TABLE OF CONTENTS

ARTIFICIAL INTELLIGENCE (AI)

01.

INTRODUCTION

A NEW BIG THING IN EDUCATION: AI

- Enhanced efficiency and language enhancement
- Concerns about the decline of original thought in students
- Research in UC San Diego
  - Tested teachers and students with pairs of essays
  - Teachers: 70%
    Students: 62%

Kiderra, Inga. "Study Shows That Even Confident Educators Have Trouble Recognizing AI-Generated Essays." *Phys.Org*, Phys.org, 28 June 2023,

MAIN OBJECTIVE

ARTIF INTEL (AI)

- **Build a model to classify AI vs. Human generated texts**
- **Assess model performance in a new dataset - *generalisability***

<<<<

OUR DATA

Data 1 - Model training & Internal validation

- From Kaggle
- Around 500k essays, articles, and letters created by AI or written by human
- 487,235 rows, 2 columns
  - *text*: essays, articles, letters ..
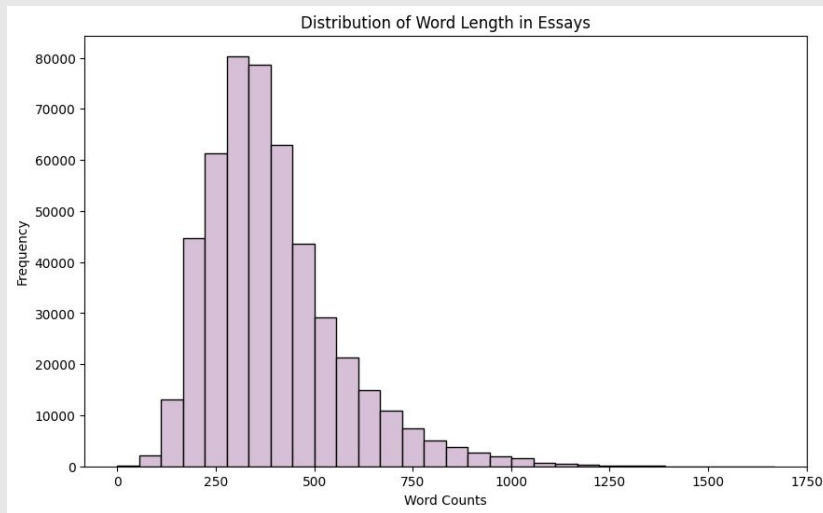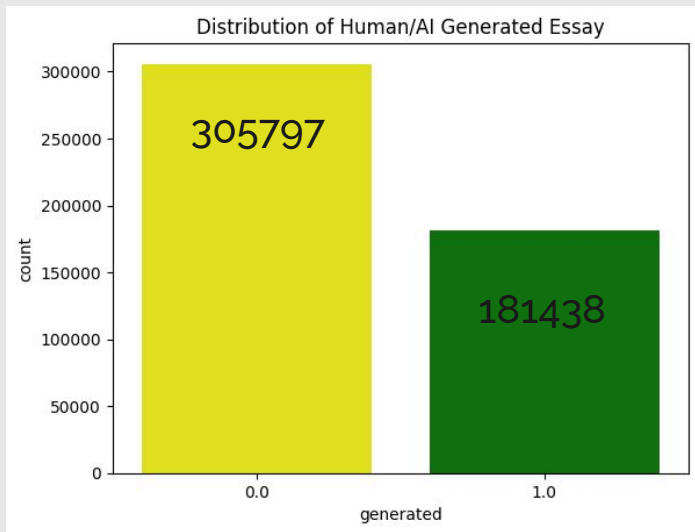 - *generated*: labeled AI(1) or Human(0)

Data 2 - External validation

- Combined data
  - 100 BBC articles (2004-2005)
  - 100 AI generated essays in various topics (Gemini)
- 200 rows, 2 columns

Distribution of Human/AI Generated Essay

Distribution of Word Length in Essays

0.0 = *Human generated* , *1.0* = *AI Generated*

ESSAY SAMPLE

Quick Quiz
AI or Human ?

