzing them, also leads to a deep understanding of the problems that the person wants to solve and find solutions for them. For example, if institutions have listened to the point of view of the people concerned in the classification processes related to artificial intelligence and used them without resorting to and experimenting with their own data and applying it, this could lead to failure in the process of using artificial intelligence because they did not use specific measures to study the model’s status and the results it came out with, but rather followed only one methodology. This will lead to the destruction of the digital transformation that we want.

## 4.2 Research Methodology

In this research, I used two strategies to collect data and then integrate them into the research to verify, confirm and find the answer to our questions. They were the use of interviews and then my use of the Python code that I did to analyze the secondary data.

**A)** I will start first with the methods of implementing the ***interviews***, which I did through these steps.

A diagram of a job interview

Description automatically generated

In each of these steps, which were for the purpose of creating and organizing an interview to answer some questions, each step had a significant impact on the result and on the research as a whole.

1. **Designing the Interview Questions: -** The questions were carefully designed with the research objectives and focused on the following (these are not the questions, but what the questions focused on)
2. *How companies adopt artificial intelligence*
3. *Benefits of using artificial intelligence in companies*
4. *Some challenges and limitations*
5. *Future improvements*

These questions were clearly chosen and in clear language, so that when the question is asked, it is understood by the people I interviewed correctly and smoothly, because any error in the formulation of the question may give us an answer that is not related to the research we want to present. All the words came in a language that technical people and specialists in the field of natural language processing can understand in a simple and smooth way. The questions were designed to be 11 questions, most of which are open-ended, with one to two closed-ended questions.

1. **Selecting the people: -** I selected 2 experts in the field of artificial intelligence, especially natural language processing, specialists and workers in it, as it was easy to find them because they are present at my university. I went to one of them and asked him to conduct an interview and he gave me an appointment. As for the other, I sent him an email and he set an appointment for the interview with me.
2. **Making the Interviews:-** I went to them at the appointed time for the interview and it was time to start the interview process, but Before that, I got their approval to record the interviews so that I could listen to them later and make sure that I didn't forget any information. I used the questions that I wrote, and the interviews were smooth and the questions were understandable except for one question that I had to clarify in one of the interviews, but the interviews were good. The questions were understandable. All the answers were in the required format. 11 questions were used in these interviews.
3. **Analyzing Responses: -** After recording the interviews, I listened to them several times to analyze the answers of the two people I interviewed and I determined the answers for each question.
4. **Integrating with Research Progress: -** Finally, after I analyzed and collected data from the experts, I will integrate these ideas with the analysis results from the secondary data. This is to ensure the validity of the research performance and to find a clear and final answer to what is required from the research.

Interviews are an effective strategy in the process of gathering insightful and sensitive information from experts. I have collected important information that I would not have reached if I had not met with NLP experts. Through discussions with them, I gained an important point of view about the possibility of companies using artificial intelligence that can classify news and how artificial intelligence will be integrated into the business of media agencies. The experts reached the light on several aspects such as the introduction of artificial intelligence media centers in their work during the next five years, in addition to their shedding a large part of the restrictions related to the use of artificial intelligence in these organizations. It is very important for organizations to adopt artificial intelligence, especially the presence of some advanced technologies that can improve their work and benefit from the infrastructure of this generation of students that can help them create structures of code with which they can achieve high profits and save some expenses.

Comparison Between The Methodologies

I chose interviews in my research as the main method for collecting data from experts. However, there was another method that could have collected data, which is surveys. So, in the table below, there is a simple comparison between these two methods.

|  |  |  |
| --- | --- | --- |
| Comparison | Interviews | Survey |
| Insights | Give us more and better answers which helps to convey the point of view in detail and in a better way | It captures more answers but less detailed and sometimes depends on the type of question if it is open or closed |
| Sample size | In this research, I limited it to a small number due to the lack of time | It is preferable to have more than 15 people |
| Flexibility of questions | Flexible questions that can be clarified during the interview | Fixed questions that cannot be explained to participants if it is online |
| Place of conduct | By phone/in person/or online | In the current era, it is used online and can also be used on paper |
| Can generalization be made | It is not possible to generalize if the sample is small and does not include a large number of people | In some cases, it can be generalized if it comes from trusted people and a large number of people |
| Time and resources | Slower and requires interviewing each person separately and organizing time for each person | Faster and several answers can also arrive at the same time |
| Methods of analysis | There are no methods for analysis | Answers can be analyzed using some programs such as Excel |

Why were interviews chosen over surveys?

1- The limited time for participants, which prevents the time to conduct a survey, as in order for the survey to be accurate, it must reach some of the people concerned and deliver it to more than one person interested in this industry, so time worked against us at this time.

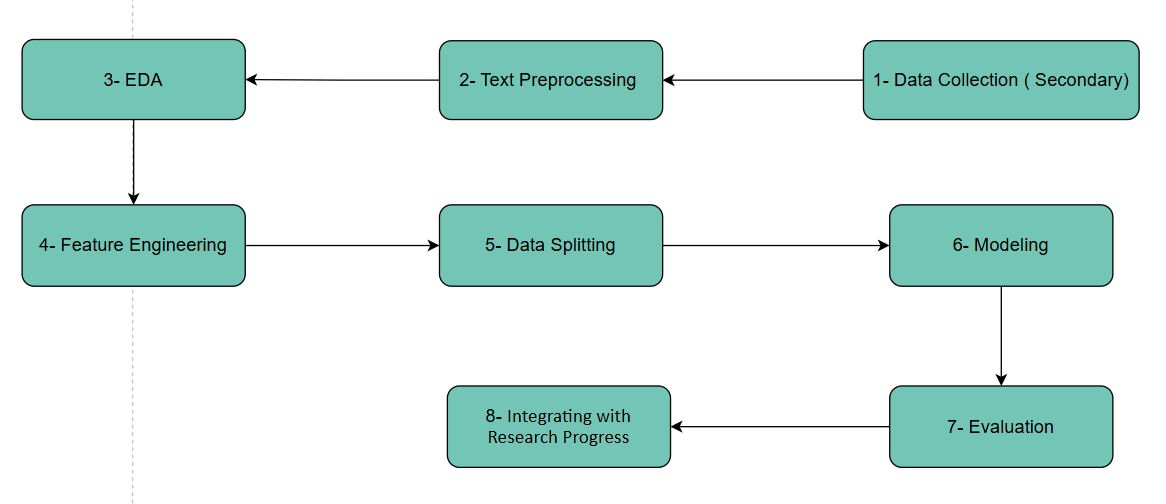
2- The type of questions I wanted to ask people allowed me to clarify some of the questions and enabled experts to give some detailed answers and expand on their responses to the questions.

3- The interviews allowed me to conduct direct conversations between me and specialists in the field of data and artificial intelligence, which helped in the process of communicating points of view better and faster.

4- Concerns about non-response and inaccuracy, as most of the surveys I have conducted previously suffered from the problem of low response and not taking these surveys seriously by those concerned with them, which makes it difficult to collect data well and also for this data to be accurate

**B)** Now moving on to **the code** that was used to collect data, enhance the search and answer the questions. It is a Python code that was created to predict the type of news that is present, whether it is political, business, sports, technology or entertainment. A data set from BBC News was used.

Below is the structure of the secondary data process:



1. **Data Collection (Secondary ):** This data was collected from the BBC News website, which contains 2127 rows. This data was available on the famous data collection website Kaggle. It also contained other features such as the ease of reading the text, sentences in the text, and the difficulty of the words in this text. It also contains ready-made summaries to implement the data on them, as they used a technique called Text rank and LSA to implement these features (but we will not use it. We will use the original text). This data set was chosen for several factors. First, it was classified into several articles for more than one category. It contains five types of articles. Second, the quality of the linguistic data set is good and from a good source, which is the BBC. Third, this data set contains features that can help the researcher in the process of analyzing the data and extracting some features from it. Fourth, some similar research was done on the same data or similar data, which confirms the quality of this data.
2. **Text Preprocessing:** The text processing step is one of the important steps that work to make the text correct and error-free before converting it and entering it into the model, at this stage several pre-processing steps were used, including:
3. *Removing punctuation marks: -* Have all punctuation marks been removed from the text.
4. *Making the text lowercase: -* This helps the model deal with all similar words in the same way without any difference between whether they are in capital letters or if they are in lowercase letters.
5. *Lemmatization: -* Returning words to their original forms
6. *Removing stop words:* Removing stop words is an important and common thing in the field of natural language processing. This comes so that the model can train in the correct words.
7. *Tokenization:* Dividing the text into words
8. **EDA:** In order to gain a comprehensive and broad understanding of the data, we conduct exploratory data analysis, where I used a set of charts to understand the data, such as the distribution of categories, where we found that the distribution is almost equal for each category of articles. I also analyzed the relationships between the readable scores and the text categories. This enabled me to see the hidden aspects of the data and analyze them more and better from within the data (I also made two plots after removing the stop words to see what words significantly affect each category)
9. **Feature Engineering:** My feature engineering was in some steps, the most prominent of which are:
   1. I did not enter a set of columns into the model because we did not need them in the study, and we were satisfied with only the columns that contain the text itself.
   2. Data conversion: - This was in the form of TF-IDF, where after pre-processing I converted the data into the form of TF-IDF Matrix, which is a technique that gives us the important words within each slice of data. The goal of conversion to this form came because the model cannot take data in the form of words but only takes it in the form of numbers. This conversion came to fill this gap so that the data can enter the model and the training and testing process occur.
10. **Data Splitting: -** I divided the data into two groups, one for training and one for testing. The distribution came at a rate of 80% to the training self and 20% to the testing data. This gives the model a large training process on a large part of the data, while also leaving a small part of the invisible data for the model to be tested on.
11. **Modeling: -** The modeling process was using machine learning models such as logistic regression, random forests, and also support vector machines, and another pre-trained model, which is BERT. This comes so that we can train the data on Several types of automated models, whether pre-trained or other models
12. **Evaluation: -** The process of evaluating models comes to know what is the best model that can be dealt with in the case of this data of ours, and I measured these models through the following methods
13. *Accuracy*
14. *Precision*
15. *Recall*
16. *F1-score:*

These are some degrees that can measure the effectiveness of this model, but I have not ended up here yet. After I saw that the percentage was very high in most values, I made sure whether the model was overfitting or not, and I used a technique called *Cross validation*. The results came out good in all stages of Cross validation. In this way, I made sure that the data was sound and good to deal with all our models.

