National College of Ireland

Multilingual Text Analysis using Natural Language Processing and Transfer Learning.

MSc Research Project

Data Analytics

Jinal Jaisukh Sarvaiya

Student ID: x19207662

School of Computing

National College of Ireland

Supervisor: Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jinal Jaisukh Sarvaiya
Student ID:	x19207662
Programme:	MSc in Data Analytics **Year:** 2021-2022.
Module:	MSc in Research Project
Supervisor:	Hicham Rifai
Submission Due Date:	31 January 2022.
Project Title:	"Multilingual Text Analysis using Natural Language Processing and Transfer Learning"
Word Count:	…6152…… **Page Count**…20……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …… *Jinal Sarvaiya* ……

Date: …31 January 2022……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Multilingual Text Analysis using Natural Language Processing and Transfer Learning.

Jinal Sarvaiya
x19207662

Abstract

It only takes one toxic comment to sour the debate of internet. Trust and safety teams at major social media companies face this difficulty of examining the behaviour and activities such as online chats. Different lexicons and languages are used to make up range of settings which contains words or phrases of these exchanges. If these toxic comments and contributions to social media or web are identified, then we would have a much safer and more interactive contributions to internet. By using Natural Language Processing and Transfer Learning, we experiment by expansion of modelling capabilities with intentions of identifying toxicity on online comments and promoting trust and safety of the user. We'll examine the accuracies of several models, such as BERT and others, and derive insights into the effectiveness of pre-trained language models on categorizing multilingual text versus traditional methods. To help assist in the industry multilingual and multi-label classification model are developed which will help to assist industry in knowing its worth by various tactics.

1. Introduction

Thankful for the rise in e-commerce and social media platforms, including the increasing growth in internet and network connection technologies which makes customer interaction with online platforms by sharing their product experiences. Nowadays people use online consultation for evaluating hotel stay or making any decision. The above evaluations are full of details of customers experience. By analysing such data, the organizations can plan on their basic/ primary areas of improvement. It is allowing businesses to filter out large chunk of data from different resources in reliable manner which is done automatically. The results of goods and services are carried out based on feedback of the customers business. Before analysing raw inputs, it should be properly structured. Process of summing text into ordered categories is called text categorization.

The research attempts to answer the following question of our project-

How efficient are pre-trained language models on categorizing multilingual text compared to traditional models?

From discussion on online platforms which has one or more labels of category will have a look after how well pre- trained language models are. While evaluating results and assess test parameters like precision, recall, F1- score and accuracy. In last decade, active users have increased in millions. For example, Facebook has around 2.6 billion users around the globe which counts 1/3 rd of the world's population. There has been significant increase in ideas clashing as there is rapid growth of active internet users with downside like frauds and hate speech, which in turn has become dangerous for users safety. Toxic comments are being widespread on social platforms like Twitter and Facebook. Trust and Safety is known here as where due to business strategies there will be limit the danger of exposing the users with toxicity acts on online platforms.

Text analytics is becoming a more important aspect of businesses since it allows for quick access to data insights and the automation of business processes. Classifiers of text are used by Natural Language Processing which automatically analyses text by assigning list of predetermined tags which is totally based on content. An appropriate predefined label has been assigned to every feedback sentence in test dataset. To make task a multilabel classification, feedback must have multiple labels given to it. The investigate is centering on techniques that can categorize multilingual and multilabel input of client that are strong. Besides, how doable is it to utilize Machine Translation (MT) to change over non-English input into English and after that categorize it utilizing English-based frameworks, in case local frameworks for each of the various dialects are fundamental to dissect input. We went through few limitations caused by unbalanced data and by describing the approaches which are pre-processed when dealing with it. With a center on most later transfer learning strategies, we did tests on assessing information on run of methodologies, which has been on consideration from past two a long time. AI transfer learning approaches which have empowered generalizability in NLP frameworks by utilizing pretrained dialect models. These systems may perform a variety of tasks, including sequence labelling, text categorization, and natural language generation. A thorough examination of the many representations of transfer learning methodologies has been conducted.

2 Related Work

2.1 Analysing sentimesnts using Natural Language Processing

(Chitra et al. 2021) Customer reviews are collected digitally and analysis was performed using NLP algorithms, which are faster and more accurate. APIs and SDK resources exposed by top organisations can be used for such purposes. APIs from firms like IBM, Parallel Points, and others, as well as Twitter, Facebook, and WordPress, can all benefit from this input review process. As sentiment analysis aids all professional and non-technical organisations in interpreting client input on how the product fits, the result is in numerical form according to the category selected, such as emotions, feelings, taxonomy, and so on. The aim was to make it simple to understand for the user comfort. (Fujihira and Horibe 2020). The analysis of the textual data, extracting sentiment of each word with a dictionary, and of text-based word sentiments are 3 different processes mentioned in this article. German, Spanish, French, and

