IST 722: Text Mining

**Analyzing University Guidelines on generative AI**

Final Project Report

A logo with leaves and text

Description automatically generated

**Group Members:**

VIPUL SARODE

RITESH VERMA

**Content**:

1. Introduction
2. Method
3. Result
4. Conclusion
5. Ethics Statement
6. Reference

**Introduction**

**Purpose:** This report is dedicated to the exploration and analysis of university guidelines regarding the use of generative AI tools. Given the rapid evolution and increasing prevalence of AI in educational contexts, our primary aim is to understand how universities are navigating this landscape. This involves identifying their major concerns, the extent to which they encourage or discourage AI use, their focus on data privacy and security, and the implementation of ethical standards in AI usage.

**Background:** The integration of AI in academic settings has become a pivotal topic in recent years. With generative AI tools becoming more sophisticated, universities are faced with the challenge of balancing technological advancement with ethical considerations, data security, and academic integrity. These tools offer immense potential for enhancing learning and research but also pose significant risks if not governed by comprehensive guidelines. This dichotomy necessitates a thorough examination of how different universities are approaching these challenges.

**Overview:** In this report, we dissect the various policies and guidelines adopted by universities in relation to generative AI. The report is organized into several key sections: After this introduction, we delve into our methodology, outlining how we gathered and analyzed data across various academic institutions. This is followed by a detailed presentation of our results, where we highlight the key findings from our analysis. The conclusion section synthesizes these findings, drawing out the larger implications for the academic community. Lastly, we include an ethics statement, reflecting on the moral considerations inherent in our research and the subject matter at hand.

**Method**

**Approach:** This study was conducted with a systematic approach to gather and analyze the guidelines on the use of generative AI by the top 70+ universities in the USA. The primary source for the selection of these universities was the Yocket website, which ranks institutions based on various criteria of academic excellence. This initial list served as the basis for our comprehensive dataset.

A black and white text on a grey background

Description automatically generated

***Dataset***

**Dataset Construction:** The dataset was meticulously compiled through a manual selection process to ensure the inclusion of relevant and current guidelines pertaining to generative AI usage. Each university's stance on AI tools, their concerns regarding data privacy, security, and ethics were carefully documented. The dataset was designed to capture the nuances of each university's policy, providing a robust foundation for our analysis.

**Data Collection:** To collect the guidelines, we employed the BeautifulSoup library, a Python package widely used for web scraping. This tool enabled us to programmatically navigate through the universities' official websites, extract the pertinent information, and ensure the data's accuracy and relevance. Special attention was given to the sections of the websites where AI policies were discussed, including academic honesty policies, IT department advisories, and official statements from university leadership.

**Data Processing:** To preprocess the data and make it ready for our analysis, we first used regex to remove any special characters and digits as they are irrelevant for our analysis. Then, we converted all the words to lowercase and tokenized them. Finally, we removed all the stop words form the text and used the WordNetLemmatizer to convert all the words into their basic lemmas.

**Quality Assurance:** To ensure the reliability of the data, we conducted a thorough verification process. This included cross-referencing the scraped data with the official policy documents available in the public domain, when accessible. Additionally, we reached out to university representatives for confirmation or clarification when guidelines were ambiguous or not publicly detailed.

**PART 1:**

**Sentiment Analysis**: Sentiment analysis in our project was conducted using an advanced transformer model. These models, known for their efficiency in handling natural language processing tasks, analyze text data to discern underlying sentiments. The transformer processes input text, such as customer reviews or social media posts, and identifies patterns indicative of positive, negative, or neutral sentiments. This is achieved through the model's layered architecture, which allows for deep contextual understanding of the language. The output, sentiment scores, aids in understanding public perception and sentiment trends, providing valuable insights for decision-making. We used transformers for sentiment analysis as we didn’t have labeled data and task at hand was an unsupervised learning task.

**K-means**: In our analysis, we employed the K-means clustering algorithm to segment our dataset into distinct groups. This method involves partitioning data into 'k' clusters based on feature similarities. Initially, we determined the optimal number of clusters using the Elbow Method. By plotting the sum of squared distances against the number of clusters, we identified a distinct 'elbow' at k=5, indicating that five clusters would be the most effective for our analysis. Subsequently, we proceeded with five clusters in our K-means algorithm. After clustering, we focused on the cluster that prominently featured the word 'guidelines'. Using a word cloud visualization, we highlighted the clustered the universities using the most used keywords in their guidelines.

