Sentiment Analysis of the Impact of AI Implementation on the Job Market

Kevin Bernardi Marte
Department of Computer, Information Science, and Mathematics
University of San Carlos
Cebu City, Cebu, Leyte, Phillipines
18200054@usc.edu.ph

***Abstract*—The advancement in the fields of AI and its increased capabilities has caused major changes in the job markets as companies see such tools as a boon to increase productivity while potentially reduce manpower costs and hassle, while many of the job workers and job seekers fear such advancement will change the job market and work condition for the worse.**

This study aims to explore the public sentiment, specifically within the English-language YouTube community, regarding the impact of artificial intelligence (AI) on job opportunities across various industries. The study analyzed comments from selected YouTube videos to measure the overall sentiment and and visualize the overall sentiments of the involved parties with this development. Comments is first scraped from the selected videos to form the basis for the main dataset. The main dataset is then manually labelled and undergoes text preprocessing procedures before being split into the training and testing dataset. The training dataset is then subjected to random undersampling due to the initial data imbalance between the positive and negative labelled comments, as well as feature extraction (TF IDF Vectorizer) and selection (chi2 test) methods. The training data is then be used to created the Naive Bayes classifier model and the model is then used to predict the labels of the testing data as well as compare the result of the classifier to that of the manual label. And a 10 - fold cross folding is used to determine the effectiveness of the class in identifying and labelling new, unseen data. 1,117 YouTube comments were used in this research with an initial ratio of 9 negative comments : 3 neutral comments : 2 positive comments.

Using Multinomial Naive Bayes with TF-IDF vectorizer feature selection, Chi2 test feature selection, and 10 - fold cross folding validation, the model achieved a macro average of accuracy 0.56, precision 0.37, recall 0.36, f1-score 0.36 with an average 10 fold cross-validation score of 0.614. The model performed well in identifying negative comments, due to majority of the comments in the dataset consisting of negative comments, while positive performed worse due to the small amount of positive comments found in the dataset. Neutral identification performance performed the worst, probably in due part also to the small amount of neutral comments in the dataset, as well as the fact that the neutral comments tend to incorporate features that usually are found in positive or negative comments, thus increasing the likelihood of neutral comments being falsely labelled as either positive or negative.

Based on this research, the majority of the YouTube comments in this shows negative sentiments towards the implementation of AI in the workplace while a few showed support for such development.

***Index Terms*—TF IDF, Chi Square, Sentiment Analysis, Naive Bayes, Python**

I. Introduction

This chapter will provide the the rationale of the research, problem statement and objectives, the significance of the study, and the scope and limitation of the research.

A. Rationale of the Study

The integration of artificial intelligence (AI) technology into modern society has caused a major transformation in the employment landscape. This shift has produced a variety of public sentiments across various industries, as AI's growing capabilities and adoption reshape various professional realms. As many navigate the job market of the current decade, the expanding influence of AI has given rise to both optimistic and pessimistic reactions from those directly impacted by its proliferation.

A real life manifestation of this paradigm shift can be observed in the realm of entertainment, particularly the Hollywood industry. The deployment of AI to recreate the youthful visages of iconic actors, exemplified by the rejuvenation of Harrison Ford in the latest Indiana Jones movie and Mark Hamilton as Luke Skywalker in the Mandalorian, reflects just one side of AI's impact. Moreover, concerns about AI-generated scripts and storylines, driven by cost-cutting measures, culminated in the 2023 SAG-AFTRA strike. Concurrently, reports of AI-driven chat bots replacing content and copy writing jobs, as well as IBM's plan to automate back-office functions, have become profound examples of AI's influence on the job market today that cultivates mostly pessimistic reactions among workers in the affected fields(Fortune, 2023; BBC, 2023; Business Insider, 2023).

Yet, amidst these challenges, AI can bring positive transformations in the job market. Research in the IT and CS sectors suggests that while automation may displace certain roles, it concurrently results in productivity gains and opens avenues for new opportunities in machine learning, natural language processing, and data science (Sakib et al., 2023). In human resource management, the introduction of AI automation presents a dual-edged sword, addressing misalignment in talent acquisition while posing challenges such as data security concerns and potential employee dissatisfaction. However, the benefits include enhanced efficiency, fair salary calculations, and a focus on developing valuable soft skills (Zhu, 2021).