English sentiments have all been classified. The performance of the classifier has been evaluated by comparing it to the previous classifier using the evaluation standards "Accuracy," "F1 Score," "Precision," and "Recall." We proposed and compared a multi-linguistic sentiment analysis approach based on translating word to word along with a sentiment dictionary to classifiers "VADER" and "GCP." (2019, Ma) For classification, he employed the conventional BERT architecture as well as many alternative BERT architectures that were designed and trained to match with the baseline bidirectional LSTM with Glove Twitter pretrained embeddings. With a base score of 3.29 percent, and average F-1 score, the best outcomes were obtained with BERT andd BERT-based LSTM. Subjectivity as well as the Ambiguity both have a big impact on how well these models work. (Araújo *et al.* 2020)They are concentrating on evaluating activities that were planned to be done in a language-specific manner using a simple baseline. For 14 human-labelled languages, the report provides thorough quantitative analyses. Our findings show that instead of using the language-specific technique, simple translation of input information into English and then using an existing best method built for English might be preferred.(Galeshchuk *et al.* 2019) In this research, supervised sentimental analysis algorithms are applied to multi-lingual twitter data in three Slavic languages: Polish Croation and Slovenian. (Sproat 1996) The text analysis model has been described in this study with text to speech synthesis, which is used as a text analysis module in multilingual Bell Labs TTS systems using finite state transducers. These transducers are built up of lexical toolkits that allow for lexicon descriptions, morphological rules, and expansion rules, among other things. English, Spanish, Romanian, French, German, Japanese, Mandarin, and Russian are among the languages to which this paradigm has been applied. (Acikalin *et al.* 2020)Because Turkish NLP has inadequate resources, it becomes a serious research disadvantage. In order to address these restrictions, researchers have offered two solutions in this work. A) After converting Turkish to English, using the BERT model. B) the multilingual BERT model is fine-tuned. In this research, they examine the hotel and movie datasets to determine whether they are positive or negative. In a movie dataset, the models attain great accuracy, with BERT outperforming the previous model. An automated B.E.R.T model used in place of the multilingual B.E.R.T model. The texts are initially translated into English and then into English alone. It is suggested that BERT models be used, as they have been shown to yield results. (Shafin *et al.* 2020) The purpose is to develop a model that uses NLP to evaluate customer feedback of online purchases and provide a ratio of positive and negative remarks made in Bangla by previous customers (NLP). With an accuracy of 88.81 percent, SVM outperformed all other methods. (Ray and Chakrabarti 2017) The framework for sentiment analysis using R has been developed in this study, which analyses Twitter data utilising API. A lexicon-based approach is utilised to analyse each user's sentiment, and data from Twitter is also collected. Sentiment analysis is the computational management of user viewpoints, sentiments, and subjectivity in text.

2.2 Deep Learning and Machine Learning Techniques on text classification.

(Ramaswamy and DeClerck 2018) They're looking into unstructured data from social media, chat rooms, and voice recordings, among other sources. To expand its commercial opportunities, such as getting to know customers' needs, improving the product, and providing marketing advice. They've employed several natural language processing and deep learning techniques to aid in the analysis of more relevant data in order to keep track of consumer comments. (Dhyani 2017) Participants in this study experimented with simple neural architectures to assess multilingual consumer feedback. The performance of language-specific nats was assessed using Exact Accuracy (EA) and Micro-Average F1 (MAF) scores. (Ghorpade and Ragha 2012), They've done it. Different sentiment analysis are gradually appearing for decision-making after being analysed and categorised. Sentiment classification, automated survey analysis, opinion mining, and recommendation algorithms are all examples of text analysis in this work. (2017, Swarnalatha) They've helped customers buy books, and authors use Amazon user reviews and Twitter hashtags to conduct sentiment research on public opinion. To identify feelings, researchers employed NLP and ML techniques such as bags of words, MNB classifiers and the famous 'n'-gram approach. Users' emotions were categorised into three groups: 'neutral', 'positive' and 'negative'. (Vidhale et al. 2021) The foundation on handwritten digit recognition neural net-works have been created in this research. A charactor recognition MAT-LAB model as well as hand-written digits recognition model were employed in this study. The matlab model is used to obtain the outcome of various created languages' barrier. All of the predicted languages, including English, Marathi, and Gujarati, are transformed to English. Optical character recognition is used to transform languages written manually or written images of text in Indian languages (such as Gujarati or Marathi) into English. (Schultz 2014) Six languages were used in this research to train and test various recognition engines in monolingual, multilingual, and cross lingual configurations. Using a global phoneme set, we developed a multilingual speech recognition system that can handle five different languages. The acoustic models of the five languages are combined into a single system, and context-dependent phoneme models are constructed using language enquiries. This task force proposal included some design variations of an ML/CL voice recognition system. The task force will be broken down into smaller, more rounded pieces under the recommended method to make it as modular as possible. Algorithms from multiple partners can be integrated into the set after the baseline platforms are completed.

2.3 Users Product Feedback analysis

(Vidhale *et al.* 2021) used natural language processing (NLP) tools on this consumer market data and provided immediate in-sights on products' pros and cons. The models used provided rapid comprehensive and understandable data resources for gaining quick useful information. (Alibasic *et al.* 2021) has revealed insights from the consumer satisfaction of typical sources of airline service of TripAdvisor company with 50000+ reviews from 2016 to 2019. Here the

customer insights gives a better picture in analyzing various area of business. It provides a quick, easy-to-use, and comprehensive data resource for quickly getting insights. As the customer reviews are very important as they can help in changing or improvising the aspects based on the consumer feedback. Hence here to analyze all the data, NLP is used which is mixture of artificial intelligence and machine learning which helps in reviewing multiple feedbacks. (2017, Akella et al.) The challenges of mapping customer verbatim, which is also a natural language text classification problem, are described in this work. The verbatim from social media conferences, the quality office's transactional system, and other sources are employed. We employ conventional methodologies for document representation, such as word counts and TF-IDF. Because of the complexity of word relationships, we employ word embedding with word2vec to build document vector representations of text. Machine learning techniques are then used to classify customer comments into worry codes. We plan to use translation techniques to broaden the scope of this English comment classification system to additional languages. (Baydogan and Alatas 2019)They must conquer the challenge; the goal was to determine the client experience and satisfaction automatically. This research was conducted on six different airline businesses, using three labels: favourable, negative, and neutral. To process this data, NLP and machine learning are utilised, and the findings are displayed in tables and graphs. K-Nearest Neighbors, Random Forests, etc. were all investigated in depth in this study across all datasets, with SMO achieving the greatest accuracy of roughly 80% for three labels, indicating success in working with three datasets.