Cluster 0: A close-up of words

Description automatically generated Cluster 1: A close-up of words

Description automatically generated

Cluster 2: A close-up of words

Description automatically generatedCluster 3: A close-up of words

Description automatically generated

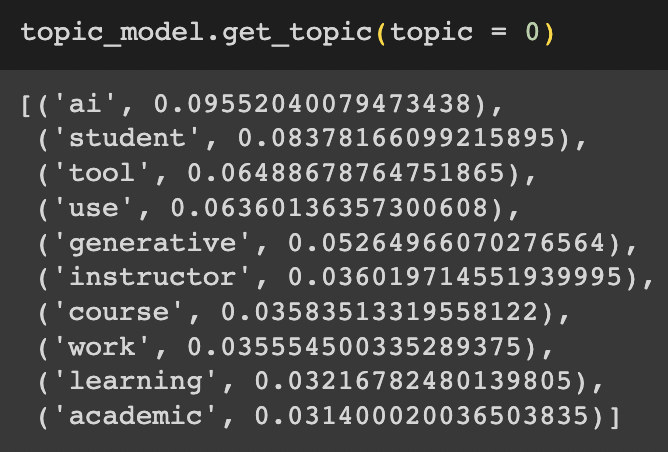
Cluster 4: A close-up of words

Description automatically generated

**Topic Modelling**

**A screenshot of a cell phone

Description automatically generated**

 A screen shot of a computer

Description automatically generated A screen shot of a computer

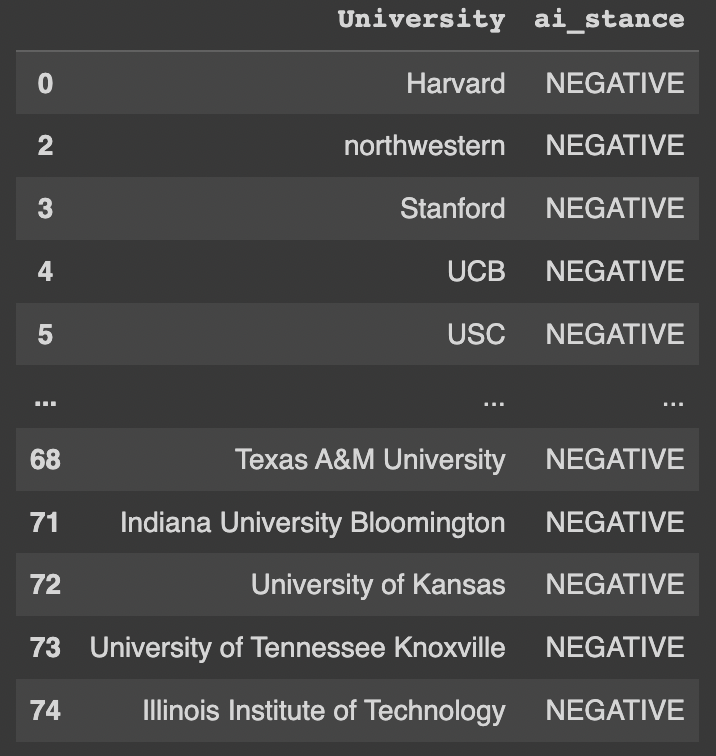
Description automatically generated

Topic modeling is a statistical method used for discovering the abstract "topics" that occur in a collection of documents. It is often used in text mining to uncover hidden semantic structures in large text bodies. The process involves identifying clusters of words that frequently occur together in documents, with each cluster representing a topic. The output image appears to show the results of a topic modeling analysis, where '0' and '1' represent two distinct topics associated with universities. These topics could be composed of words frequently mentioned in texts related to different areas of study or administrative aspects of these universities. The '-1' denotes outliers, which are data points or terms that do not fit well with the identified topics. This suggests that topic modeling not only helps in categorizing data but also in detecting data points that diverge from the main themes.

**Part 2:**

**Results**: The sentiment analysis conducted for this study utilized advanced transformer-based models to assess the stance of various universities towards the use of artificial intelligence (AI). The initial dataset comprised public statements, official publications, and academic papers from each institution. Through the sentiment analysis, we categorized universities into two groups: those that demonstrate a positive stance toward AI utilization, and those with a negative stance. The analysis revealed that universities such as Yale and Cornell University are advocates for AI, promoting its use and development. In contrast, institutions like Harvard and Stanford exhibit a cautious or negative disposition towards AI proliferation. The differentiation of these stances underscores the diverse institutional perspectives on the implications and ethics of AI integration into academia and society at large.

A screenshot of a cellphone

Description automatically generated ****

***Universities who encourages******Universities who discourages***

Our examination of university priorities regarding AI ethics and security was conducted through a word matching methodology. This technique involved scanning institutional documents for specific keywords associated with 'ethics' and 'security' within the context of AI. The process revealed that universities such as Duke University, ETH Zurich, and the California Institute of Technology explicitly emphasize ethical considerations in their AI programs. Similarly, Harvard, Yale, and Northwestern were identified as institutions that prioritize security in their AI initiatives. The use of word matching enabled us to efficiently highlight the universities that not only engage with AI but also take a conscientious approach towards the ethical and security dimensions of its implementation.

A screen shot of a list of college ethics

Description automatically generated A screenshot of a cell phone

Description automatically generated

***Universities who focus on ethics Universities who focus on security***

**PART 3:**

**Conclusion**

In conclusion, our comprehensive analysis utilizing sentiment analysis and word matching techniques has provided a clear panorama of the academic landscape regarding AI. It is evident that while a significant number of universities maintain a positive stance towards the adoption and advancement of AI, there is a contrasting group that holds reservations, underlining the need for a cautious approach. Furthermore, our findings underscore a strong institutional focus on ethics and security, suggesting that these universities are not only embracing AI but are also deeply committed to its responsible use. This balance of enthusiasm and prudence reflects a mature and multifaceted academic response to the challenges and opportunities presented by AI technologies. According to our K-means analysis there is a significant number of universities who are reluctant use open source AI Because of data privacy and security issues. Such universities may possibly use and promote their own self-made AI

**Part 4**

**Ethics Statement**

Our analysis heavily relies on the current guidelines published by universities. While we possess extensive knowledge regarding current trends, attitudes, and hesitations of universities towards AI usage, we are unable to deploy our models for production at present. The reason being, in the dynamic landscape of AI, there's a possibility that many universities might alter their stance on AI usage as new technologies emerge. For instance, if there's a new extension or plugin for an open-source AI that restricts the AI model from retraining or collecting data from prompts, universities that currently abstain from AI due to privacy concerns might reconsider their stance. Therefore, deploying the model and assuming its results to be indicative of the future might not be ethically sound.

**PART 5**

**Reference**

BERTopic :- <https://www.pinecone.io/learn/bertopic/>

Sentiment Analysis :- <https://www.geeksforgeeks.org/what-is-sentiment-analysis/>

Topic Modelling :- <https://monkeylearn.com/blog/introduction-to-topic-modeling/>