Chat-GPT Based Entities Classification - ESWA 2023

Taliya Shitreet, Renana Rimon, Eran Levy, Daniel Zafrir ,Or Haim Anidjar, Michelle L. Oren

[1]School of Computer Science, Ariel University, Israel
,[2]Faculty of Architecture, Technion - Israel Institute of Technology

orhaim@ariel.ac.il, Taliyashitreet@gmail.com, renana1414@gmail.com, eranlevy9@gmail.com, danieltzafrir96@gmail.com, michelle.oren@gmail.com

Abstract

renana Urban planning in Bogota's Savannah, Central Colombia, poses challenges for land provision, urban projects, public services, social integration, environmental protection, and transport network consolidation. Collaborating with a local non-profit agency, they leverage technology to develop the region's 2051 vision. Their strategic plan guides decision-makers in sustainable development, involving stakeholders and residents. Traditional surveys are impractical at such a large scale, necessitating a robust technological solution to understand collected data, including the nuances of different sentences conveying the same meaning, and are both scalable and replicable to be employed in additional regions worldwide. One of the limitations in the world of data and AI, especially in NLP, is the scarcity of labeled data. In our research, we propose an innovative solution for automatic data tagging based on ChatGPT. Furthermore, we present two approaches: unsupervised and supervised. In the unsupervised approach, we utilize the language-agnostic BERT model, which enables model generalization across additional languages, going beyond individual words to comprehend sentence-level meanings. The model represented in the research learns the distribution from the ChatGPT labels and improves the performance by labeling new sentences even more effectively.

1. Introduction

Eran Urban planning has always been complex. What has changed is the scale, as humanity experiences unprecedented growth. Never in history have human settlements reached such a large population size that we can no longer use the same tools we used 100 years ago. The source of the problem is the limited capacity of the agents leading public participation processes in reaching these populations to collect their feedback. this is a global problem, not a local one. Our intervention is local but this is a global problem and we are providing a global solution. Obtaining the maximum participation of citizens in the drafting process is of major importance. Participation in the preparation, diagnosis, formulation, monitoring, and evaluation of territorial management plan gives citizens in general, directly or through various forms of social and economic organization such as unions, boards, local administrators, or ecological, civic, and community entities the possibility of incorporating their proposals to improve their quality of life, through the promotion of public space and urban construction projects that optimize land use, mitigate the conflicts derived from it and allow the improvement of essential public facilities such as hospitals, public parks, libraries, and government service centers.

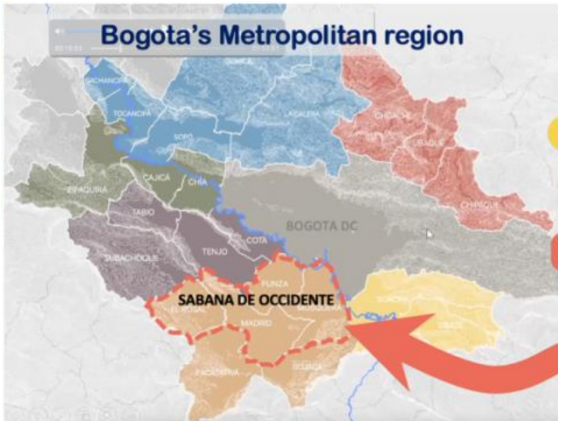

Figure 1: *Bogota's 'Sabana de Occidente planning area comprises the municipalities of Funza, Mosquera, Madrid, Facatativá and El Rosal (only part of the district).*

Taliya While the area is known for its vibrant culture and burgeoning business scene, it is also grappling with high levels of poverty, unemployment, and income inequality. A significant portion of the population lives in a formal settlement, houses that were built legally but they are not planned, meaning each person built his own house with proper authorizations, but no one was overseeing the development at the macro

level where access to basic services such as clean water, sanitation, and electricity can be limited. These disparities are further exacerbated by the limited public transportation options available to residents, which can hinder their ability to access education, healthcare, and employment opportunities.