This paper aims to explore of public perceptions surrounding AI's proliferation in the job market. Diverging from common conventional sentiment research relying on typical platforms such as Twitter, review sites, and other social media platforms, the study looks into the comments section of YouTube videos discussing the topic. Leveraging TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and chi-square testing, the relevant features extracted will be input into a Naive Bayes Classifier. The choice of Naive Bayes stems from its simplicity, ease of evaluation, and feature independence, rendering it resilient to noise . The research aims to determine the sentiments within a section of the YouTube community by scrutinizing comments, employing a Naive Bayes Classifier with TF-IDF vectorization and chi-square testing

B. Statement of The Problem

This study aims to evaluate the sentiment prevalent within the English language YouTube community represented by the selected videos regarding the impact of AI on the job market on different industries by achieving the following objectives :

- Determine YouTube videos that will serve as the source of the datasets for the research and scrape a certain amount of comments off each video
- Apply relevant text-preprocessing techniques to the datasets to eliminate irrelevant comments and noises from relevant comments.
- Design a model based on the chosen classifier to be trained in the manually labelled dataset and label the test dataset using the trained classifier model.
- Evaluate the result of the model by displaying the label distribution of the dataset labeled by the model and the performance of the model-labeled dataset with the true label.

C. Significance of the Study

The significance of this study is important to key stakeholders in the field of the job market who might be looking for concrete data on their peer's reaction to AI proliferation, specifically the individuals who are directly impacted by automation on the job-seeking side of the job market, that includes:

- **Job-seeking individuals**
  Job seekers can gain valuable insights into the collective sentiments expressed by the public, particularly within the English-language YouTube community, concerning the impact of AI on the job market. This knowledge enables them to make more informed decisions as they navigate the evolving job market and prepare them for potential up-skilling or shift to a different industry as a response
- **Individuals in Careers Affected by AI**
  Those whose careers are directly influenced by AI technologies will benefit from a better understanding of the sentiments and concerns of their peers and the broader public discourse on this subject. It allows them to also make informed decision on whether to upskill themselves or shift to another field
- **Researchers in the Field**
  This research serves as a solid reference for researchers interested in investigating the same topic using alternative methods and implementations. It provides insights into sentiment analysis techniques, data collection, and the utilization of YouTube comments as a valuable source of information and hopefully encourages them to utilize other public social media as sources of their dataset such as YouTube and Reddit.

By focusing on these key beneficiaries, this study offers a more concise and purpose-driven understanding of its significance within the context of AI's influence on the job market.

D. Scope and Limitation

In this study, the primary objective is to measure the sentiment expressed in English-language YouTube comments related to the impact of AI on the job market using Naive Bayes classifier. To achieve this objective, the researcher aims to select and analyze English-language YouTube videos specifically discussing this topic as the source of their dataset. He only considers videos starting from 2023, as that was the year the AI boom started. These comments will serve as a proxy for understanding the voice of the English-speaking public within the scope of this research.

II. Methodology

This chapter of this research proposal describes the methods and steps that are used to address the objectives of the study.

A. Data Extraction

YouTube was chosen due to the ease of gathering comments for dataset for free, the fact that each video has its own distribution of positive and negative comments and discusses the topic in a positive, neutral negative light, and the platform is accessible to nearly all regions globally. The videos that will be scraped will be selected based on the following criteria :

- **Content / Topic**
  Also known as topicality, it refers to the degree to which a video matches the user's query or interest, based on the video content and the user's information need or goal. It is generally the main starting point for video selection (Yang Marchionini, 2004) and is usually considered the most dominant criterion (Albassam Ruthven, 2019; Ondego Komlódi, 2017).
- **Author / Authorship / Familiarity**
  It refers to the creator or producer of the video, and what credentials, reputation, or style they have, while familiarity refers to how knowledgeable or interested a viewer is with regards to the topic or the creator of a video (Yang Marchionini, 2004). It is usually regarded as another major criterion in video selection or relevance as viewers are more likely to watch videos from sources they are already familiar with or sources

with reputable background (Albassam Ruthven, 2019; Ondego Komlódi, 2017)