1. **Integrating with Research Progress: -** In the end, the information obtained from the interviews was integrated with the data that was analyzed using secondary data. This is in order to solve all the questions raised in the research and find the best answers to these questions.

The BBC data set was used, and a comprehensive code was created for it on Python. This method was effective and good for evaluating the performance of all the models that were used and studying the data structure and what it contains secrets. It included pre-processing and also EDA, and then it was separated. Data, transforming it and merging it with models to give us the final result that we evaluated through the measurement tools used.

# 5.Results and Discussion

Discuss the results. Include all tables and figures.

Explain how the results meet the research question and objectives.

Describe merits and limits of the analysis.

After using the two strategies through which I collected data to solve the problem of news classification, I will now detail all the results that I came out with from these two strategies.

## *5.1 Interviews: -*

I conducted interviews with 2 university professors who are specialists in natural language processing and the use of artificial intelligence. I presented questions to them and listened to their opinions and points of view. How will news classification supported by artificial intelligence be utilized? What are the benefits that institutions can obtain and the challenges that they can face? Also, what are their future expectations for them?

The questions were as follows: -

* 1. ***How is the process of utilizing classification models supported by artificial intelligence carried out in media institutions?***

The first expert answered that in the Arab world, up to this moment, she has not read a study or seen a topic that talks about the use of artificial intelligence-supported classification of news, meaning that at this time, no type of artificial intelligence has been used in news newspapers, and that some newspapers still rely on humans themselves to classify news, meaning that the writer of the news is the one who classifies it into its content. The second expert said that media institutions can classify media news, benefit from data science and artificial intelligence by understanding the title and using some techniques in natural language processing, then creating a classification model, then classifying this title into several categories under which it falls. It is also possible to go further than this, such as going to the content of the text itself, and through it is possible to classify it as well. It is also possible to analyze texts and benefit from the main words in each text. It is also possible to do sentiment analysis on the text itself, and this is due to the fact that most of the news has become widespread on the Internet, and thus the news that has some problems is known (this matter indicates that The major technologies that media organizations can benefit from)

The first expert pointed out that the Arab world does not yet have a widespread classification of texts, as most of what is currently used is only by humans and is often classified based on the writer's own classification. The second expert suggested some beautiful content that could help news organizations in the process of classifying texts, such as identifying key terms, classifying titles, classifying news, and analyzing the sentiments of texts. Let's not forget that he also highlighted the use of some techniques related to natural language processing. The first question came as a door that led us to the main road that contains the complete reasons behind not using news classification techniques. Through this question, we discover that there is a great deficiency in the use of news classification in the Arab world and in the world as a whole, and there are techniques that could help us in this regard.

* 1. ***What are the reasons that make media organizations want to use classification models supported by artificial intelligence: -***

The first expert answered and answered that there may be an error in the process of human classification of news (since it is by people and not by machines), so it is possible for some errors to occur with the people who classify this text and gave us an example of the Al Jazeera Company containing a lot of political news, but it is possible that there is news that includes sports, so this news should not be placed (by human error) in the political content instead of the sports content

The two experts mentioned important reasons for using these technologies, the most prominent of which was human errors, which could cause a major problem in the classification process, as well as reducing the time and effort spent in these institutions, which means increasing the profits of companies due to the small number of employees who could play several roles in the process of vaccinating, classifying and publishing the correct article, as well as reducing time and focusing on publishing news without paying attention to the linguistic aspect of the news and the classification aspect.

The second expert answered when doing automatic categorization, this matter reduces the time that may occur when classifying it manually using humans or when holding a comment on their websites and analyzing it one by one and she also touched on the topic of fake news that artificial intelligence may help to detect it well

* 1. ***What are the best models for machine learning and artificial intelligence that are used for classification?***

The first expert answered that we must start with the basic steps such as LSTM and then build a solid foundation after which it is possible to branch out to models that are pre-trained such as BERT/BERTA/GPT.

The second expert answered Transformers, which can also be like BERT. She also added that vectorization techniques can be used with the use of a traditional model such as machine learning, random forest, or SVM.

The two experts agreed on the usefulness of BERT and also on the usefulness of Transformers in general, and each of them suggested some models that could help us in the process of classifying texts.

* 1. ***How are models improved after their application?***

The first expert answered that there are many ways that can be done to improve models, such as improving through.

The second expert answered that it is possible to modify the hyperparameter of the model itself. She also touched on the fact that if we use BERT, it is possible to work, change and modify their architecture, and if nerve cells are used, it is possible to make our own improvements by increasing the number of neurons. She also touched on the topic of feature selection, so that it is possible to use optimizations on the data itself, which helps the model to enter selected features into the model itself.

Improving was a main focus for us so that adjusting the parameters and improving the algorithms related to the models themselves came at the heart of this topic.

* 1. ***How do artificial intelligence tools help us in making operational efficiency decisions in media institutions?***

The first expert answered that it is possible for parents to relieve the burden on employees and reduce their operating time. She justified this by saying that behind every news item there are many people who work to review, publish, and improve the news, as well as classify it based on its type. This increases working time, effort, and cost for the organization. However, when applying these tools, it is possible for us to reduce these burdens on institutions. For example, we can create a model to classify the news and another model to detect whether this news has been reviewed or not, which helps reduce effort and time in this process. The second expert answered that it is also possible to reduce the burden on employees and the company and increase the opportunities to improve texts using artificial intelligence.

The two experts agreed that artificial intelligence helps news organizations greatly whether by reducing the operational burden of the organization in general and also automating the sites using artificial intelligence, which allows organizations to publish and review the article and publish it in a simple way and it is also possible to improve the quality of the text through artificial intelligence and this matter could help companies increase their own productivity

* 1. ***What are the challenges and limitations that organizations may face when integrating classification models?***

The expert answered: First, some small companies cannot afford large models that can help in their work, such as BERT, which requires large and huge servers to be handled, which means a significant increase in costs for small companies. Second, in terms of privacy, she believes that companies can create their own model without the need to use external sources (such as the cloud) to store it, which means that she believes that the privacy problem is not big for companies that can work on their own model and remain within the scope of the company.

The second expert answered that one of the problems that artificial intelligence may face is bias, as it may affect users' views of themselves due to the presence of some simple problems in the data on which the model is trained. She also touched on the issue of privacy, so that if the model contains data, it must be maintained very well, especially if it is on the cloud, because the model is training of data only, meaning that it contains the data itself.

The restrictions came in general in privacy issues so that privacy constitutes a major barrier for each of them. The first expert tells us that organizations can create their own system without the need to upload it to the cloud, while the second expert says that if the model is uploaded to the cloud, it contains some data, which means It is possible that this data could be hacked and bias was also greatly expressed in the opinion of one of the experts, as there was a fear that the model would distort some of the model's predictions.

* 1. ***Do you think that the process of classifying texts should be completely automated, or should there be human intervention?***

The expert answered that she divided her answer into two parts. The first part is in the early stages of the project. This is because the model cannot learn without human intervention. It will not be able to recognize the type of text. For example, a political text. The model cannot recognize it as political unless a human tells it that it is political. As for the second part, which was in the final stage, which is the stage in which there may be no human intervention but only monitoring without significant or direct human intervention.

The second expert answered that there must always be human intervention, whether in the training stage or in the validation stage. This comes so that a person can ensure that the model is able to correctly determine its tasks, such as correctly classifying news, detecting fake news, or analyzing sentences. There must always be human intervention at some stage of the model and ensure that the model is good.

The two experts stressed that there should be human intervention in the stages of using artificial intelligence in institutions, such as monitoring the project, in order to achieve reliability for this model and also supervising training to ensure proper training for models.

* 1. ***What improvements would you like to see in artificial intelligence models?***

The first expert answered and her answer was within the seventh question, as she saw that there is no 100% complete model yet without significant human intervention, and she wants the monitoring process of the model to be as little as possible, so that it is more automated than it is now.

This question was not asked by the second expert.

They stressed the need to automate work while reducing significant human intervention and making models work independently.