3 Research Methodology

This study used the CRISP-DM (Cross-Industry Standard Process for Data Mining) technique. As noted in Tie et al., it will be explained in the business understanding, data understanding, data preparation, and modelling sections below (2011).

3.1 Business Understanding

Social media have grown by millions in past decades. In 2021, over 2.6 billion users worldwide on Facebook with the speed of rising online users' significant growth in ideas and ideals clashing with some unexpected drawbacks like radicalization violence, hate-speech which has become threats to personal safety of users. On online discussing websites such as Facebook and Twitter, occurrence of toxic comments is very common. Trust and Safety are set of business strategies by decreasing danger of users on online platform by being exposed to toxicity or behaves outside of community norms. While promoting consumer acquisition, engagement and retention, which is increasingly becoming crucial function for the users who are seeking to protect their users. Technology professionals have been collaborating with Machine Learning researchers to build strategies for maintaining online safety users because one of challenges faced by social media companies is to determine toxicity in online conversations. We conduct this research by training Transfer Learning-based pre-trained classification models on a huge number of texts in multiple languages that have been rated as toxic, severe toxic, obscene, threat, insult, and identity hate by human rates. We'll examine the

accuracies of several models, such as BERT and others, and derive insights into the effectiveness of pre-trained language models on categorizing multilingual text versus traditional methods.

3.2 Data Understanding

This dataset is important aspect of this research, and it should be made to its ethical use. We have taken this dataset for our research project from Kaggle. This dataset contains around 2,23,548 rows in each CSV file of 6 different languages (Portuguese, Russian, French, Turkish, Spanish and Italian) wherein we have just considered 3 languages each with limited number of rows due to limitations in translating the comments into English. In this research, we have considered 3 languages- Spanish, French and Italian with limited rows from each dataset. This multilingual structure consists of 6 different categories such as severe_toxic, insult, toxic, obscene, threat and identity_hate. All null/missing values are removed from the dataset using Python in this step. The data has been cleansed and the unwanted columns have been eliminated. Because the data comprises text in multiple languages, pre-processing procedures include translation, tokenization, stemming, and lemmetization. Following graph representing the imbalance data.

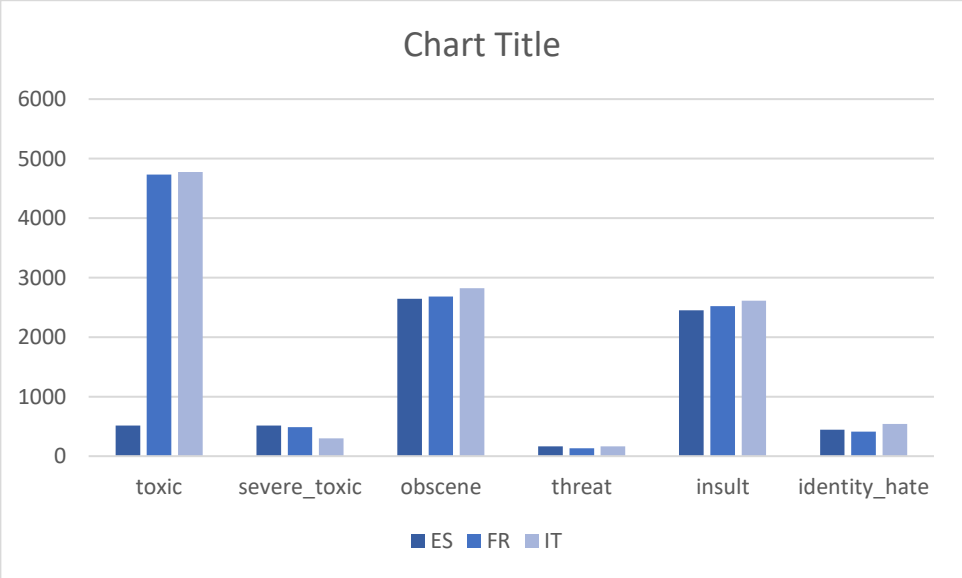


Figure 1: Data Statistics per language

3.3Data Preparation

3.3.1 Translation

Google Translator was used in this project wherein multilingual machine interpretation models and the techniques help in decoding one textual content to another which means converting all

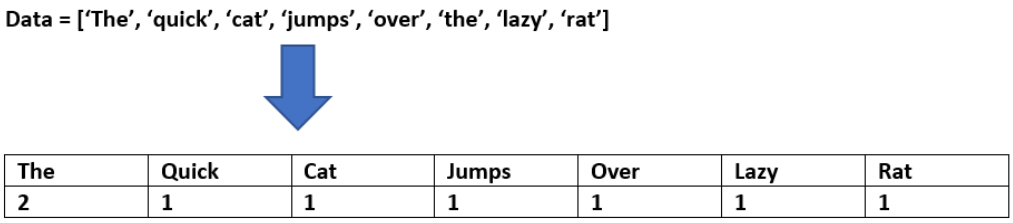
the different languages in the research into English. Google Translator API was used on each of the three language datasets; however, only partial data was translated due to the access limitations on the freeware API.