Eran These municipalities attracted mostly populations who could not afford to live in Bogota and would commute to the capital daily. Still, the region has also developed its own industries, among them 'export-quality' flowers - the biggest crop is roses. This also means that a large percentage of the population living in the area are seasonal workers, one of the strongest explanations as to why the region did not develop a strong identity and has been so shortsighted when it comes to designing the future living conditions in the area.

Taliy The planning agent, is a non-profit organization composed of 40 different companies that donate funds to the organization every month. The organization is staffed by young, visionary planners with a high degree of motivation and a strong desire to make a difference and involve the community in the process of ultimately designing the city while representing the interests of the population and contributing to the community.

Eran The challenge of better city planning, particularly in areas heavily reliant on seasonal workers such as the region under consideration, highlights the need for public participation to gather diverse perspectives and harness the crowd's collective wisdom to develop more effective strategic plans.

However, large cities can present challenges when it comes to the logistics of organizing, collecting, and interpreting the required information from everybody.

Michelle However, despite the extensive research in the PD field and the availability of several commercial technologies, organizing a significant participatory process remains a challenging task because of the low participation rate [1] [2]. Other pertinent challenges include intra-community politics [3], bureaucracy [4], professional language and terms [5], cultural and religious aspects [], knowledge gaps [6], and the loss of public trust in politicians and local authorities [7]. Historically, PD has involved user engagement methods focused on collecting information from citizens to form a better understanding of their needs, such as storytelling, workshops, and interactive design tools [8] [9]. However, these engagement methods do not differ from traditional user-centered design (i.e., asking people what they need). Therefore, it has been argued that, in the effort toward the improvement of products and services, PD should be more motivated by the belief in the value of democracy and empowerment of disenfranchised groups [10] [11]

Eran In [12] there are suggestions about best practices for collecting data through crowdsourcing one of them is using a well-designed crowdsourcing platform. A well-designed crowdsourcing platform can help to ensure that the data is collected in a systematic and accurate way. Second, validating the data, the data should be validated to ensure that it is accurate and reliable. Protecting the privacy of individuals, the privacy of individuals should be protected by collecting only the data that is necessary. And of course, securing the data, the data should be secured to protect it from cyberattacks.

Potential solutions to the challenge of engaging citizens in city planning could be to utilize platforms that are already widely used and accessible to the public. For example, in Colombia, WhatsApp is prevalent due to its accessibility through free social network packages offered by smartphone providers. Leveraging such platforms for communication and engagement could increase the likelihood of citizen participation and help ensure a more representative range of perspectives is incorporated into the city planning process.

Michelle The ChatBot that will be collecting the data will be operated on the Whatsapp platform under user [+57 3227778403] to allow greater access as it is a free application that does not depend on the availability of domestic internet connectivity or data and sold separately in unlimited use (timed) packages to mobile line holders in Colombia. Global Web Index's 2021 Social Media User Trends Report rated Colombia as the sixth highest country in the world in Monthly WhatsApp users as a percentage of total internet users aged 16-64, scoring 92%. WhatsApp also makes one of the top ten identified applications as 'highly accessible' for persons with disabilities based on the Web Content Accessibility Guidelines (WCAG). For those with impairments or literacy barriers, participants have the option of sending recorded voice messages instead of typing their answers.

By merging new technologies with existing ones such as WhatsApp chatbot and ChatGPT, the planning agent can gather diverse perspectives from a broader range of residents, addressing the challenges posed by Colombia's social disparities and limited mobility. The incorporation of digital tools helps create a more inclusive and representative strategic planning process, empowering communities to have a direct impact on the future development of their city. The insights gathered from this process will be invaluable in shaping a strategic plan that truly reflects the needs and aspirations of the people of Bogotá.

Eran But the question arises, how can the capabilities of the ChatBot be used to generate concrete insights from the people that the planners can use?

On [13] the paper discusses the concept of insight building. Insight building is the process of us-

ing data to generate insights. Insight building can be a complex process, but it can be made easier by using crowdsourcing. The paper also discusses the following methods for getting insights from the data that is collected through crowdsourcing: Data mining is a process of extracting knowledge from data. Data mining can be used to identify patterns in data, make predictions about future events, and generate insights. Machine learning: Machine learning is a field of computer science that focuses on the development of algorithms that can learn from data. Machine learning can be used to develop models that can predict future events, generate insights, and make decisions. Natural language processing: Natural language processing is a field of computer science that focuses on the interaction between computers and human language. Natural language processing can be used to extract meaning from text data, to generate insights, and to make decisions.