* 1. ***After five years from now, how do you see the development of the classification of models based on artificial intelligence in news and media institutions?***

The first expert answered that she hopes that artificial intelligence will be fully integrated into news sites, and this comes in order to reduce the burden on companies and companies, as there are many people who are behind the screens who do a comprehensive reading of all the topics that are circulated in media institutions and on their news pages, and this matter requires a great deal of effort on the part of the institution and the people themselves. It is also possible that we will see a model that makes improvements to the text itself that the writer writes and that it erases spelling errors from it. Last question

The second expert answered that she expects that most companies and media institutions will activate dual intelligence to a large extent in the next five years, especially after the current boom in the world of artificial intelligence that our world is currently experiencing. Its presence is no longer optional, but rather it was mandatory. There is a very important data science department in every company, but she confirms that within five years there will be some modern technologies that will appear in our world and make the matter easier and more effective.

Future expectations were good, as there could be great benefit in the future in the process of classifying data using artificial intelligence and also increasing productivity through the availability of some tools that could help news organizations, such as improving text quality.

* 1. ***Answer yes or no. Do current media institutions rely on artificial intelligence in classifying news?***

The first expert answered no.

The second expert answered no.

The two experts agreed that most organizations currently do not use artificial intelligence in news classification processes, and this indicates a gap in adopting the progress that is happening in the world, which could help improve productivity in media organizations and could also improve their income.

* 1. ***Answer yes or no. From your point of view, do you think that artificial intelligence has worked or may work in the future to greatly improve business, especially in news organizations, compared to old traditional methods?***

The first expert answered yes, and this matter lies in reducing the burden on institutions, especially since most institutions have people behind the screens who analyze, classify and prepare the text, and artificial intelligence can perform these steps quickly, which also helps institutions to focus on one thing only, which is writing news without paying much attention to the rest of the other operations due to the presence of artificial intelligence to solve these problems here, all of them. The second expert answered yes, and this matter comes due to the prediction process that may occur due to the presence of artificial intelligence and also improving the quality of news by delivering news to people interested in it, for example, athletes are interested in sports news, so a recommendation system is created for them so that the news reaches them, and the expert confirms that if work is done to reduce the bias of the model, this matter will lead to a major leap in the world of data classification so that the user will feel that the news is not biased towards a certain party, and it is also possible to make a classification of fake news, which helps the user avoid this news, and all of these technologies can make the end user feel very comfortable with the media that These techniques are used

The two experts stressed that artificial intelligence greatly improves business. It is specific to media institutions and can reduce burdens and increase the company’s profit by reducing the work burdens on employees, increasing productivity, and targeting people by using a recommendation system that displays targeted news to them.

***What is the merits and limits for my interviews?***

***Merits: -***

1. Capturing information directly from our helper helped us understand the information provided by the people.
2. The answers we were given addressed the main research questions I asked and based on them I selected the interview.

***Limits: -***

1. The sample size is small as it was only one participant so we may need more interviews and interviews.
2. Lack of diversity as the participants are university professors which may lead to less accurate insights from people immersed in the media industries.

These were the questions; answers and analyses I did in the interview with the experts in the field of natural language processing.

The impact of Small Sample Size For Interviews

When I started doing this research, I interviewed only two experts in the field of natural language processing and artificial intelligence, and they gave us good answers to the questions we asked, as they provided us with very important insights, but the small sample size of only two experts poses some problems and limitations that may affect the results we came up with because of this sample. One of the main concerns that are always taken into account is the bias in the opinions of the participants, with only two people we interviewed, their opinions may not be generally directed to our industry, which is the classification of news texts, as these two experts may have been influenced by their experiences and work in the field in which they work, and this matter may work to distort the understanding of the new updates of what may be happening in media institutions currently and their classification of news. I had to provide a more comprehensive, broader and larger sample in order to include the different people working in these fields, and also to ensure finding an accurate answer that we come up with from more than one party, and it may be More generally

We also have a problem, which is the inability to generalize, as having two experts cannot be aware of all matters related to the challenges facing media institutions. There may be several problems that media institutions may face that these two experts are not aware of. Therefore, conducting interviews with more than 20 people in various fields related to media institutions, for example, journalists, writers, and also artificial intelligence engineers, as they are currently involved in those institutions in which we want to integrate artificial intelligence, or in the process of selecting more than 20 people, it is possible for them to allow us to conduct analyses in various fields in which dual intelligence may help those institutions, and this linking process that is between non-technical people such as journalists and media professionals, as well as technical people in this field, may give us different visions and directions that may help organizations with future improvements and the integration of artificial intelligence into them.

Therefore, future research must include more than several people at different levels of work fields, as I mentioned, technical people and non-technical people familiar with the field of media institutions

Alternative Methods For Data Collection

As I mentioned several times in this research, only two experts in the field of natural language processing and artificial intelligence were interviewed and some insights were obtained from them. However, there are some alternative methods that could have been used or combined with the interviews, such as surveys, whether with journalists or specialists in the field of natural language processing, which could have worked on a real spread of views in the research.

Potential benefits from my point of view for surveys

1- Increased ability to generalize the problem, as surveys can reach more people at the same time. Finding 20 journalists or people interested in artificial intelligence in a long time may be easier than using surveys and sending a link for them to fill out.

2- This point is related to the previous one, as it may include broader points of view, as the large spread may lead to giving their opinions (even if it is through surveys and finished questions, there will be a discrepancy between the answers due to the lack of participants in the survey), and this matter may clarify We have a comparison between the point of view of technical people versus the points of view of non-technical people in the same survey

But why was a survey not conducted in this research

Despite the many advantages that surveys provide us with, such as the rapid spread of the survey and also the finished answer that we can use some programs and analyze them and extract graphs that may benefit us in the research, there are several factors that prevented me from using the survey in this research

1- The type of questions that I wanted to ask people were related to their answers sometimes, so I saw in this research that I should conduct the interview process

2- Fears of non-response and inaccuracy, as most of the surveys that I had previously conducted suffered from the problem of low response and not taking these surveys seriously by the people concerned with them, which makes it difficult to collect data well and also for this data to be accurate

3- The time frame in the research did not allow me to use the survey, as the survey must reach specific targeted people and their answers have a specific meaning, so I cannot give this survey to unqualified people to follow up on it and reach The people concerned require a lot of time for me to send them the file and fill it out and also take and analyze their answers

4- This research came to explore the visions of experts in the field of classification of news texts for institutions instead of conducting a comprehensive and broad study at the industry level and in a very large way so that all categories of contributors to the news production process participate greatly, which allowed the interviews to ask the specialists and provide detailed explanations to me and to ask them based on their own answers and clarify some questions for them that they could not understand and this matter would not have been present in the process of their use of the survey

How can I complete my research in the future and I found that the best way is to combine interviews (to obtain detailed and in-depth answers from experts) and combine them with surveys (to measure the views of institutions and trends specific to the industry that I am studying and to deliver this survey to a larger group that can give us some answers through surveys)

The research can move to a much further level if it happens There is a combination of qualitative and quantitative methods in the research process, and therefore I found that in the future the combination process is the best option.

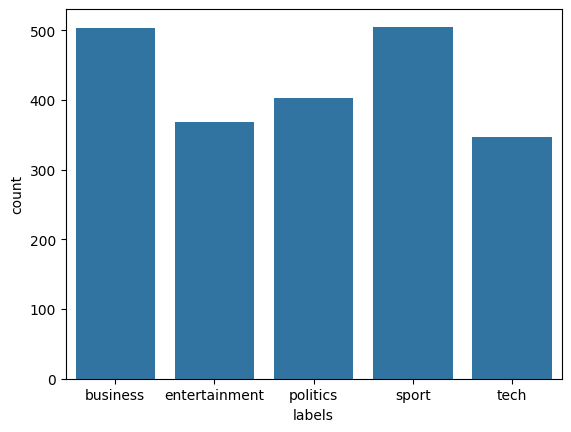
Now we will move to the next section (the code) in which we will have to discuss the results of the automated models I did and their predictions for what our text classification will be.

## *5.2 Python code*

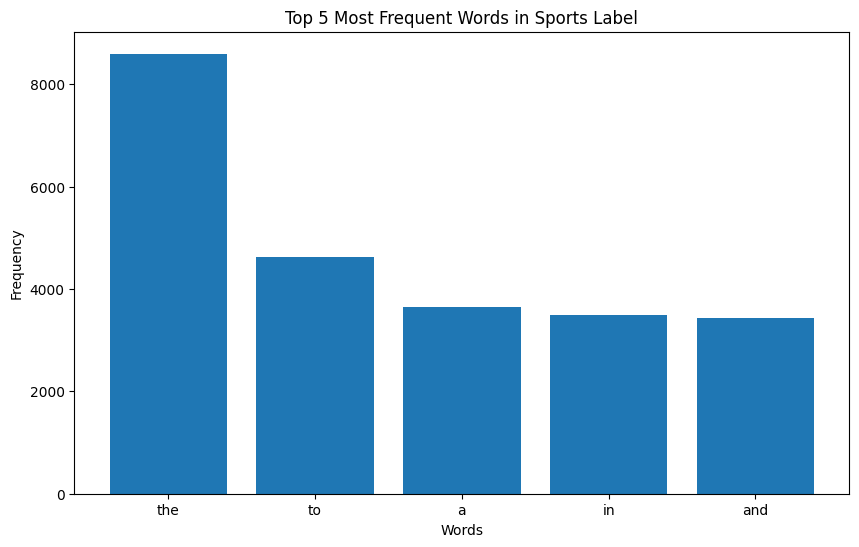
* + - * 1. **Data analyze**: - I will now analyze the data without going into the models in this step, i.e. just a general analysis of the data.