3.3.2 Lemmatization

Cleaning the data and extracting features were included as substantial pre-processing steps with traditional models. Part-of-speech tagger which is NLTK's recommended baseline also known as Perceptron Tagger in which the verbs, adjective, nouns etc is assigned to each word. It provides a dictionary of feature-weighted weights that is used to predict the proper tag for a set of features. During the training phase, the tagger guesses a tag and adjusts the weights based on whether the guess was correct. Lemmatization is preferred after stemming which can convert more words into more meaningful format instead of removes the last few characters leads to erroneous interpretations.

3.3.3 Count Vectorizer

This converts the text into vector which in terms of counts. Before converting into vector representation, it allows text to be pre-processed. It is the feature representation of text with lot of flexibility.



The	Quick	Cat	Jumps	Over	Lazy	Rat
2	1	1	1	1	1	1

Figure 2: Example for count vectorizer

3.3.4 TF-IDF Transformer

The Term Frequency — Inverse Text Frequency (TF-IDF) statistic seeks to better determine the value of a word in a document by taking into account its link to other documents in the same corpus. This can also be done by looking at the same term in other texts from the same dataset and counting how many times it appears in the document. Every word in the dataset is given a TF-IDF score. And each word's TF-IDF value rises with each appearance in a document but falls progressively with each appearance in other documents. Tfidf (w, d, D) = tf (w,d) * idf (w, D) Where, N: Number of documents present in dataset d: Current given document for dataset D: All document collection w: word given in document.

3.4 Modelling

We approached the project by using two techniques a which are mentioned in detail below. They are used in phase of model building and results are obtained and discussed in Evaluation section.