- **Date / Recency**
  It refers to how recent or up-to-date the video is, or how closely it matches the user's temporal interest or need. It is an important criterion for viewers who places value on the time context of their topics or videos, such as film history research, news videos (Yang Marchionini, 2004), and technological advancements.
- **Language**
  It refers to the the linguistic system or code that the video uses to communicate its content and meaning. It is an important criterion as it can affect decision making and judgement based on the language context of the viewer and the video itself (Ondego Komlódi, 2017) , with a significant amount of mentions to be considered a video selection criterion (Albassam Ruthven, 2019)

Comments from videos that fits the aforementioned criteria is then scraped using YouTube API and a custom-built Python Scraper based around that API. 5,433 comments were initially obtained during this process, but due to the extreme imbalance of the dataset towards the negative sentiment, negative comments were removed to make the dataset balanced, resulting in a final dataset size of 1,632 comments with a a sentiment ratio of 1:1.

B. Data pre-processing

The comments then undergoes further data pre-processing to reduce noises in the collected data. The dataset is first removed of any emoticons, bot, empty, or non-essential comments, urls, tagged usernames, non-ASCII characters, and non-English comments. Then any HTML character codes is replaced with its ASCII equivalent. Next, any contractions or abbreviations, formal or informal, is replaced by their appropriate non-contracted or non-abbreviated counterpart. Furthermore, any grammatical error or spellings is corrected.

Finally, lemmatization and tokenization is applied to the cleaned data . lemmatization transforms any English text into its root form (ie. "Stealing" would turn to "steal" after lemmatization), while tokenization splits a given sentence or text into smaller fragments (ie. Tokenizing "I love tocino silog" will result in ["I", "love", "tocino", "silog"].).

C. Data Annotation

Manual Data Annotation is performed on the 1,117 comments. A human annotator assisted by lexicon labelling is used to speed up and correctly annotate the sentiment of each comment. After annotation, the dataset is split into training and testing dataset with a train : test ratio of 7:3, resulting in 1,163 training data, and 499 testing data.

D. Feature Extraction and Selection

TF-IDF Vectorizer, or Term Frequency-Inverse Document Frequency Vectorizer, is the feature extraction method selected for this research. It is a feature weighting measure used in text mining and information retrieval systems where it quantifies the importance of a word in a document which is part of a corpus. It consists of :

- **Term Frequency**
  This measures how frequently a term (t) occurs in a document. The operation is represented by the following formula :
  TF(t) = number of times term t appears in a document / total number of terms in a document
- **Inverse Document Frequency**
  Inverse Document Frequency : This measures how important a term (t) is. The operation is represented by the following formula :
  DF(t) = log(total number of terms in a document / number of times term t appears in a document)

Chi-squared test is used to select the top 500 features of the vectorized features. It is a statistical hypothesis based on the Pearson distribution test that is used to determine whether there is a significant association between two categorical variables. It is represented by the following formula :

$$X^2 = (Oi–Ei)2/Ei \tag{1}$$

Where Oi is the observed value and Ei is the expected value.

E. Sentiment Analysis

This Study used Multinomial Naive Bayes as the classifier model of choice. It is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features of a sample and gives the feature with the highest probability as the output , which works well for data which can easily be turned into counts, such as word counts in text. Thus, it is suitable for text classification with discrete features. Furthermore, a two-class approach, which classifies data into either positive or negative sentiment, will be used since according to research, Multinomial Naive Bayes classifier performs the best with the approach (Vaseeharan and Aponso, 2020).

F. Pipeline Parameter Tuning

GridSearch is then applied to the training data in order to find the most optimal parameters for the pipeline to work properly. The pipeline in question consists of the Feature Extraction and Selection Methods, and the Classifier Model

G. Model Evaluation

An accuracy metrics will be used to display the accuracy performance of the classifier with regards to the comparison of the generated labels to its true label counterpart. The following metrics shall be used in the research :

- **Accuracy**
  Accuracy simply measures how often the classifier correctly predicts. It is defined as the ratio of the number of correct predictions and the total number of predictions.
- **Precision**
  Precision explains how many of the correctly predicted cases actually turned out to be positive .

- **Recall**
  Recall explains how many of the actual positive cases the model was able to predict correctly
- **F1 Score**
  F1 Score gives a combined idea about Precision and Recall metrics, in other words, it is the harmonic mean of precision and recall used to measure a model's accuracy on a dataset .