Before beginning the planning process, it is essential for the planners to have a thorough understanding of the area they are going to work on. This requires conducting a statutory study, which involves obtaining information such as building permits, land use patterns, and legal ownership in the region. In addition, geophysical research is crucial, including geography, geology, hydrology, and aspects related to environmental hazards. The quality of built and open areas, road infrastructures, and service infrastructures are also studied alongside other relevant investigations.

To proceed further, the data was classified into six thematic categories identified by the planning agent prior to the participation process, as key areas requiring intervention. Without knowing if those were the actual dominant thematic categories de facto discussed by the participants.

Eventually, the ChatBot must gather the data and classify it into six key themes. These themes are as follows:

- Environment and climate resilience
- Mobility (transport)
- Local identity
- Future of work
- Land use
- Other

Now a short explanation about each category will follow, and immediately after that will come an example for each category:

- **Environment and climate resilience:** This category focuses on protecting and preserving the environment, as well as building resilience to the impacts of climate change. It involves initiatives to reduce greenhouse gas emissions, promote renewable energy, conserve natural resources, and adapt to changing climatic conditions. Example sentence: "The city council implemented a comprehensive recycling program and installed solar panels on public buildings to enhance environment and climate resilience."
- **Mobility (transport):** This category deals with improving transportation systems and enhancing mobility options for individuals and goods. It includes efforts to reduce traffic congestion, promote sustainable modes of transportation such as walking, cycling, and public transit, and embrace emerging technologies like electric vehicles and autonomous transportation. Example sentence: "The city introduced bike-sharing programs and expanded the public transit network to encourage sustainable mobility among residents."
- **Local identity:** This category focuses on preserving and celebrating the unique cultural, historical, and social characteristics of a specific locality or community. It involves initiatives to protect heritage sites, supports local arts and crafts, promote traditional cultural events, and foster a sense of belonging among residents. Example sentence: "The city organized an annual festival that showcased local traditions, music, and cuisine to celebrate and strengthen the local identity."
- **Future of work:** This category explores the evolving nature of work in the face of technological advancements and changing economic landscapes. It encompasses initiatives to foster innovation, promote entrepreneurship, upskill and reskill the workforce, and create inclusive and flexible work environments. Example sentence: "The city established a co-working space and launched entrepreneurship programs to support the future of work and encourage startups."
- **Land use:** This category involves the planning, management, and allocation of land resources in urban and rural areas. It includes initiatives to ensure sustainable land development, protect natural habitats, promote mixed land-use patterns, and create efficient and inclusive spaces for living, working, and recreation. Example sentence: "The city implemented a zoning policy that encourages mixed-use developments, allowing residents to live, work, and access amenities within walking distance."

Taliya The chatbot gathers all this information, but the question arises, how will this clustering be performed? How can insights be extracted from this data?

We have data from numerous conversations with various clients, and our research deals with classifying these conversations into the five main topics while considering the extraction of entities from the sentences. We argue that there is an influence of an entity in a sentence on its classification into a specific topic.

Since we are dealing with unlabeled data, we are talking about unsupervised learning [14] [15] using an entity extraction model. In this context, we used the language-agnostic BERT Sentence Embedding (LaBSE) [16] [17] technique to generate the embedding vectors for the sentences and categories. LaBSE is an advanced natural language processing (NLP) method that allows for the creation of vector representations of sentences in a way that is agnostic to the input language, enabling semantically meaningful comparisons and analyses of sentences across different languages. By using these word embeddings, each sentence and category will be represented by a vector, and we will employ mathematical metrics such as cosine similarity and Euclidean distance to estimate the distance of each sentence from each of the categories, taking into account the nouns in the sentence.