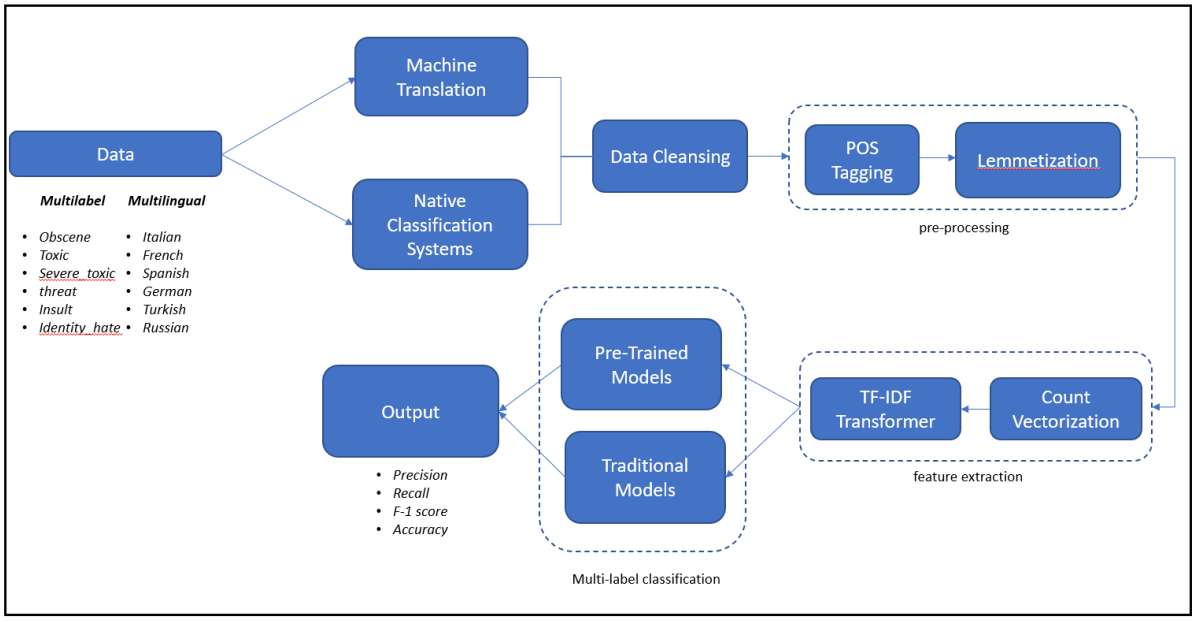


Figure 3: Design process flow

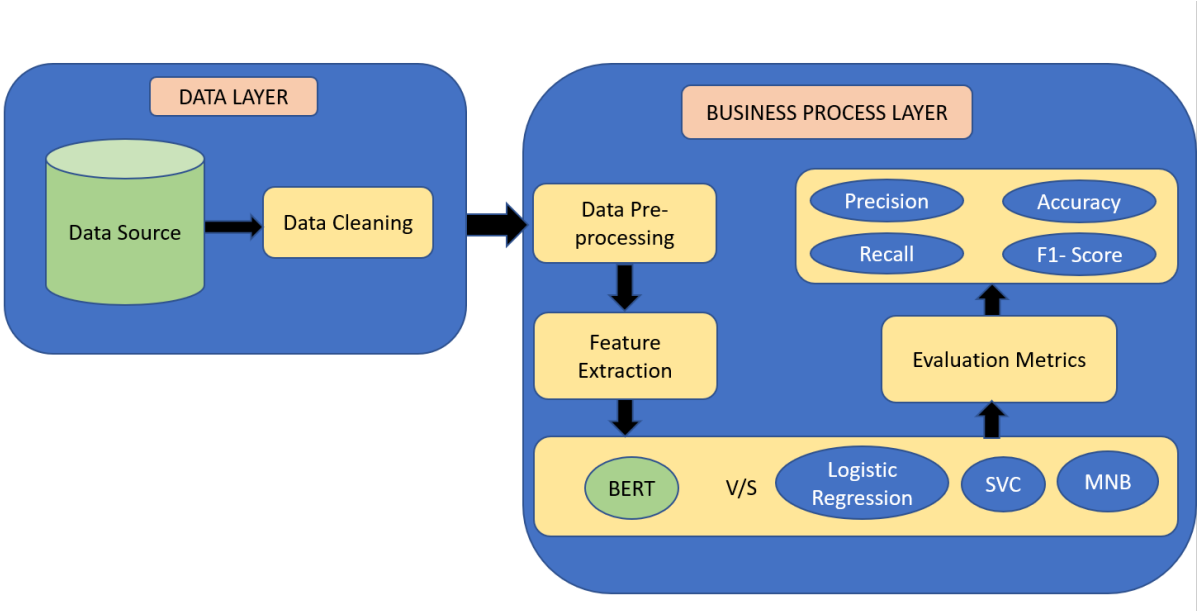


Figure 4: Architecture of Proposed Methodology

3.4.1 Binary Relevance Strategy

Binary Relevance, also known as One-vs-rest classifier is most instinctive method for solving problems of multilabel classification. Multilabel classification is divided down into many binary classification issues using this method. This is accomplished by assigning one classifier to each class, and the class competes with the other classes for each classifier. With MNB, SVC, and Logistic Regression classifiers, we employed SKLEARN's OneVsRestClassifier.

3.4.2 Traditional Classification Models

Traditional text classification predictive modelling involves labelling inputs data to group them into various classes. These are machine learning based models which do not comprise of neural networks. A model would use the training dataset and mathematical techniques to map examples of input data to a set of classification labels. The training set here must be representative of the population in a way that it needs to have enough examples of every class label to produce best accuracy. Logistic Regression Classifier, Multinomial Naïve Bayes and Support Vector Classifier are few examples of such traditional classification models.

3.4.3 Transfer Learning

We compared the results produced by advanced pre-trained models such as BERT with traditional techniques such as Multinomial Nave Bayes (MNB), Support Vector Classifier (SVC), and Logistic Regression, motivated by recent developments in Transfer Learning (TL), where pre-trained models are used as a starting point in natural language generation and classification tasks. Transfer learning is the process of improving learning in a new activity by transferring knowledge from a previously acquired related task. Model is developed is used to perform one task and later reused as starting point for other tasks in TL approach using deep learning. The mode stores knowledge which is gained from training one dataset and applies to another dataset of similar cases.

This, however, only applies to deep learning. If the characteristics learnt in the first task are universal. The contexts and linkages in our scenario are characteristics human communication. Word embedding is used for such problems, in which mapping of words where different terms of same meaning having similar vector representation into high dimensional continuous vector space. There are efficient algorithms for learning these distributed representations of concepts, and it is common for research organizations to release pre-trained models that have been trained on a large corpus of text texts under a permissive license. This is acted as optimization technique which saves on hours of training which produces better results.

1) Higher starting performance- Before optimizing the source model, the initial knowledge gain is greater than it should be.
2) Higher rate of improvement- During social model preparation, the rate of skill- enhancement is steeper than it would otherwise be.

3) Higher asymptote of curve- Trained model with converged skill is better than it otherwise would be.

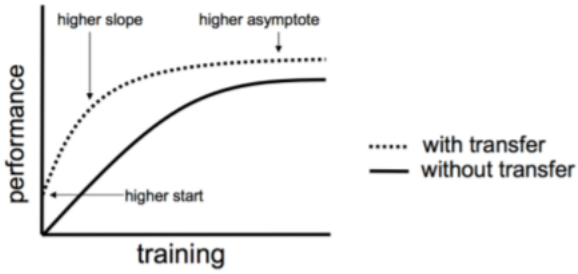


Figure 5: Improved learning through TL

4 Data Specification

The architecture is depicted in figure form in this section. The architecture is two-tiered and consists of a data layer and a business process layer. Tier 1: Data Layer- The data has been cleaned and sent to the next tier from the data source at this layer. Tier 2: Business Layer- This is the most important layer, since it specifies all data pre-treatment, feature extraction, model evaluation, and evaluation methodologies. Finally, in terms of accuracy, precision, recall, and F1-score, the output was compared to established machine learning and NLP methodologies in Figure 4.

5 Implementation

5.1BERT

Bidirectional Encoder Representations from Transformers (BERT) is multilingual transformer-based model, created and published in 2018 by Jacob Devlin and his colleagues at Google. It was pre-trained on BookCorpus and English Wikipedia was first release of English language. The open-source implementation which is original and is based in Tensorflow and later compensated by PyTorch releases. In this project, we have used BERT base uncased variant with consists of components of 12 transformed blocks or layers, 12 attention heads and pretrained on millions of parameters. Using words and both the directions, each word is contextualized by representations of pretrained bidirectional built-up context. It becomes easier by eliminating 15% of input words and running the entire sequence by encoder which is deep bidirectional transformer and later predicting only words which are masked.