Starting with the 1st chart that we had it is a bar chart that contains each label that we had

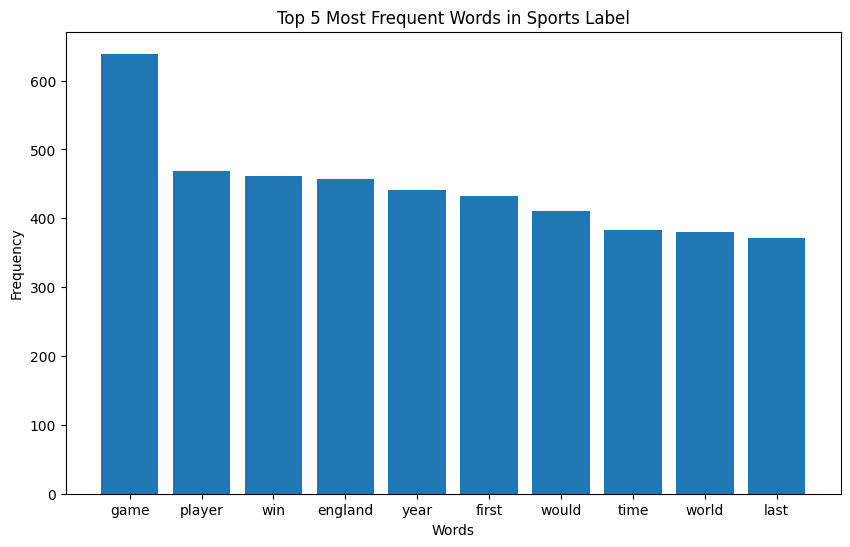
1. This chart shows that the data is distributed reasonably and almost equally among them, as the business and sports categories dominated the number of numbers present, while the entertainment and technology categories came in slightly less than them. It is important to ensure that the data set is balanced or not in the model and training stage, as it prevents bias towards one category over the other.



1. We have this chart (which I did before preprocessing the text) with the top five most frequent words in the sports category and as shown in the figure most of the words are stop words so having high frequency words like this can lead to bad results in our model and this is because words like And, to have very high frequency in the text which affects our model so we have to preprocess to remove these unwanted words



1. After analyzing the processed data, the distribution of words in the sports classification became more meaningful, as unwanted words were eliminated, which could affect our model, such as everything, but there were also many words that did not have a meaning, such as the word “said” and the word “mr,” so I removed them manually (just to know the most common words in this class). We saw the presence of several words that contain sports, such as “dame,” “win,” and “player,” as they are all words that indicate sports events. The pre-processing step of the data succeeded in making a set of improvements to the data, such as removing unwanted words and noise words that could affect our data.



* + - * 1. ***Modeling analyze: -***

Several types of models were used to classify the news. We used regular machine learning models, and we used pre-trained models. We will now see what these models are.

1. ***Logistic Regression:*** It is a simple model, but it is one of the most powerful models we have in machine learning. It is a statistical model used for classification tasks. It predicts the probability of whether this point is one of the possibilities presented in it, which are usually between zero and one as probabilities. It is an effective, simple, interpretable and executable model. It is an excellent choice for solving linear problems between features and targeted variables.
2. ***Random Forest:*** It is one of the machine learning models that depend on decision trees, where there is more than one decision tree that is trained with each other, then their outputs are collected and predictions are reached. Each tree is trained on some specific features, and then a majority vote is made for most of the trees to make a decision to predict this model. It is a strong model and is often able to overcome overfitting. Random forests are also strong against extreme values. This model is one of the most powerful models in the world of ML models.
3. ***SVM: -*** It is one of the machine learning models that works on classification and regression by finding the appropriate hyperplane that separates the points from each other using one of the techniques it uses. It is excellent for data that is in more than one. It easily separates this data using its hyperplane.
4. ***BERT (Bidirectional Encoder Representations from Transformers): -*** It is one of the most powerful models in the world of NLP. It is a pre-trained model based on what is known as transformers. It uses what is called bidirectional in its basic structure. This means that it goes through the data more than once and adjusts its weights, which helps it understand the context of the sentence more deeply and determine the meanings of the words in it. This model is one of the latest and most advanced in the field of artificial intelligence and the field of NLP.

But what is the difference between ***pre-trained*** models and ***traditional*** machine learning models?

1. ***Definition of each of them***

Pre-trained model: - It is a method of using artificial intelligence, which is the use of a model that has already been trained on a very large set of data and is different in its characteristics from regular models. These pre-trained models are widely used in operations that require the use of deep learning models, as they are very powerful and very effective in terms of the tasks that they can perform, and they are very useful in cases that do not require a large training process for their model (for example, finding a ready-made model trained on the meanings of words in the English language)

Regular models: - Which are also called traditional machine learning(ML) models, and they are models that are created and developed and then taught and trained from scratch for tasks related to the institutions themselves or for the purpose of using them for some goals related to artificial intelligence such as logistic regression, random forests, and also SVM

1. ***Advantages and Disadvantages for each of them***

***Advantages***

|  |  |  |
| --- | --- | --- |
| Model type | Advantage 1 | Advantage 2 |
| Pre-trained Models | Good start: - Pre-trained models are one of the good starters for developing models because they are pre-trained, which helps reduce the time and effort required to build other models | Excellent performance: - The performance of these models is good because the person who uses these models benefits from the pre-training of these models, which helps increase the efficiency of the tasks (and to get the maximum benefit, it is preferable to fine-tune them appropriately) |
| Traditional ML | Design and low requirements: - Models can be designed specifically for tasks, which makes them better in some scenarios, and their computational requirements are less expensive than using pre-trained models. | Ability to interpret: - Some models provide a simple ability to understand and interpret, which helps in decision-making processes and also for stakeholders to understand how these models work. |

***Disadvantages***

|  |  |  |
| --- | --- | --- |
| Model type | Disadvantage 1 | Disadvantage 2 |
| Pre-trained Models | Resource-intensive: - Some models require very high resources to use this model | Not knowing the trained data:- This matter can be negative sometimes, especially if the models are unknown and we do not know the data on which they were trained, as this may lead to bias in the data due to our lack of knowledge of the self on which the model was trained |
| Traditional ML | Limited performance: - You may face some problems when using some large data or when using them for complex tasks. | Lack of scalability: - These traditional models may face problems when using large data, so this makes them bad for the expansion process for large and large companies when using them. |

***Now for my results:-***

***Accuracy***

|  |  |  |
| --- | --- | --- |
| Model | Training | Testing |
| Logistic regression | 100% | 99% |
| Random forest | 100% | 96% |
| SVM | 100% | 99% |
| BERT | 100% | 98% |

As we can see, logistic regression and also SVM form the best comparison ratio between training and testing, as the training ratio came to 100% for them and the testing ratio was equal to 99%, which means that it learned well on the data and obtained a reasonable percentage of unseen data. As for random forests, it got the worst result inferring, which came to 96%, while BERT came to 100% for training and 98% for testing.

***Precision***

|  |  |  |
| --- | --- | --- |
| Model | Training | Testing |
| Logistic regression | 100% | 99% |
| Random forest | 100% | 99% |
| SVM | 100% | 99% |
| BERT | 98% | 97% |

The ratio came high here, which was 100% for both logistic regression and random forests and also SVM, and in the test also came to them 99%, which means that it correctly identified the correct cases without large false positive results, and this is good in relation to the accuracy of our models. As for BERT, the ratio came to 98 for training and not testing, it came to 97%.

***Recall***

|  |  |  |
| --- | --- | --- |
| Model | Training | Testing |
| Logistic regression | 100% | 99% |
| Random forest | 100% | 96% |
| SVM | 100% | 99% |
| BERT | 98% | 98% |

The ratio came to 100% in the test for both logistic regression and also SVM, and 99% For each of them in the data testing process, the random forests achieved a percentage of 100% in the training process and 96% in the testing process, which is a noticeable decrease from the other two models. As for the pre-trained cities BERT, the percentage for both testing and training was 98%.

***F1-score***

|  |  |  |
| --- | --- | --- |
| Model | Training | Testing |
| Logistic regression | 100% | 99% |
| Random forest | 100% | 96% |
| SVM | 100% | 99% |
| BERT | 98% | 98% |

The percentage was also as follows: 100% for both logistic regression and SVM in training and 99% in the testing. As for the random forests, as before, the percentage came to 100% in training and 96% in the testing, while BERT obtained a percentage of 98% in both tests.

Based on the results that came out to us from the tests that I conducted through the code, we see a clear superiority for both logistic regression and also SVM, which used TF-IDF with them, compared to BERT, where logistic regression and also SVM achieved in most percentages between 100% and also 99%, while BERT achieved a percentage ranging around 98%, although BERT is considered one of the strongest models we have in the world of natural language processing and also in the world of artificial intelligence, we see through these results that its performance is slightly lower than the other models, and the reason may be that BERT needs a large amount of data in order to be used, as it is specially designed to deal with large data sets that often require working with huge amounts of data in adjusting them, and also if we want to combine the words of the second expert with the results that came out to us, as she said, we find that in most cases when BERT is used, a modification is made to its infrastructure, on the other hand, dealing with small-sized data is very suitable for using a machine Learning models, even if they are slightly complex (such as having texts and some words in the data).