Additionally, we can use supervised learning [18] technology by labeling the data using ChatGPT. By giving the chat a sentence, he will have to tag it into one of the 6 categories that we mentioned above. Upon obtaining the labeled data from ChatGPT, we pursued two distinct methods for generating sentence embeddings, aimed at making the dataset suitable for Multi-Layer Perceptron (MLP) model training [19]. These two methods were the Sentence BERT [20] and Term Frequency-Inverse Document Frequency (TF-IDF) [21] [22] via the TextVectorization layer in TensorFlow, marking two distinct approaches for text classification in our work.

TF-IDF is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. It measures the significance of each word by considering its frequency in the specific sentence and its scarcity in the entire corpus. This emphasizes words that are frequent in specific sentences but are generally rare in the corpus, thus highlighting the more informative words in the text.

In the first approach, we transformed the labeled sentences into vector representations using the bert-base-nli-mean model. This yielded embeddings capable of capturing the context-aware and language-agnostic characteristics of the sentences.

In the second approach, we utilized TF-IDF for generating embeddings, focusing on the word frequency within the sentences, thus underlining the most informative and distinctive words in the text.

In both approaches, the vectors derived from these embeddings laid the foundation for training the supervised MLP models. This process enabled the models to predict categories based on the syntactic, semantic, and word importance characteristics of the sentences, as captured by the bert-base-nli-mean and TF-IDF embeddings, respectively.

In summary, the Bogotá metropolitan region and its municipal cities are dynamic urban areas facing significant socioeconomic challenges that need to be addressed through inclusive and representative urban planning processes.

1.1. Our contribution

Talya This paper presents an innovative approach to tagging data using ChatGPT. In addition, a chatbot is used to collect and organize resident feedback to inform the urban development team. The contributions in this paper are organized as follows:

- **Improving data collection and processing with deep learning technique.** The implementation of the chatBot eliminates the need for manual data collection from seven million residents. Instead, it leverages deep learning techniques [23] [24] to aggregate, analyze, and summarize this information efficiently, without the need for a human team to carefully analyze and summarize the responses.
- **Avoiding human intervention by using ChatGPT.** Using ChatGPT, the researchers skip the labor-intensive process of human data tagging. This advancement overcomes the logistical challenges and potential errors that humans would have made in tagging data.
- **Ensuring consistent data tagging using ChatGPT.** This approach eliminates the potential differences and inconsistencies that might have arisen from employing different taggers for the data, thereby improving the quality and reliability of the project's data.
- **Adopting a dual approach consisting of supervised and unsupervised methods for scalability.** The proposed solution combines supervised and unsupervised approaches to facilitate seamless expansion to additional categories. This two-way approach allows for immediate expansion to new categories in the unsupervised approach, while the supervised technique allows for easy adjustments to expand categories.
- **Offering a replicable model for public participation processes in urban and regional planning with global potential.** This research contributes to the generation of more inclusive and sustainable urban and regional development that can be scaled according to needs

adapting to a variety of population sizes and geographic extent. It figures out the best way to develop cities in the long run, and it uses Bogotá's occidental Savannah as an example.

1.2. Paper structure

The remainder of this paper is structured as follows: Section 2 focus on ChatBot in combination with NER [25] and, embedding sentences with BERT, and also the employment of chatbots in urban planning processes such as public participation and crowdsourcing. Section 3 discusses the dataset used in this paper and the pre-processing procedure. Section 4 presents the approach employed in this paper which is an unsupervised and supervised method with different embedding techniques. Section 5 presents the results of the unsupervised and supervised methods with BERT and TF-IDF embedding and their conclusions. Finally, Section 6 provides a fundamental discussion, and concludes and summarizes this paper.

For ease of reading, Table 1 provides a list of abbreviations that are commonly used in this paper.

Table 1: *List of abbreviations*

Abbreviation	**# Meaning**
BERT	Bidirectional Encoder Representations from Transformers
LaBSE	language-agnostic bert sentence Embedding
NLP	Natural Language Processing
NER	Named Entity Recognition
PD	Participatory Design
AAIRL	Architectural Artificial Intelligence Lab
RASA	Real-time Application Support and Automation
NLU	Natural Language Understanding
LLM	Language Learning Model
TF-IDF	Term Frequency-Inverse Document Frequency
NLI	Natural Language Inference
STSb	Semantic Textual Similarity Benchmark
IS-BERT	Info Sentence BERT
STS	Semantic Textual Similarity
GPT	Generative Pre-training Transformer
MLP	Multi-Layer Perceptron
ReLU	Rectified Linear Unit
POS	part-of-speech