INPUT: she likes to read [mask1], her favourite [mask2] is romance.
LABELS: [mask1] = books; [mask2] = genre

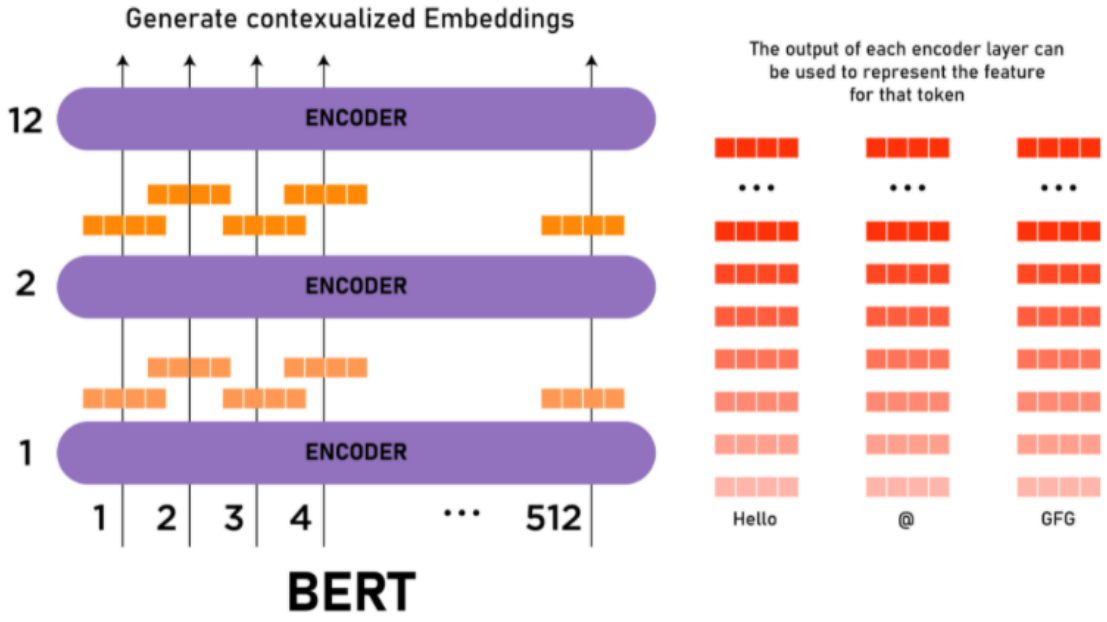


Figure 6: BERT Architecture

Any monolingual texts are generated from any tasks which are straightforward and trained by sentences to learn relationship between them by BERT. Full Tokenizer have been used as BERT's default that implements strategy of WordPiece tokenization. Most likely variations of the words which are currently used in vocabulary are inserted and the vocabulary is initialized with all individual characters in language. This method is useful for dealing with 'unusual words' in a text. In the final model, we used a maximum sequence length of 128 tokens; nevertheless, the BERT base uncased gives a maximum sequence length of 512 tokens. No layers are frozen during fine adjustment in default settings. The task-specific parameters are applied to both pre-trained layers at the same time. There is same learning rate which are tuned with all the parameters and for our model we have hyperparameters tunned with 0.5 dropout, batch size of 16, 2e-5 learning rate and 3 epochs.

5.2 Multinomial Naïve Bayes (MNB)

Rather than the precise distribution of each variable, the name Naive Bayes refers to the model's unambiguous expectations of freedom. A Naive Bayes model assumes that each of the features it employs is conditionally independent of the others for any given class. Under the Naive Bayes assumption, the following holds for calculating the probability of observing features f1 through fn for any class c.

$$p(f_1, \ldots, f_n|c) = \prod_{i=1}^{n} p(f_i|c)$$

As a result, working with the posterior probability is significantly easier when using a Naive Bayes model to categorize a new case:

$$p(c|f_1,\ldots,f_n) \propto p(c)p(f_1|c)\ldots p(f_n|c)$$

These independence assumptions are rarely fulfilled, yet Naive Bayes models have done surprisingly well in practice, even on difficult tasks when the strong independence assumptions are clearly untrue. The MNB classifier is a variant of the NB classifier that use a multinomial distribution for each of the characteristics. The name MNB simply means that each p(fi) is a multinomial distribution, which is useful for data that can be easily converted into numbers, such as word counts in text or fractional counts like tf-idf.

5.3 Support Vector Classifier

The technique is built from the ground up for binary classification problems, and it tries to paint a linear relationship between two classes to make the difference between them as large as feasible. The SVM algorithm categorizes the data points in an N -dimensional space by the determination of a hyperplane. This results in distinct and effective identification of data points. The SVM algorithm creates a virtual boundary for determining which vectors belong to a particular group or category and groups them accordingly. SVM algorithms can be typically applied to all forms of data that involve some kind of encryption. When SVM is used for text classification, the first step is to perform the transformation of text blocks into vectors. In this research, binary relevance method using One-vs-the-rest (OvR) multilabel strategy was employed to transform the text into meaningful vectors. The Grid Search technique is a method to perform Hyperparameter tuning wherein cross-validation is performed and hence evaluation of the value to be set for hyperparameters that offer optimum accuracy is also done. Through the Sci-Kit pipeline utility the following tasks were completed-Data transformations, hyperparameter tuning, model training, tf-idf transformation.