1

Smaller states tend to hold significant power because their two votes for president and vice president add up more than the votes of larger states that have many electors. This is because of the split of the electoral votes. Some argue that electors are not bound to vote for the candidate who won the most votes nationally. They do not have to vote for their own state's nominee unless their state has a winner take all system. However, there are states that have adopted laws that force their electors to vote for their state's candidate. It seems that, no matter how, electors are not bound to vote for the candidate who won the most nationally. This is not always the case because of state legislatures who can overrule the electors and vote for the alternative candidate their citizens have selected for them, even if the voter lives in a state without a winner take all system.

2

The electrol college system is an unfair system, people don't have the right to select their own president, they dont have the right to select a president. Because, when people vote they are technically voting for the electors for a candied. That candidate can be a democratic or a republican. In source two it states that electors can choose there opposing candidate. Which declines the the whole voting process. Why do we vote? we vote to select a leader who will defend this country and make america a place of opportunity. The most important reason why the electrol cllege is unfair because of the " winners it take all" rule.

02.
STATISTICAL METHODOLOGY

VECTORIZATION TECHNIQUES

TF-IDF

- Transforms text data into numerical representations
- Calculates the importance of a term within a document relative to entire corpus.
- Combines TF and IDF, effectively highlighting terms that are frequent in a document but rare in the corpus.

DTM

- Represents documents as matrices where rows correspond to documents and columns to terms
- Captures the occurrence of terms in documents and enables the analysis of textual data at a granular level.
- Valuable for capturing the context of terms within documents,.

> > > >

N-GRAM ANALYSIS

- MONOGRAM - Single words in a text, representing the most basic unit of analysis. They provide insights into individual word frequencies and distributions.

- BIGRAM - Sequential pairs of words in a text, capturing relationships between adjacent words. They offer a more contextual understanding of language by considering word co-occurrences.

NAIVE BAYES

- Naive Bayes: Probabilistic classification algorithm based on Bayes' theorem.

- Assumes independence between features given the class label.

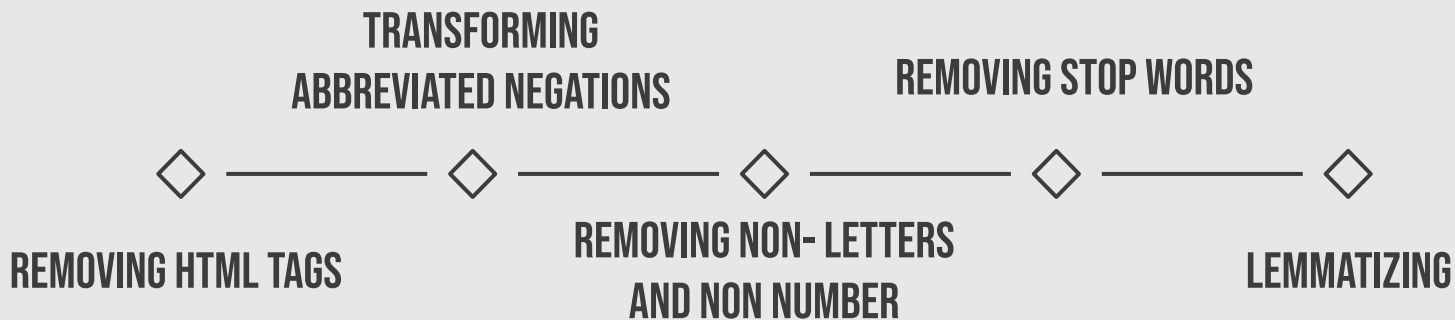- Particularly effective for text classification tasks due to its simplicity and efficiency.

> > > >

03.