Overfitting Challenges on Models

Overfitting is one of the problems that every AI engineer faces, as a model works excellently on training data but has problems in dealing with unseen test data.

For example, overfitting occurs if a model achieves 99 to 100% on training data but gets 70% on unseen test data, this means overfitting.

Overfitting in traditional machine learning models

Overfitting can occur in some cases, such as if the model is too complex for the data that the model is trained on, or when using many features in the data, which may have a lot of noise and this data is not very pure, or if this data is unbalanced or too small for its model, which leads to the model memorizing patterns instead of learning them.

How can we reduce overfitting for traditional machine learning models?

It is possible to use Feature Selection, which can reduce our own data, for example, it can reduce Dimension of words that can be entered into TF-IDF and also Regularization can be used which works to penalize complex models and reduce their complexity and finally Cross-validation can be used which works to ensure that the model performs well across the data set and mixes it well (I used Cross-validation to verify the data in this project and it became clear based on Cross-validation that the data works well according to one of the machine learning models that we used)

Overfitting in pre-trained machine learning models

Machine learning models are huge models and contain very large parameters. For example, BERT contains from 110 to 340 million parameters (depending on the version that is used) and this huge number of parameters becomes clear to us when training BERT on some types of data, especially those that are very small and that have not been well pre-processed, it is possible for us to overfit

How can we reduce overfitting for pre-trained machine learning models

It is possible that We do the Dropout layers process, which is the process of disabling some neurons in the model in order to stop overfitting. It is also possible to use additional data and increase the articles that can be entered into the machine learning models in advance, which helps to increase the total number of data inside it and learn better. It can also use Regularization. We can also do fine-tuning, which helps to adjust our model and prevent it from overfitting.

## *5.3 Were the questions answered after the analysis?*

***How do News businesses and institutions use the trained models for their applications?***

Based on the experts I interviewed, we concluded that companies and media institutions can benefit from artificial intelligence models and natural language processing in their news and news classification in a way that enables the institution to make major operational improvements in the institution itself using artificial intelligence, as artificial intelligence can classify media news in several ways, whether by processing the titles, studying and analyzing them, and then classifying the news based on them, or by classifying the content of news articles in full and then knowing what this topic contains and then analyzing the content on it. Institutions can go further than this, while news is currently widespread on social media, the public can comment on this news. Artificial intelligence can also help analyze users’ feelings, analyze these texts, process them, target viewers, and deliver the news they want. Media institutions can benefit from artificial intelligence and the news classification process that was and still is done manually, and this matter is exposed to human error in news classification. Institutions can benefit from artificial intelligence for several reasons that make them want to own artificial intelligence technologies, such as It helps us reduce time and effort within institutions, which also reduces operational costs and relieves employees and makes the media institution focus all its thoughts on news, news production, news presentation, and leaving classification to artificial intelligence. Artificial intelligence can also work to detect incorrect classifications of articles in the institution and work to correct and reduce them. Artificial intelligence can also detect fake news, which is very important for people who receive news. This preserves the integrity of the media institution. There are many technologies that media institutions can use artificial intelligence in, whether at the classification level or at the level of correcting spelling errors or detecting complete errors. The presence of artificial intelligence technologies has become a matter of great importance in media institutions, especially since they deal directly with people, and if an error occurs, this matter will be very bad for the reputation of the institution. Despite these advantages and the widespread use of artificial intelligence in our current world, experts have found that there is a shortage in the use of artificial intelligence in media institutions at this time, especially in the Arab world. Experts have found that there may be some challenges that have posed this obstacle, such as the fact that it is difficult for small institutions to use artificial intelligence on Wide range due to the high computational cost on them and also fear of privacy problems that some users of artificial intelligence fear, especially when storing sensitive data on cloud computers.

If media organizations can benefit greatly from the technologies that are currently emerging and generated in our world, media organizations are among the entities that can work to benefit from artificial intelligence and natural language processing, as transformational models such as BERT can help them improve their news, improve the quality of the text, and also automate the organization's operations such as classifications. Artificial intelligence can also reduce the burden and reduce the cost on media organizations. There may be some obstacles, but these companies can solve and address them with a thoughtful strategy. With the continued spread of artificial intelligence in the world and the presence of big data in media organizations, it becomes imperative to use artificial intelligence tools in these institutions. Accordingly, institutions must use trained models in their applications. Thus, we have also made a complete understanding of the objective so that it has been understood how institutions can use artificial intelligence and language processing. Natural in its companies

***Can traditional methods such as TF-IDF with machine learning models outperform pre-trained models such as BERT for the BBC dataset?***

The answer is yes, we can get better results for the BBC dataset using traditional methods such as TF-IDF in conjunction with using them with machine learning models such as logistic regression, and this is superior to BERT, but what are the reasons that made BERT fall behind logistic regression? In my opinion, it goes back to one of the following reasons.

1- ***The data set size*** is theoretically small and pre-trained models such as BERT always prefer to be used in big data, as they need a huge amount of data to give us excellent and very accurate results, and with a small data size like this, it appears that there is superiority through using traditional methods in machine learning

2- ***Structure and efficiency: -*** Logistic regression is simple and uncomplicated and is used for simple tasks and is also faster in terms of implementation, while pre-trained models such as BERT are slow and complex at the same time, which makes it difficult to deal with small selves such as the one we have here

Here I also implemented the ***objective*** for this stage, where I built machine learning models and compared their performance through measurement tools such as accuracy precision recall F1 score and compared the results we got with the results that came out to us from BERT

Link AI Model Findings to Industry Applications

Should News Organizations Train BERT on Regional Languages?

The BERT model is a pre-trained model, so it has many languages ​​that have been trained automatically, but what about some regional languages ​​to ensure improving the quality of the results obtained from the model? It is preferable for organizations to adjust it and provide it with data specific to the regional location, in order to ensure understanding the context, or even some different types of BERT can be used that provide us with support for multiple languages ​​and dialects. Also, special training can benefit from the special dialect that is used in the regions to improve the quality of the news that BERT is used in its operations, whether it is classification or correction or what is related to machine learning models and their uses in media institutions.

Even at the level of the Arab world, BERT can be used for the Arabic language and it can also train some different dialects on it.

Accuracy and Computational Costs

Institutions can use some types of models specific to natural machine learning that are not pre-trained, and it is possible that they will give them good results, as happened with the equipment that we have. If the institution is small and does not have many uses for artificial intelligence, they can use traditional models.

As for Using BERT for large and large media organizations BERT provides us with very high accuracy, but it also requires very high resources and computing power, so some of the following solutions can be used, whether for small or large organizations

It is possible to use some types of versions in which BERT is lightened, which can maintain its accuracy but reduce computing costs

Or you can use cloud-related artificial intelligence services such as AWS, which are currently widespread in the labor market and are very widely used, especially here in Jordan, where AWS is widely used due to the provision of computing power and saving money that could have been spent on using BERT, which is also inexpensive and saves us some expenses such as maintenance, follow-up and huge infrastructure

Implementing Models in Real World Applications

There are many ways in which artificial intelligence can be used in news companies and also on the ground

As the interviews conducted with experts revealed to us some points that could lead to helping institutions increase their productivity and also save costs on them, and The applications came as follows

1- Artificial intelligence can be used in media institutions to automatically classify news. The news keeper comes and does not classify it, but rather passes it to a system. This system comes and understands this news and what it is classified for, and automatically places this news under the appropriate classification for it. This step can save institutions some errors that humans may make during the human classification process, in which a human error may occur and the news is classified into other content.

2- Artificial intelligence can make specific allocations for people based on what this person prefers and what he follows. For example, the artificial intelligence system can know that the person does not prefer economic news, but prefers only sports news. The system makes an allocation for this person in order to reduce the percentage of appearance of economic news and increase the percentage of sports news. It is also possible for the system to understand the sports news that he prefers. For example, a person who supports a certain club, the system studies this person and knows his love for this club, so it shows him the news of the club he follows.

3- Artificial intelligence can filter and detect false information and news because false news can be annoying Readers in media institutions and news pages on social media can be trained by artificial intelligence to detect this misleading news

4- Artificial intelligence can also help companies reduce the financial burden on them by reducing the human production process in institutions. According to experts, media institutions have many people who perform operations that artificial intelligence can do in a shorter time and at a lower cost, such as people who review spelling errors in news. Artificial intelligence systems can reduce the time required to do this step and also do it smoothly and quickly and at a lower cost than human hands.

All of this comes from knowing what the organization wants and what the organization wants to use artificial intelligence in, and then choosing the model of artificial intelligence to train it on the language to be used. After the tests that will be done, it will be used within the organization effectively according to the organization's need for it

# 6.Conclusion and Recommendations

### 6.1 Conclusion: -

This study came to analyze the use of traditional methods in machine learning along with TF-IDF against pre-trained models such as BERT in news classification to use data taken from BBC which classifies the news into one of the five whether it is technology, sports, entertainment, business or politics. The research also sought to conduct a study on how media organizations use artificial intelligence in their work and business.