2. Related Work

Eran The topic of chatbots has been explored in quite a few works. In [26] the authors proposed a way to create smart chat-bots that can better understand what the user is asking for. They do this by identifying key information in the user's input and using it to improve the chatbot's ability to recognize the user's intention. The method uses a combination of RASA NLU and neural network techniques. RASA NLU is a library for natural language processing. It extracts entities from user input by using techniques like tokenization [27], POS tagging [28], and NER. In our work, we use these similar methods to extract entities and keywords that relate to our 5 categories.

In [29] describes the development of an academic chatbot that uses natural language processing (NLP) and entity extraction to provide students with access to academic information. The chatbot was developed at ISB Atma Luhur, a university in Indonesia, and is designed to be used by students on mobile devices. The chatbot uses NLP to understand the intent of a student's query and to extract entities from the query. For example, if a student asks 'What is the course schedule for next semester?' The chatbot will use NLP to understand that the student is looking for information about course schedules. The chatbot will then use entity extraction to identify the entities in the query, such as "course schedule" and "next semester". When users will use our chatbot we will need to understand what they say and in which category to put it among the 5 we have. Understanding chatbot users is critical. If the users feed the chatbot with unrelated things we must filter them accordingly or alternatively give them an appropriate response based on what they write to the chatbot.

We can see another approach in [30] where the authors discuss the different approaches to knowledge entity extraction and text mining. One approach is to use supervised learning. Supervised learning is a machine learning technique that requires labeled data. Labeled data is data that has been manually annotated with the correct answers. Supervised learning can be used to train a model to extract knowledge from text. Another approach is to use unsupervised learning. Unsupervised learning is a machine learning technique that does not require labeled data. Unsupervised learning can be used to find patterns in unlabeled data. These patterns can be used to extract knowledge from text. In this article, both of the techniques will be used.

When we talk about the benefits of crowdsourcing in the context of urban places there are also possibilities like in [31] when they used to collect data in smart cities with mobile crowdsourcing. Mobile crowdsourcing is a type of crowdsourcing that uses mobile devices to collect data from a large number of users, Just like in our work. This data can be used

to improve a variety of urban services, such as traffic management, public safety, and environmental monitoring and conditions, such as air quality and water quality.

Daniel The literature on public participation and integration of citizens inputs in plans is vast. In [32] The authors believe cities Should be increasingly planned bottom up employing more inclusive approaches, and labeling top down processes as outdated and ineffective in meeting contemporary urban living demands. In another study [33], the authors implemented a four-step methodology to design and evaluate chatbots. They collected a dataset of user self-reported data on various topics, trained language models on this dataset, developed chatbots powered by the trained models, and evaluated the chatbots by having users interact with them. The findings revealed that these chatbots effectively collected data in a natural and engaging manner, producing accurate and reliable results. Importantly, they reached a broader user base, including individuals who may have been hesitant to participate in traditional surveys. The study suggests that employing language model-powered chatbots can significantly enhance the efficiency and effectiveness of data collection, potentially revolutionizing the field by facilitating greater data sharing and understanding of human behavior.

Renana Rimon In a different context, in [34] the focus is on classifying words and expressions into data categories. That is, the classification is not in the context of a sentence but rather for a specific word. The classification is done using the following methods: Term Frequency - Inverse Document Frequency (TF-IDF) [35], and Linear Support Vector Machine [36]. The results of the research showed that the linear support vector model achieved the highest accuracy rate of up to 96.82%.

Sentence-BERT (SBERT) [37] is a modification of the BERT (Bidirectional Encoder Representations from Transformers) model [38] that is designed to generate fixed-length sentence embedding. The main idea behind SBERT is to fine-tune the pre-trained BERT model on a sentence-level task, such as sentence similarity, instead of a word-level task, as done in BERT. SBERT generates sentence embeddings by fine-tuning a pre-trained BERT model on a sentence-level task, using a Siamese or triplet network architecture, various pooling strategies, and data augmentation techniques. The SBERT sentence embedding were evaluated using the SentEval toolkit [39], which is a popular tool for assessing the quality of sentence embeddings. The evaluation yielded an impressive score of 87.69% for the SBERT-NLI-large model. This model was pre-trained on NLI datasets and then fine-tuned on the STSb (Semantic Textual Similarity Benchmark) dataset [40]. The model has good results, but its limitation is the lack of multilingualism. Each model is trained for only one language.