NLP & MODELING

TEXT PREPROCESSING

TRANSFORMING
ABBREVIATED NEGATIONS

REMOVING STOP WORDS

REMOVING HTML TAGS

REMOVING NON- LETTERS
AND NON NUMBER

LEMMATIZING

> > > >

ORIGINAL ESSAY EXAMPLE

Cars. Cars have been around since they became famous in the 1900s, when Henry Ford created and built the first ModelT. Cars have played a major role in our every day lives since then. But now, people are starting to question if limiting car usage would be a good thing. To me, limiting the use of cars might be a good thing to do.

In like matter of this, article, "In German Suburb, Life Goes On Without Cars," by Elizabeth Rosenthal states, how automobiles are the linchpin of suburbs, where middle class families from either Shanghai or Chicago tend to make their homes. Experts say how this is a huge impediment to current efforts to reduce greenhouse gas emissions from tailpipe. Passenger cars are responsible for 12 percent of greenhouse gas emissions in Europe...and up to 50 percent in some carintensive areas in the United States. Cars are the main reason for the greenhouse gas emissions because of a lot of people driving them around all the time getting where they need to go. Article, "Paris bans driving due to smog," by Robert Duffer says, how Paris, after days of nearrecord pollution, enforced a partial driving ban to clear the air of the global city. It also says, how on Monday, motorist with evennumbered license plates were ordered to leave their cars at home or be fined a 22euro fine 31. The same order would be applied to oddnumbered plates the following day. Cars are the reason for polluting entire cities like Paris. This shows how bad cars can be because, of all the pollution that they can cause to an entire city.

Likewise, in the article, "Carfree day is spinning into a big hit in Bogota," by Andrew Selsky says, how programs that's set to spread to other countries, millions of Columbians hiked, biked, skated, or took the bus to work during a carfree day, leaving streets of this capital city eerily devoid of traffic jams. It was the third straight year cars have been banned with only buses and taxis permitted for the Day Without Cars in the capital city of 7 million. People like the idea of having carfree days because, it allows them to lesson the pollution that cars put out of their exhaust from people driving all the time. The article also tells how parks and sports centers have bustled throughout the city uneven, pitted sidewalks have been replaced by broad, smooth sidewalks rushhour restrictions have dramatically cut traffic and new restaurants and upscale shopping districts have cropped up. Having no cars has been good for the country of Columbia because, it has aloud them to repair things that have needed repairs for a long time, traffic jams have gone down, and restaurants and shopping districts have popped up, all due to the fact of having less cars around.

In conclusion, the use of less cars and having carfree days, have had a big impact on the environment of cities because, it is cutting down the air pollution that the cars have majorly polluted, it has aloud countries like Columbia to repair sidewalks, and cut down traffic jams. Limiting the use of cars would be a good thing for America. So we should limit the use of cars by maybe riding a bike, or maybe walking somewhere that isn't that far from you and doesn't need the use of a car to get you there. To me, limiting the use of cars might be a good thing to do.

CLEANED ESSAY EXAMPLE

car car around since become famous 1900s henry ford create build first modelt car play major role every day life since people start question limit car usage would good thing limit use car might good thing like matter article german suburb life go without car elizabeth rosenthal state automobile linchpin suburb middle class family either shanghai chicago tend make home expert say huge impediment current effort reduce greenhouse gas emission tailpipe passenger car responsible 12 percent greenhouse gas emission europe 50 percent carintensive area unite state car main reason greenhouse gas emission lot people drive around time get need go article paris ban drive due smog robert duffer say paris day nearrecord pollution enforce partial drive ban clear air global city also say monday motorist evennumbered license plate order leave car home fin 22euro fine 31 order would apply oddnumbered plate follow day car reason pollute entire city like paris show bad car pollution cause entire city likewise article carfree day spin big hit bogota andrew selsky say program set spread country million columbians hike bike skate take bus work carfree day leave street capital city eerily devoid traffic jam third straight year car ban bus taxi permit day without car capital city 7 million people like idea carfree day allow lesson pollution car put exhaust people drive time article also tell park sport center bustle throughout city uneven pit sidewalk replace broad smooth sidewalk rushhour restriction dramatically cut traffic new restaurant upscale shop district crop car good country columbia aloud repair thing need repair long time traffic jam go restaurant shop district pop due fact le car around conclusion use le car carfree day big impact environment city cut air pollution car majorly pollute aloud country like columbia repair sidewalk cut traffic jam limit use car would good thing america limit use car maybe rid bike maybe walk somewhere not far not need use car get limit use car might good thing