### 6.2 From the findings: -

1. ***Performance comparison: -*** A general comparison was made between machine learning models using TF-IDF. These models were logistic regression, random forest and also SVM compared to BERT. The results were in favor of traditional machine learning models and the accuracy was slightly higher than BERT on BBC data. This result shows that pre-trained models can lose to traditional methods when dealing with relatively small data volumes, as BERT requires a huge amount of data to benefit greatly from its capabilities.
2. ***Natural language processing applications in Media organizations: -*** The interviews I conducted with experts in natural language processing showed that media organizations can benefit greatly from the current boom in our world using artificial intelligence, as there are many applications in which artificial intelligence can be used in their organizations by reducing human errors, improving the quality of news, detecting errors, reducing operational efficiency, increasing profits, targeting people, and so on. However, organizations must understand that there are some drawbacks that must be reviewed before using artificial intelligence tools in their organizations, such as high accounting costs, privacy, and news bias.
3. ***Gaps: -*** The interviews revealed a large gap in the use of artificial intelligence in some media organizations in the Arab world, as some tasks are still performed by humans, and media organizations must keep pace with development and adopt artificial intelligence technologies that are developing every day in our world.

### 6.3 Recommendations: -

1. ***Developing the artificial intelligence sector in media organizations: -*** Where artificial intelligence is integrated with their daily work and using artificial intelligence applications that work to improve operational efficiency and reduce errors Humanity, and it is also possible to increase the satisfaction of news readers.
2. ***Use in small institutions: -*** I recommend that small institutions create light learning models as they can be low-cost, fast-production and do not require additional loading beyond the institution's needs. They are suitable for them, and they can benefit from artificial intelligence technologies in a light and simplified way without the need to use strong methodologies such as the pre-trend model.
3. ***Integrating artificial intelligence into news systems: -*** Media institutions must install their own systems and integrate artificial intelligence into them. This gives some tasks greater accuracy and a significant reduction in time, such as correcting spelling errors in texts and classifying articles within their news site.
4. ***Detecting fake news: -*** With the spread of fake news, it is recommended for media institutions to create a feature to detect fake news. This helps increase the reliability of people in the media, which benefits the institution itself.
5. ***Strengthening recommendations: -*** Artificial intelligence can do By recommending people based on the content they prefer to watch, when this feature is done, the mood of news readers improves because important news reaches them, and this helps improve the income of institutions to increase the number of readers interested in the news within the institution, so I recommend making recommendations through artificial intelligence for people
6. ***Sentiment analysis: -*** Analyze the comments that are on the news on social media and study them and study the opinions of the public on the news.

Expand My Recommendations

To ensure the successful use of artificial intelligence in companies, it is necessary to use some organizational strategies for this matter and create some teams that can help ensure the success of this process.

It is possible to create some teams that specialize in ethical monitoring processes for artificial intelligence.

Any artificial intelligence model can be biased towards a certain category, even if this model is a pre-trained model such as BERT. This can lead to losses in companies due to multiple processes such as poor classification of news bulletins. It is also possible for the model to be biased towards some news categories and only recommend these categories to people. Therefore, to address this matter, companies must take some important steps such as creating a specialized team to follow up and review this model and also conduct some data audits to ensure that there are no problems related to the news published by the institution that the model can be biased based on. It is possible to use explainable artificial intelligence (xai), which helps explain the model's classification decisions and based on what the classification process was done for everything.

When using these procedures, the classification process in journalism can become a balanced process free of errors. And free from the biases that the model may fall into

How can the process of *integrating artificial intelligence into the workflow of journalism* take place at the present time

With the development that is happening all the time, media institutions must currently use artificial intelligence in their work. Artificial intelligence has come to improve the news process in general. Artificial intelligence can perform some tasks that help journalistic institutions increase their productivity through

1- The classification process. News can be classified automatically without any problems. This comes in addition to the possibility of human verification that this system supported by artificial intelligence has correctly classified it

2- Artificial intelligence can perform the process of linguistic correction and correct spelling errors that may occur in the text provided by the writer. This comes very quickly and with high accuracy

3- Detecting fake news. The system on news pages can track and detect fake news, which helps to prevent the public and people from being stressed by fake news that may occur

4- Recommendation processes. Recommendations for people are very useful for institutions so that Increasing consumer confidence and their sense of comfort when dealing with the news organization

Journalists and press organizations can use artificial intelligence and benefit from it, as the journalist can scrutinize, classify and review the news in a very short time and without much effort. Consumers (news readers) can also feel confident with the organization they listen to or take news from because there is no fake news and also allocating news specifically for them that attracts them. These methods can work on journalism and increase their own production and increase the readers and consumers of the organization that implements artificial intelligence in it, as artificial intelligence can increase the speed of production while maintaining the integrity of honest journalism

How can journalists and editors (non-technical people) interact with artificial intelligence when they do not know anything about it

The integration process that can occur in media organizations is undoubtedly that there are people who do not know how to deal with artificial intelligence tools that can be used in those organizations, so they must first provide training for employees to familiarize them with artificial intelligence and familiarize them with it (Even if it is non-technical) on how the process they want to use is done, in other words, how should journalists and news editors trust artificial intelligence? This is an important step for them to trust artificial intelligence, and then secondly, teach them how to use these tools.

Solving some of the problems that small media companies may face when using artificial intelligence tools.

Small media companies and institutions often lack the resources that can work to use artificial intelligence through them, so these institutions can use some of the following solutions.

1- If these institutions are small, they can use light models, i.e. traditional models that can do their work if it is simple.

2- If they need to use pre-trained machine learning models, they can use some light models that do not require high resources to use.

3- Using platforms that support artificial intelligence in the cloud, for example, using AWS, which is a platform that is currently widely used in institutions that want to introduce artificial intelligence into their work.

These steps can provide a simple strategy that helps institutions to make technical improvements to them and integrate artificial intelligence and how to integrate artificial intelligence into their work.

### 6.4 Future Work

1. ***Using other sources of data: -*** Testing several sources of data and several types in several languages. This matter can work to analyze the strength of models in different scenarios.
2. ***Using interpretable artificial intelligence XAI: -*** It is a technique that helps us understand and explain how the model works on classification. It is important to discover how and why the thing was classified to its place.
3. ***Fine-Tuning Techniques: -*** This helps improve the model and make it more useful and able to deal with diverse types of data.
4. ***Using real-time data: -*** To verify the work of models while using data in real time Actual
5. ***Changing models: -*** Trying several types of models, whether traditional models or pre-trained models, and arriving at the best model that can deal with data in the form of news texts.

Future Research

Structured Roadmap

To ensure the continuity of the process of classifying news texts using artificial intelligence and ensuring better results and better sustainability in future projects and research, future studies should focus on several things such as making the data set larger and improving the model through several factors and then increasing the transparency of artificial intelligence and knowing how its classification process works and understanding why this classification happened and the main roadmap for the future is as follows

1- Expanding the scope of the data set: - The data set that I used came from BBC New and it is very effective data and gave excellent results, but it is small and limited in scope, so future research should do the following

Use data from different sources and different languages, and it is preferable for it to have several news networks (in order to give us the best results and to train on different languages ​​and dialects. For example, the data set comes from Al Jazeera Channel, Al Mamlaka Channel, and Roya Channel). Therefore, this step comes to make linguistic improvements in the context of the study and improve the model's capabilities and make this model able to generalize its results on different platforms. It is also possible that In different regions

They can also use data via social media, and this can help the model recognize the real languages ​​that humans use in their languages ​​and communication

It is also important that there is a balance between the data sets used, and this comes so that we do not have bias from one category to another and give us high accuracy

2- The process of adjusting hyperparameters: - Adjusting the hyperparameters is important in any company when using machine learning models, whether pre-trained or traditional, and this comes so that the model is better adjusted for the problem that the institution needs to solve or a goal in the institution from using artificial intelligence, as each institution differs in its needs and capabilities in the model it needs

Future work should include the possibility of using one of the methods through which hyperparameters can be used and adjusted, for example, Grid Search & Bayesian Optimization can be used, which are some methods for determining the best hyperparameters for any existing model, so they can be used BERT

For note only, the definition of each of Grid Search & Bayesian Optimization

Grid Search:- It is a method by which the artificial intelligence engineer puts the parameters that he wants to test on the model he has, where he creates a training network that takes each parameter and tries it with a parameter and gives us the best results available based on the results we got from each parameter, although it is good sometimes, it can be computationally expensive, especially if it is used with models that contain a large number of hyperparameters

Bayesian Optimization:- It is the process of adjusting the hyperparameters by approximating the function, where he determines the parameter based on the previous results that he obtained through the previous results [14]

Explainable Artificial Intelligence XAI: The special classification comes in artificial intelligence models such as the black cube, which makes non-technical people such as media institutions and journalists unable to understand how the classification occurred and the reason for classifying articles within the categories specified for them. This leads to a lack of transparency between non-technical people and machine learning models. This affects some decisions that may benefit the institution and for which artificial intelligence will be used. To address this issue, what is called explainable artificial intelligence XAI was created. It is based on principles such as SHAP and LIME, as it provides us with readable results and an understanding of how the classification occurred and what this decision was based on.

SHAP: It is a technology used to set a specific value for each feature in the data to indicate how much it contributed to producing the decision taken. For example, in the process of predicting the risk of heart disease, the machine learning model may use some features such as age, body mass index, physical activity, and family history of injury. Here, the amount of each feature's contribution to it is determined. [15]

LIME: It is a technology that generates some specific approximations for predictions. The model, such as when predicting sentiment in a sentence, LIME mixes the light on the important words to indicate them in the specific prediction process that occurred. [15]

Some visualizations can also be made of how the prediction occurred and what we have come out with, which increases the confidence of journalists in the artificial intelligence technologies that we provide.