Taliya Shitreet The paper [41] proposes a method called Info-Sentence BERT (IS-BERT) for unsupervised sentence representation learning. The goal is to learn meaningful sentence embeddings without relying on labeled data. The method builds on top of BERT and uses a combination of 1-D convolutional neural networks (CNNs) and a new self-supervised learning objective based on mutual information maximization to derive meaningful sentence embeddings. The approach maximizes the mutual information between sentence-level global representations and token-level local representations, using a discriminator to decide whether pairs of sentence and token embeddings are from the same sentence or not. Experimental results show that IS-BERT significantly outperforms other unsupervised sentence embedding baselines on common semantic textual similarity (STS) tasks and downstream supervised tasks. The paper evaluated the performance of their proposed IS-BERT model using the SentEval toolkit. The evaluation showed that the IS-BERT model achieved an accuracy of 85.91% .

3. Datasets

Renana Rimon The research aims to reach a diverse and broad audience to obtain the most comprehensive and genuine public opinion. When working on a project based on collective intelligence [42], it is crucial to gather abundant data that represents the true target audience. However, collecting data, especially in a domain lacking tagged data, is not a straightforward task. The data needed consists of conversational sentences, such as WhatsApp messages, in the architecture topic. To collect this data, the researchers utilized a chatbot based on chatGPT, designed to engage in conversations with residents and gather as much relevant information as possible regarding the topic of constructing a public structure that suits the real needs of the residents. The initial dataset consisted of several dialogues conducted through the chatbot. The chatbot conversed with the client as an architect, asking targeted questions. The initial dataset comprised approximately 850 sentences from both the architect and the customer. The focus was on the responses received from the customer, totaling 360 sentences. The initial dataset size was relatively small, posing a challenge for the model to learn and accurately label the data. The conversations were held with a wide spectrum of people - women, men, young and old, as well as a broad range of socio-economic statuses, ranging from academics to illiterates. Individuals with low socio-economic status constitute the majority of the population in Bogotá Savanna, with low levels of education or no education at all. Consequently, it was not un-

common to receive irrelevant answers, answers with spelling errors, or incomplete sentences. The goal was to obtain clean data, containing only the relevant sentences. Furthermore, the field of architecture is extensive and encompasses various topics. Each sentence was to be labeled with the appropriate category it belongs to, including environment and climate resilience, mobility (transport), local identity, future of work, or land use. A sentence may be related to multiple categories or none of them, in which case it would be labeled as "other". Since the data was collected in an unlabeled format, suitable approaches were required to deal with it. Both unsupervised and supervised approaches were considered. The supervised approach involved data labeling. For the initial 300 sentences, could be considered to label the data manually, which was feasible for the initial small dataset, but would require a more efficient method as more data was accumulated. Therefore, an innovative approach for data labeling was proposed. The labeling process would be performed by utilizing chatGPT. Each sentence and the five categories would be inputted to chatGPT, which would then assign the sentence to the relevant categories. It was emphasized to ChatGPT that each sentence could belong to more than one category or none of them. If chatGPT did not assign any category to a sentence, indicating its irrelevance, it would be labeled as "other" [Table 2].

Table 2: *Sentences and Categories*
Part of the dataset consists of sentences labeled by the chatGPT.