FEATURE EXTRACTION

TF-IDF

Dimensions

Unigram: (341064,20000)

Bigram:

DTM

Dimensions

Unigram: (341064,20000)

Bigram:

ARTIFICIAL INTELLIGENCE (AI)

MODEL RESULTS - ACCURACY

LOGISTIC REGRESSION

TF-IDF
- Bigram : 0.9944
- Unigram : 0.9897

DTM
- Bigram: 0.9989
- Unigram: 0.9970

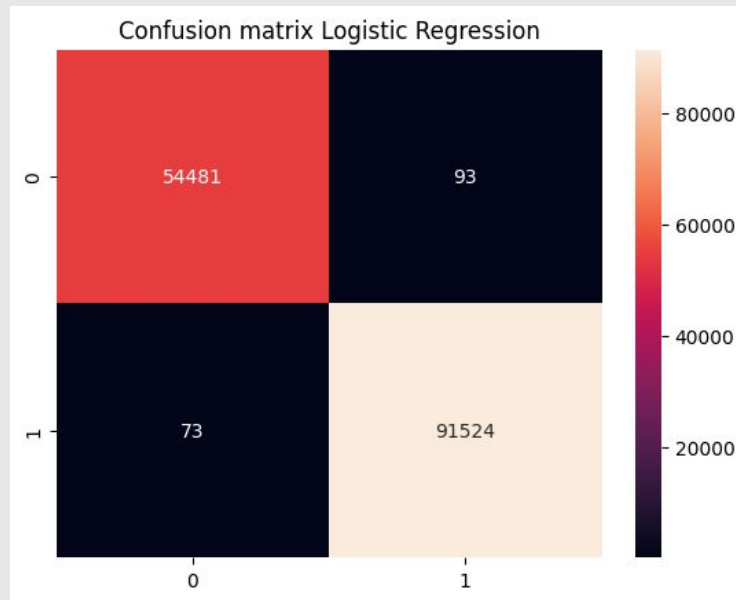## NAIVE BAYES

TF-IDF
- Bigram: 0.9670
- Unigram: 0.9506

DTM
- Bigram: 0.9646
- Unigram: 0.9500

Logistic Regression with **DTM Bigram**

Accuracy: **0.9989**

Confusion matrix Logistic Regression

04.

EXTERNAL VALIDATION

NEW DATA

- Created the new dataset by combining
  - Data from Kaggle: 2004-2005 BBC articles
  - 100 AI generated essays in various topics using Gemini
- 200 rows, 2 columns

> > > >

GEMINI GENERATED ESSAY
EXAMPLE

The Paradox of Progress: Technology's Promise and Peril

In the relentless march of human progress, technology has emerged as both a beacon of hope and a source of unease. While it has undeniably transformed our lives in countless ways, bringing unprecedented convenience, connectivity, and knowledge at our fingertips, it has also raised profound questions about our future and the nature of our humanity.

The promise of technology lies in its potential to solve some of the world's most pressing challenges. From eradicating diseases and addressing climate change to connecting people across vast distances and empowering individuals with access to information and education, technology has the power to create a more just, equitable, and sustainable world.

However, the peril of technology lies in its potential for misuse and unintended consequences. The same tools that can connect us can also isolate us, the same algorithms that can personalize our experiences can also manipulate our behavior, and the same advancements that can save lives can also be used to wage war and inflict harm.