Through the integration process that will be developed in the future, this helps people understand, whether for people who want to conduct future research, understanding why this prediction occurred for the model and based on what the prediction process occurred. This matter is viewed from two sides: the technical side for people who want to conduct future research that confirms to them how to predict and what are the words that can have an important impact on the news that is classified, and on the other hand for non-technical people, journalists or media institutions, they can get a clear explanation of why this news was classified into the appropriate category for it

Real World VS News Text Classification Models

The news classification process has many benefits that can be provided by artificial intelligence, but deploying these models in real life poses many challenges related to several issues such as computational costs, ethical considerations, and potential risks from misleading information. Therefore, if we want to deploy the use of artificial intelligence models in media institutions, we must face these challenges.

1- Resources and computational costs: - Artificial intelligence models, especially pre-trained models, require very high computational resources, which makes their use and deployment costly for institutions, especially small and large ones. Institutions may need very high graphics processing units, which may be special, and costs for maintenance, air conditioning, cooling, and energy consumption operations are also a big burden on institutions, but there are some solutions that can benefit institutions, such as using light and light models. There are some models that give us high accuracy, but without great computational complexities. They can also use artificial intelligence models that are linked to cloud computing and use some platforms that provide them with this service, such as AWS, which We currently see it widely spread in the markets and it is used and some models are trained on it and the use of such platforms can work to save severe costs on institutions such as energy costs, maintenance, air conditioning and cooling. All of these costs are also reduced for future research, conducting experiments and tests on the work environment that companies can use such as AWE and experimenting with it and studying their future research papers on it.

2- Artificial intelligence, biases and ethical problems: - Artificial intelligence models that are trained incorrectly and have bias can cause problems in classification processes such as inappropriate classification of economic news when political news. This happens unintentionally from the model and due to several factors. One of the challenges that the model may face is bias in the type of data it was trained on, whether this data is unweighted or contains bias in it. Among the solutions that can be used are: First, it is possible to use some techniques that help us detect bias in the data. Second, a team specialized in ethics and oversight can be created to review the institution's model and ensure that it is compliant. Thirdly, for the journalistic standards of the institution, it is possible to use multiple data types and contain equal representation of all data within it, which does not cause us any bias towards viewpoints in any way. Researchers in the field of data and artificial intelligence can use some methods that can be studied in detail to see if they are useful for detecting bias and trying to reduce it through modern methods.

3- Detecting misleading information: - Artificial intelligence tools must be able to detect misleading information for news. This is a major challenge in light of the spread of social media in our hands and the ease of spreading fake news and incorrect information from unreliable sources. The challenges in this area come in that there is some unreal news that classification processes cannot detect whether this news is false or incorrect or not. Therefore, there may be some models capable of detecting this misleading information. Future research should explore this aspect and ensure finding an appropriate way through which this news is detected and using modern techniques to verify facts and claims that spread quickly.

Because journalistic work is a very important job And it must integrate artificial intelligence into it to ensure higher and greater efficiency and practice in this field. Media institutions must first solve some of the problems that they may face, such as computational costs, which can be addressed by using some solutions such as going to the cloud, as well as some ethical concerns that may occur and biases that can be addressed by creating some teams or training the model better. Finally, the risks of misleading information. Some models must be found that work to detect this news quickly. By addressing these challenges, media institutions can ensure the creation of new platforms that rely on artificial intelligence in their work and by ensuring that artificial intelligence can be subject to journalistic standards and journalistic integrity and reliable in terms of the training processes that the model undergoes.

# 7.Reflections

Avoid generalization and focus on personal development and the research journey in a critical and objective way.

### 7.1 Selected Research Methodology

The adventure of writing the research was a unique experience that allowed me to develop my view and thinking about how to choose, apply and think about some methodologies that could achieve meaningful results in the study. I conducted this study to address the topic of news classification using artificial intelligence tools, especially natural language processing, by putting it under theoretical and practical tests. I was surrounded and aware of some limitations that could hinder the methodologies that I would choose, as for all positive things, there must be something bad or something that the researcher did not take into account.

1. **Positivism and Pragmatism: -** I chose Positivism as a basic phrase that I relied on in my research because I was able to use some measurement tools such as accuracy and F1-score, through which I measured the results of machine learning models and also to compare them with the performance of pre-processed models. I also relied on Pragmatism by conducting a study with experts in the field of natural language processing, and this allowed me to collect realistic visions from experts with Analyses that I can get out of the data analysis process for models ,One of the advantages that I noticed in my research is that I was able to measure and compare models clearly, and I was also able to use the real data that I collected from experts and use it in this research along with the data that I used to study the performance of models in the classification process. One of the bad things that I faced is that Positivism can give us very good results, but without going into much detail about how it works.
2. **Deductive Approach: -** I used it through the hypotheses that were previously present and that I extracted through related work that I read, for example, comparing the performance of pre-trained models against traditional models, where I tested these hypotheses on the same data set for all models to extract insights and reveal the results through them. One of the advantages that was in this methodology is that it allowed me to conduct experiments on the scope of some hypotheses that were previously imposed by other researchers. One of the negatives is relying on some hypotheses that were previously present, so that if you do not find this hypothesis They are hard to find a match for
3. **Mixed-Method Simple:** The process of merging quantitative and qualitative techniques came to analyze all the data we have, where the quantitative analysis came to evaluate the work of the models using tools based on measurable numbers to ensure the correct comparison process between the models, while the data came from qualitative data through the interviews I conducted with experts. Among the advantages of choosing this methodology is not relying on a single source of data, which helps to increase and develop ideas and increase knowledge related to the research process. This is a joint approach between data on qualitative forms and on comprehension forms, which enhances the results and validity of the research. Among the limitations that may be related to this methodology is that it requires a lot of work to achieve the goals of these two types of data, and it may also take a long time in the process of collecting, analyzing and extracting results in this method.
4. **Archival research & Case Study:-** Archival research and also Case Study helped me to view some historical data that was studied through previous research, whether this data was from articles and references that I reviewed or through experiences The experts and their visions that I interviewed and knew what is the best way to use news-related classifications in news organizations and transform them using artificial intelligence tools and knowing the needs of companies and also classifying the best model that can be used and I was able to use these two methods to verify the validity of my research and also increase knowledge of modern technologies and how to use them in organizations, but one of the challenges that occurred in this research is how to integrate this information with each other
5. **Cross-sectional time: -** The process of collecting data, analyzing it, studying it and conducting experiments on it came at a fixed and clear time, which made the research in a stable process, but the problem with this methodology is the lack of development in modern technologies that appear over time

### 7.2 Alternative Research Methodologies

The selection of research methods presented here yielded excellent results in the process of classifying news and how media organizations can use artificial intelligence, but there are some other methodologies that can enhance the research process that I am doing in the future, which helps researchers like me to expand their knowledge more and better in their future research and to address new ideas with the discovery of new Methodologies.

1. **Action research:** It is applied when there is joint cooperation between the researcher and stakeholders. This leads to more efficient results and greater and deeper analysis and gives us a closer look at the problem in the work environment and his own research environment. For example, with regard to the process of classifying texts for news, going to a news satellite channel and working with stakeholders will lead to a deeper analysis of how texts are analyzed and how they work in their institutions.
2. **Longitudinal study:** In this methodology, the study is conducted over time periods to track the variables that may occur. This can improve the performance of models over time and modify them, as well as extract insights about the classified texts.
3. **Ethnography:** The process of collecting data, analyzing data, and monitoring people working in the media field can be especially important. Collecting more data about how institutions classify and conducting a comprehensive study about how to classify texts that contain several news items in the same text.

Throughout this journey that I took in studying the classification processes for news, there were lessons learned and new things that were gradually explored. Using the onion model was important for this research and for me, as it led to raising my skills in the research process. Three important things were noted in the research process.