5.4 Logistic Regression

Comparison of input features were used for logistic regression and then later they were passed through sigmoid function i.e it converts every real number input into binary number between 0 and 1. This ensures that the result is always between 0 and 1, as the numerator is always one less than the denominator. In a Logistic regression model, the GridSearchCV method was utilized to do hyperparameter tweaking. There are no important hyperparameters to tweak in logistic regression. To tune, 'C' and'max iter' were employed. In our trials, we also used this model in conjunction with the binary relevance technique. Probabilistic categorization was used for discrimination of model in Logistic Regression. In logistic regression, the horizontal value represents the value of x and vertical one has likelihood for existing classification. y|x is supposed to be a Bernoulli distribution assuming distribution function. Sigmoid function is used in logistic regression’s formula.

6 Evaluation

Classification algorithms in machine learning may use various metrics to evaluate the output. We used the major ones such as accuracy, precision, recall and f1-score on the overall set of outputs averaged across the class labels. The metrics are computed based on the true positives, false positives, true negatives and false negatives.

In this case the positives and the negatives refer to the predicted output vs the actual output, as explained below.

a) when output is negative and is correctly predicted negative, then True Negative (TN)
b) when output is positive and is correctly predicted positive, then True Positive (TP)
c) when output is positive and is incorrectly predicted negative, then False Negative (FN)
d) when output is negative and is incorrectly predicted positive, then False Positive (TP)

As per the general convention in text classification, accuracy is not considered the best parameter to evaluate the results, we computed F1 scores on a micro level.

F1 macro-averaged scores take into account the results independently for each label taking the average by treating every label equal. On the other hand, F1 micro-averaged score aggregates the efforts towards the result from all labels in a weighted manner. Micro-average is hence preferred in our multi-label classification task. It also takes care of class imbalance in the dataset, however, that was not a challenge in our study.

6.1 Results based on F1 score

We computed F1 micro-averaged scores for evaluation even though we did not face any class imbalance in our experimentation. All the models have been trained in a way that all the sets of data are equally proportioned across labels.

Language set	Logistic Regression (LRC)	Multinomial Naïve Bayes (MNB)	Support Vector Classifier (SVC)	BERT
All	75.01 %	76.89 %	**81.26 %**	79.97 %
Spanish	82.39 %	77.45 %	**87.51 %**	74.41 %
French	70.76 %	72.21 %	76.12 %	**78.22 %**
Italian	77.69 %	78.82 %	79.96 %	**80.67 %**

Table 6.1: Classification results (F1 scores)

Table 6.1 shows metrics produced in the evaluation step. As a general consensus based on the literature that we previously reviewed, F1 micro-averaged scores were considered apt for text classification evaluation can be calculated as

$$\text{F1 Score} = \frac{1}{\frac{1}{2}\left(\frac{1}{Recall} + \frac{1}{Precision}\right)} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

6.2 Results based on Accuracy

Accuracy measure may not be as representative of the output of the classification model as F1 scores. However, for a generic model evaluation metric we computed and parked the scores. Table 6.2 showcases the accuracy scores for each model in our study and are derived as the total count of correct predictions divided by the total count of predictions.

Language set	Logistic Regression (LRC)	Multinomial Naïve Bayes (MNB)	Support Vector Classifier (SVC)	BERT
All	83.44 %	84.48 %	88.13 %	85.52 %
Spanish	87.22 %	79.72 %	91.67 %	81.65 %
French	78.06 %	87.28 %	84.22 %	79.29 %
Italian	81.13 %	83.74 %	85.84 %	87.32 %

Table 6.2: Classification results (Accuracy)

6.3 Discussion

Table 6.1 provides an overall comparison of the models both traditional as well as pre-trained. As pictured initially while studying pretrained models based on transfer learning, their overall performance was expected to be higher than the traditional ones, which is partly evident by the results in table 6.1. On the combined set, SVC performed surprisingly better not only than the traditional but also compared to the pretrained model, although the F1 score is only slightly higher (~ 1.5 percent).

BERT performed well over the combined dataset in just about 5 epochs, although the overall computation time was much higher compared to LRC, MNB and SVC. Results came out to be better than LRC and MNB.

On the individual language sets, evaluation results show comparable performances of SVC and BERT across the three sets (as shown in table 6.1). For Spanish corpus, SVC produced highest scores among all the models while BERT performed best on French and Italian corpora. Scores on individual sets are higher than the overall scores when the best model is chosen for a set. This is perhaps due to limited sample size and reduced error rate in model training.

7 Conclusion and Future Work

Our study produced expected output on the machine translation aspect, where only limited data was translated using the freeware google APIs over weeks of efforts in batched translation runs. Overall final translated set of three languages was found to be a well enough sample for our study.

One-vs-rest classification strategy worked well for multi-label classification, as a future scope we would like to experiment with models which work directly on the multi-label aspect of the problem, eliminating the additional step of bringing in the One-vs-rest technique.