One of the most pressing paradoxes of progress is the tension between efficiency and humanity. Technology has undoubtedly made many tasks more efficient, freeing up our time and resources for other pursuits. However, it has also led to a culture of constant stimulation and distraction, where we are perpetually connected to our devices and bombarded with information. This can take a toll on our mental health, our relationships, and our ability to focus and engage with the world around us.
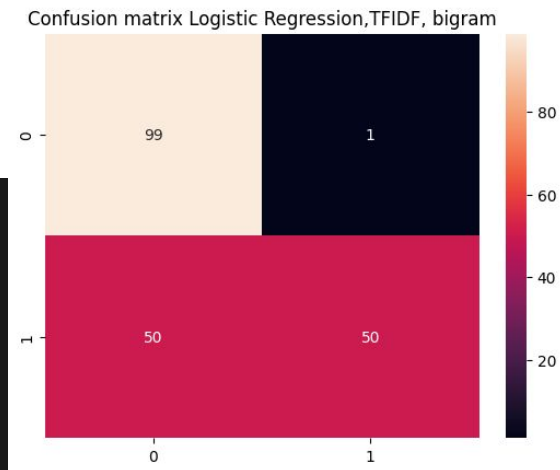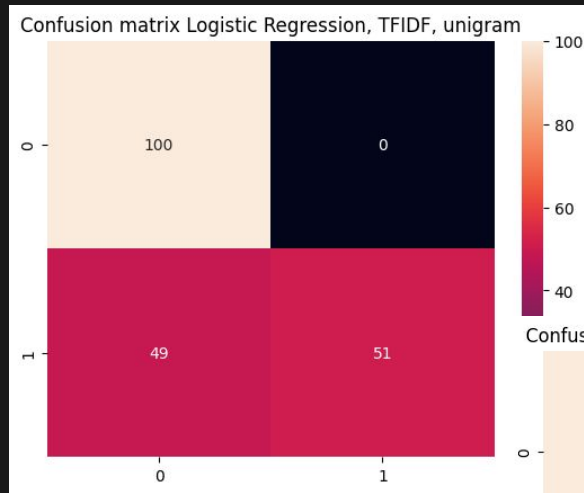
> > > >

MODEL RESULTS - ACCURACY

NAIVE BAYES

TF-IDF
- Bigram : 0.88
- Unigram : 0.85

DTM
- Bigram: 0.89
- Unigram: 0.84

Confusion matrix Naive Bayes, TFIDF, bigram

Confusion matrix Naive Bayes, DTM bigram
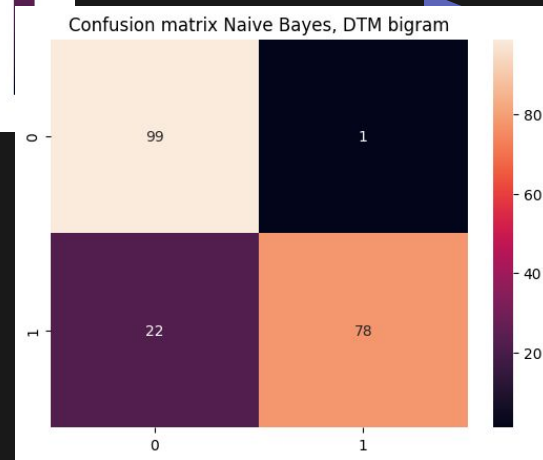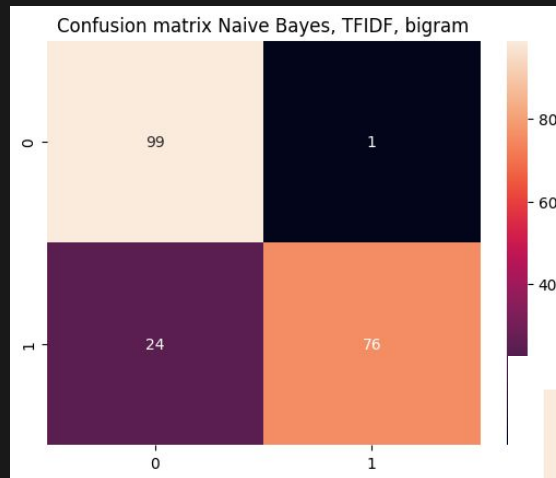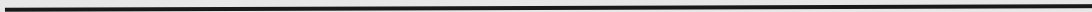
05.