1. Choosing **methodologies is very necessary**, as adapting to methodologies is very important for the researcher, while knowing the negatives and positives of the methodology, because it is the key to good research.
2. The process of **integrating different data sources**, such as Qualitative and quantitative data were excellent in making this research and it was a good opportunity to explore new ways of collecting data and information and this enhanced the research process in general
3. There are several ways that organizations can use to address their deficiency in the **news classification** process as mentioned with the experts and help them raise the level of the company

### 7.3 Recommended Actions and Future Considerations

After conducting research and understanding the need for institutions to classify news, some recommendations came to mind that some people in the media field recommended to follow, especially media companies and institutions, which are:

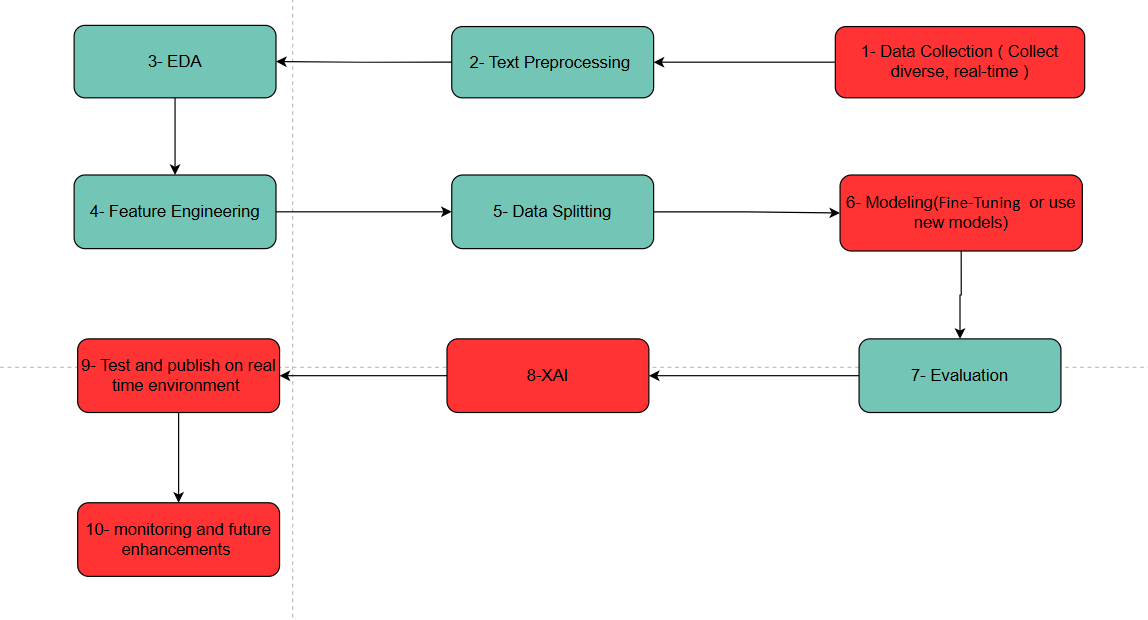
1. The media sector **must be developed** and artificial intelligence must be introduced to it so that it facilitates the affairs of companies and improves their operational efficiency.
2. Small companies and institutions can create **lightweight** models that are **low-cost** and do not require very high computing capabilities.
3. This is to help them in their operations and introduce artificial intelligence into them. **Integrating** artificial intelligence into institutions can provide greater accuracy and significantly reduce the time spent on the tasks that are used.
4. Media institutions can increase the **trustworthiness** of people in them by using artificial intelligence to **detect fake news**, and this increases the credibility of people in them.
5. Companies can make **recommendations** to people who prefer a certain type of news.
6. Artificial intelligence can be used to do **Sentiment analysis** in order to analyze the comments that are on the news on social media, and this gives them a general study of the opinions of the public.

With these points, I recommend that companies use artificial intelligence in them because they lead to increasing the trustworthiness of people It also leads to increasing its operational efficiency in various fields

Now I will recommend some recommendations to researchers and some procedures that I am thinking of doing in the future to increase the value and validity of research and to increase the development of skills in the research process and to conduct broader development in research operations

1. I recommend using some techniques and methods that can work to understand how this model can be classified based on what is called **xAI**. This can help us understand how the model works to classify this news and whether the classification process came correctly with the presence of several types of news within one news
2. I recommend conducting more and more **comprehensive interviews** with people who can give us new opinions and **not only** address the technical aspect with these people, but also go to people who work as businessmen in companies and not as technical people. This matter can give us greater knowledge and broader knowledge of research in general
3. Explore some **methodologies** that can be flexible and smoother and it is possible in the future to find some new and more advanced methodologies
4. I recommend researchers to try to use **real-time** data because most companies deal with real-time data and this can also give us a comprehensive verification of our models.
5. We can make a comparison with other models and try **several models** that can give us other results or researchers can also modify some models and change their structure.
6. Using **several data sources**, several types and several languages. This gives the researcher a stronger evaluation and a more comprehensive analysis of the models and in different scenarios.

### 7.4 Recommended Methodology



1. ***Data Collection:***

Collecting data from multiple sources and using real-time data to ensure that models work in real time and to ensure that they are tested in realistic scenarios and under different data sources, whether in multiple languages ​​or different types of data sources. This step helps us apply the results of the models in a real form in realistic scenarios for the model.

1. ***Text Preprocessing:***

Performing preprocessing steps such as removing unwanted words, splitting the sentence, tokenization, lemmatization. This step helps us enter the data in a clean and good form and ready to extract features to the project model.

1. ***EDA:***

Data analysis to understand the data environment, its distribution, extract insights from it, and know whether this data may have bias or not, and understand patterns and data distributions.

1. ***Feature Engineering:***

Extracting features using techniques such as TF-IDF. The researcher can use modern and more advanced techniques to compare models in different techniques such as using word2vec and glove, which are among the developed techniques. Which helps in the process of understanding sentences better

1. ***Data Splitting:***

The researcher can use several methods to divide the data, which can make the model robust and prevent overfitting

1. ***Modeling:***

At this stage, different models can be used that I have not tried, or researchers can change and manipulate the model's infrastructure. They can also experiment with using neurons and compare them with each other

1. ***Evaluation:***

After the modeling process, we must evaluate the model using measurement tools to verify whether the model is robust or not

1. ***Explainable AI (xAI):***

This technology can save us the process of understanding and explaining how the news classification process was carried out and why each news was classified into its category. This is one of the very important things that helps non-technical people understand whether the model is correct or not and understand the classifications on which the model made the decision

1. ***Testing and Publishing in Real-Time Environments:***

Testing the model after preparing it using real-time data and also until this model can be launched and used In the institution itself, it must be ready to receive data in real time. Monitoring and Future Enhancements: Monitoring the models must come continuously to make a comprehensive confirmation of the model and ensure that there are no problems with it, and then comes the process of future improvements and updates to it.

# 8.References

Wongso, R., Luwinda, F.A., Trisnajaya, B.C., Rusli, O. and Rudy (2017). News Article Text Classification in Indonesian Language. *Procedia Computer Science*, 116, pp.137–143.

DOI: https://doi.org/10.1016/j.procs.2017.10.039.[1]

Agarwal, J., Christa, S., Aditya Pai H, M. Anand Kumar and Prasad, G. (2023). Machine Learning Application for News Text Classification.

DOI: https://doi.org/10.1109/confluence56041.2023.10048856.[2]

‌

Zhang, M. (2021). Applications for Deep Learning in News Text Classification. *Scientific Programming*, 2021, pp.1–9.

DOI:https://doi.org/10.1155/2021/6095354.[3]

‌

Khan Md. Hasib, Nurul Akter Towhid, Kazi Omar Faruk, Jubayer Al Mahmud and Mridha, M.F. (2023). Strategies for enhancing the performance of news article classification in Bangla: Handling imbalance and interpretation. 125, pp.106688–106688.

DOI: https://doi.org/10.1016/j.engappai.2023.106688.[4]

‌

Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R. and Sattar, A. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. *Computers*, 12(1), p.16.

DOI: https://doi.org/10.3390/computers12010016.[5]

‌

Ahmed, J. and Ahmed, M. (2021). ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES. *IIUM Engineering Journal*, 22(2), pp.210–225.

DOI: https://doi.org/10.31436/iiumej.v22i2.1662.[6]

‌

Chen, W., Liu, B. and Guan, W. (2023). ERNIE and Multi-Feature Fusion for News Topic Classification. *Artificial Intelligence and Applications*.

DOI: https://doi.org/10.47852/bonviewaia32021743.[7]

‌

Mehta, D., Patel, M., Dangi, A., Patwa, N., Patel, Z. Jain, R., Shah, P. and Suthar, B. (2024). *Exploring the Efficacy of Natural Language Processing and Supervised Learning in the Classification of Fake News Articles*. [online] Ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4721587.[8]

‌

Bangyal, W.H., Qasim, R., Rehman, N. ur, Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z. and Ahmad, J. (2021). Detection of Fake News Text Classification on COVID-19 Using Deep Learning Approaches. *Computational and Mathematical Methods in Medicine*, 2021, pp.1–14.

DOI: https://doi.org/10.1155/2021/5514220.[9]

Byjus (2019). *Top 6 Difference Between Primary Data and Secondary Data*. [online] BYJUS. Available at: <https://byjus.com/commerce/difference-between-primary-data-and-secondary-data/>. [10]

Ferretti, J. (2023). *BBC Articles Dataset with Extra Features*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/jacopoferretti/bbc-articles-dataset [Accessed 16 Jan. 2025]. [11]

15writers (2019). *Understanding the Research Onion*. [online] 15 Writers. Available at: <https://15writers.com/research-onion/>. [12]

A. Melnikovas, “Adapting Research Onion Model for Futures Studies,” [https: //www.semanticscholar.org/paper/Towards-an-Explicit-Research-Methodology-% 3A-Adapting-Melnikovas/dccf54a5a4312ceb2261e1989cda01f73989d735,](https://www.semanticscholar.org/paper/Towards-an-Explicit-Research-Methodology-%3A-Adapting-Melnikovas/dccf54a5a4312ceb2261e1989cda01f73989d735) 2018, [Online; accessed 10-Jan-2024].[13]

www.run.ai. (n.d.). *Bayesian Hyperparameter Optimization: Basics & Quick Tutorial*. [online] Available at: <https://www.run.ai/guides/hyperparameter-tuning/bayesian-hyperparameter-optimization>. [14]

‌

Fathima, S. (2024). *LIME vs SHAP: A Comparative Analysis of Interpretability Tools*. [online] www.markovml.com. Available at: <https://www.markovml.com/blog/lime-vs-shap>. [15]

‌