III. Evaluation of Results

Based on the label distribution and the content of the training dataset, the following pipeline parameters produce the best result based on accuracy :

```
# Current Setup
pipeline.set_params(
    chi2__k=1000,
    nb__alpha=2,
    nb__class_prior=None,
    nb__fit_prior=True,
    tfidf__max_df=0.5,
    tfidf__ngram_range=(1, 1),
    tfidf__smooth_idf=False,
    tfidf__sublinear_tf=False,
    tfidf__use_idf=True
)
```

Fig. 1. Best Pipeline Parameters based on accuracy

The following parameters, when fed with the training data, can perform up to 0.75 accuracy.

After feeding the training data to the pipeline based on the aforementioned parameters, and feeding the test dataset to be labeled by the model, the prediction came out to 280 of them labeled as negative, and 219 labeled as positive, the performance of each respective sentiment is presented on the figure below when the predicted test labels are compared to their actual sentiment label :

	precision	recall	f1-score	support
Negative	0.71	0.79	0.74	252
Positive	0.75	0.67	0.71	247
accuracy			0.73	499
macro avg	0.73	0.73	0.73	499
weighted avg	0.73	0.73	0.73	499

Fig. 2. Performance report of trained MNB model

Based on the Accuracy Report, the model performs fairly well with a balanced accuracy across both classes. However, it is slightly better at identifying the 'Negative' class than the 'Positive' class because there are slightly more negative comments to train on than positive ones. However, this is achieved at the cost of significantly reducing the number of negative comments and thus, the size of the dataset compared to the original dataset count to make the dataset balanced and not skew towards the negative sentiment.

Further research in this topic involves trying to get similar or better performance with the original dataset, which is most likely to be skewed towards the negative sentiment. Possible approaches that can be taken to solve this issue include implementing SMOTE or Undersampling to balance the training dataset, Using other word embedding methods such as Word2Vec, implementing more text preprocessing methods like POS tagging, exploring other hyperparameter tuning methods, and integrating feature engineering for that specific topic.

References

[1] Dalton, A. (2023, July 24). Writers strike: Why A.I. Is such a hot-button issue in Hollywood's labor battle with SAG-AFTRA. Fortune. https://fortune.com/2023/07/24/sag-aftra-writers-strike-explained-artificial-intelligence/

[2] Mancini, J. (2023, August 14). IBM Plans To Replace Nearly 8,000 Jobs With AI — These Jobs Are First to Go. Business Insider. https://www.businessinsider.com/chatgpt-openai-ai-replacing-jobs-content-writer-2023-6

[3] Nolan, B. (2023, June 5). Content writer says all of his clients replaced him with ChatGPT: 'It wiped me out'. Business Insider. https://www.businessinsider.com/chatgpt-openai-ai-replacing-jobs-content-writer-2023-6

[4] Rose, I. (2023, June 16). The workers already replaced by artificial intelligence. BBC. https://www.bbc.com/news/business-65906521

[5] Sakib, N., Anik, F. I., Li, L. (2023). ChatGPT in IT Education Ecosystem: Unraveling Long-Term Impacts on Job Market, Student Learning, and Ethical Practices. ACM. https://doi.org/10.1145/3585059.3611447

[6] Zhu, H. (2021). Impact of Artificial Intelligence on Human Resource Management and Its Countermeasures. ACM. https://doi.org/10.1145/3495018.3495367

[7] Yang, M., Marchionini, G. (2004). Exploring users' video relevance criteria—A pilot study. Proceedings of the Association for Information Science and Technology, 41(1), 229–238. https://doi.org/10.1002/meet.1450410127

[8] Ondego, V. K., Komlódi, A. (2017). Web search selection criteria of foreign-language searchers. Proceedings of the Association for Information Science and Technology, 54(1), 511–514. https://doi.org/10.1002/pra2.2017.14505401059

[9] Albassam, S.A., Ruthven, I. (2019). Dynamic aspects of relevance: differences in users' relevance criteria between selecting and viewing videos during leisure searches. Inf. Res., 25.

[10] Vaseeharan, T., Aponso, A. (2020). Review On Sentiment Analysis of Twitter Posts About News Headlines Using Machine Learning Approaches and Naïve Bayes Classifier. ACM. https://doi.org/10.1145/3384613.3384650