CONCLUSION

SUMMARY OF OUR STUDY

- Successfully built models to classify AI v.s Human generated texts
  - Highest Accuracy: Logistic Regression with DTM Bigram (0.9989)

- Tested our models on new dataset and confirmed the performance
  - Naive Bayes with TF-IDF Bigram (0.88)
  - Naive Bayes with DTM Bigram (0.89)

LIMITATIONS

DATA BIAS

The dataset might be biased towards certain characteristics of human/AI-generated text

IMBALANCE CLASSES
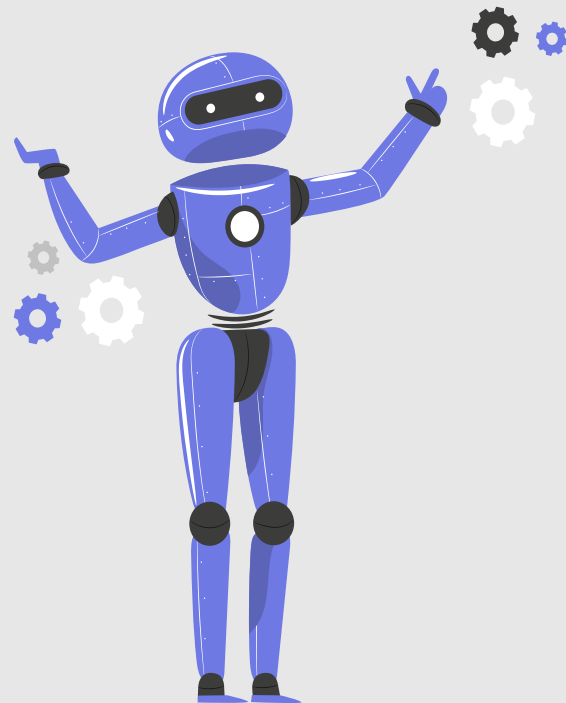
Model might be biased towards majority of classes

RECOMMENDATIONS

DIVERSE DATA

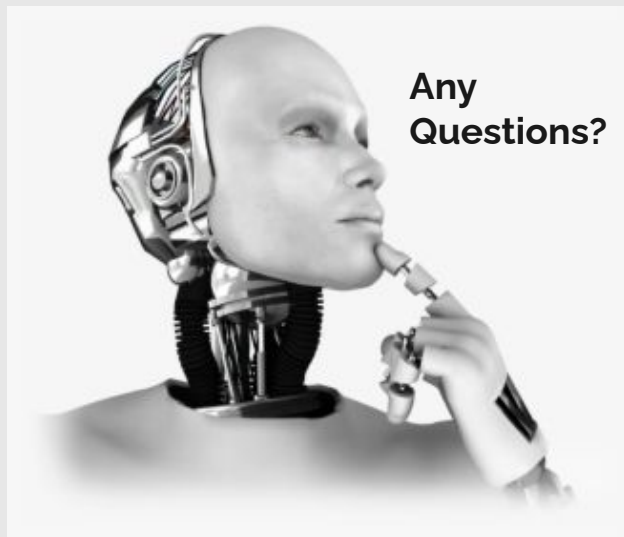Gather more extensive data from other AI models

OTHER FEATURE REPRESENTATION AND MODELS

Word2Vec,GloVe, Transformers