On the evaluation end, BERT as expected performed better than most traditional models, however, SVC outperforming every other model was something unusual to our expected findings. This was a powerful finding in our experimentation. We are yet to research deeper on the underlying SVC mathematical model to better understand what made it perform better on our dataset. One possibility while comparing BERT and SVC in our case is we used the basic parameters for BERT and skipped most pre-processing due to limited free GPU availability. BERT might outperform the traditional model on combined as well as the Spanish dataset if sufficient pre-processing and hyperparameter tuning is considered, as a base for any machine learning model in production in an enterprise environment. We would keep it as part of our future work.

Overall we found sufficient evidence to answer our research question, "*How efficient are pre-trained language models on categorizing multilingual text compared to traditional models?",* the results of our study show that pre-trained models are powerful compared to traditional classification models since a basic version of BERT could produce comparative output without any preprocessing and hyperparameter tuning on the same dataset which required POS tagging and Lemmetization in case of Logistic Regression, MN Bayes and Support Vector Classifier.

Such pre-trained models are recommended to be refined and made production ready on very large data enabling the detection of toxicity falling outside the community norms to protect users' Trust and Safety in social media platforms.

8 Acknowledgement

I would like to convey my sincere gratitude to Hicham Rifai for mentoring me through the research proposal and addressing my queries and concerns in depth. I'm also grateful to my parents and friends for their continuous support and encouragement during the completion of my research proposal report.

Reference

A. Alibasic and T. Popovic, "Applying natural language processing to analyze customer satisfaction," *2021 25th International Conference on Information Technology (IT), 2021, pp. 1-4, doi: 10.1109/IT51528.2021.9390111.*

Acikalin, U.U., Bardak, B., Kutlu, M. (2020) 'Turkish Sentiment Analysis Using BERT', in *2020 28th Signal Processing and Communications Applications Conference, SIU 2020 - Proceedings*, Institute of Electrical and Electronics Engineers Inc.

Akella, K., Venkatachalam, N., Gokul, K., Choi, K. & Tyakal, R. (2017), 'Gain customer insights using nlp techniques', *SAE International Journal of Materials and Manufacturing 10(3)*, 333–337.

Araújo, M., Pereira, A., Benevenuto, F. (2020) 'A comparative study of machine translation for multilingual sentence-level sentiment analysis', *Information Sciences*, 512, 1078–1102.

Baydogan, C., Alatas, B. (2019) 'Detection of Customer Satisfaction on Unbalanced and Multi-Class Data Using Machine Learning Algorithms', *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings*, 1–5.

Chitra, P., Karthik, T., Nithya, S., Jacinth Poornima, J., Srinivas Rao, J., Upadhyaya, M., Jayaram Kumar, K., Geethamani, R. & Manjunath, T. (2021), '*Sentiment analysis of product feedback using natural language processing*', Materials Today: Proceedings

Dhyani, D. (2017) 'Exploring Neural Architectures for Multilingual Customer Feedback Analysis: IJCNLP 2017 Shared Task', available: http://arxiv.org/abs/1710.06931.

Fujihira, K., Horibe, N. (2020) 'Multilingual Sentiment Analysis for Web Text Based on Word to Word Translation', *Proceedings - 2020 9th International Congress on Advanced Applied Informatics, IIAI-AAI 2020*, (978), 74–79.

Galeshchuk, S., Qiu, J., Jourdan, J. (2019) 'Sentiment Analysis for Multilingual Corpora', (January), 120–125.

Ghorpade, T., Ragha, L. (2012) 'Featured based sentiment classification for hotel reviews using NLP and Bayesian classification', *Proceedings - 2012 International Conference on Communication, Information and Computing Technology, ICCICT 2012*, 1–5.

Ma, G. (2019), Tweets classification with Bert in the field of disaster management, Stanford University, (February), pp. 1-5.

Ramaswamy, S., DeClerck, N. (2018) 'Customer perception analysis using deep learning and NLP', *Procedia Computer Science*, 140, 170–178, available: https://doi.org/10.1016/j.procs.2018.10.326.

Ray, P., Chakrabarti, A. (2017) ‘Twitter sentiment analysis for product review using lexicon method’, *2017 International Conference on Data Management, Analytics and Innovation, ICDMAI 2017*, 211–216.

Schultz, T. (2014) ‘Multilingual and Crosslingual Speech Recognition MASPER – Multilingual and Crosslingual Speech Recognition MASPER – Multilingual and Crosslingual Speech Recognition COST 278 task force proposal Andrej Žgank , Zdravko Ka i University of Maribor , Slovenia ’, (August 2000).

Shafin, M.A., Hasan, M.M., Alam, M.R., Mithu, M.A., Nur, A.U., Faruk, M.O. (2020) ‘Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language’, *ICCIT 2020 - 23rd International Conference on Computer and Information Technology, Proceedings*, 19–21.

Sproat, R. (1996) ‘Multilingual text analysis for text-to-speech synthesis’, *International Conference on Spoken Language Processing, ICSLP, Proceedings*, 3(November 1996), 1365–1368.

Swarnalatha, P. (2017), ‘Analyzing customer sentiments using machine learning techniques’, 8, 1829–1842.

Vidhale, B., Khekare, G., Dhule, C., Chandankhede, P., Titarmare, A., Tayade, M. (2021) ‘Multilingual Text Handwritten Digit Recognition and Conversion of Regional languages into Universal Language Using Neural Networks’, *2021 6th International Conference for Convergence in Technology, I2CT 2021*, 4–8.