Sentence	**Category**
"thank you and bye"	['other']
"Where will the building be located?"	['Mobility (transport)']
"safe"	['Environment and climate resilience']
"to create a public space in front of the building"	['Land use', 'Local identity']
"Perhaps more accessible entrances larger public spaces more communal areas"	['Environment and climate resilience', 'Land use']
"Could you describe in your own words the example I provided?"	['Future of work']
"There should be many student gathering areas."	['Land use']
"I hope there will be some green plants in the building, plants always make people feel more comfortable"	['Environment and climate resilience']
"I want a garden on the roof"	['Land use']
"The building should be made of recyclable materials"	['Environment and climate resilience']

Given the multi-labeled data, despite having 360 examples, the number of labels was 405. The data was divided into train, and test sets, resulting in a very small number of examples in each set. This small dataset size could potentially hinder the model's ability to learn effectively. Additionally, the limited number of examples in the test set made it challenging to measure the results accurately, although some indicative results could still be observed. The dataset prepared for the research included 360 client sentences with labels generated by ChatGPT, which were considered the ground truth for the research purposes [Table 3], [Table 4].

Table 3: *Data Distribution*
Statistics of number of sentences & number of labels in the data. The data is multi-labeled, so #labels >#sentences. The statistic show in the table is 1) all data. 2) run with split 80% train, 20% test.

	Total # of sentences	**Total # of labels**
All data	360	405
Train (80%)	288	325
Test (20%)	72	80

The research utilizes an additional dataset for training a Named Entity Recognition (NER) model to extract entities. The dataset used is the GMB (Groningen Meaning Bank) corpus [43], which has been annotated and tagged for entity classification. Natural Language Processing techniques are applied to this dataset, incorporating enhanced and popular features. The GMB corpus provides context and serves as a valuable resource for training the classifier to predict named entities such as names, locations, and more. The dataset contains essential information about different types of entities, including geographical entities, organizations, persons, geopolitical entities, time indicators, artifacts, events, and natural phenomena. The total word count in the dataset is 1,354,149.

4. Framework

Renana Rimon In order to train a deep learning model to classify sentences into different categories, there are two approaches to solving the problem. One approach is the unsupervised method, where untagged data can be used to train the model. In the unsupervised method, a smaller variant of the LaBSE (Language-agnostic BERT Sentence Embedding) model was employed to generate sentence embeddings. To prepare the input for the model, we used a NER (Named Entity Recognition) model to extract entities from the sentence. For each sentence, the input to the LaBSE model consisted of three versions: 1. The original sen-

Table 4: *Data Distribution by Category*
Statistics of number of labels For each label category, The statistic shown in the table is 1) all data. 2) run with split 80% train, 20% test.

	Other	**Local Identity**	**Future of Work**	**Environment and Climate Resilience**	**Land Use**	**Mobility (Transport)**
All data	88	73	48	76	108	12
Train (80%)	70	59	39	61	86	10
Test (10%)	18	14	9	15	22	2

tence. 2. A sentence that includes only the entities from the NER 3. Send each noun by itself and then average them all for each category divided by the number of nouns that were in the sentence. These embeddings capture the semantic meaning of the sentences in a numerical representation. By calculating the cosine similarity between the sentence embeddings and predefined categories, the method determines the relevance or similarity of the sentences to each category.

The second approach is the supervised method, which requires labeled data. The research proposes an innovative approach to labeling the data using the ChatGPT model. The input provided to the model is a sentence and the five categories that represent subcategories in the architecture domain. It is emphasized that the labeling is multi-label, meaning each sentence can be classified into multiple categories or none of them. If ChatGPT does not assign the sentence to any of the categories, the label will be "other," indicating it is not relevant to the architecture domain.

When labeled data is available, various deep learning models can be trained. The research utilizes two approaches for sentence embedding: the first is Language-agnostic BERT, and the second is TF-IDF. After embedding the sentences using these approaches, the data enters a fully-connected neural network for training.

To evaluate the model's results using different metrics such as accuracy, precision, recall, and F1 score, the data is initially divided into train and test sets. After training the model, the test data can be fed into the model for prediction and evaluation of the results. On the other hand, in the unsupervised approach, there are seemingly no labeled data. To evaluate the results, the ChatGPT's labeling is considered as the ground truth and compares the results.

In Section 4.1, we present the unsupervised solution, while in Section 4.2, we demonstrate the supervised solution, which includes the two embedding methods. Section 4.3 describes how we simulate a scenario where the unsupervised solution seemingly received ground truth labeling to evaluate the results.

4.1. Unsupervised Method

Eran When we talk about the unsupervised method in the context of our research, we will first explain 2 main things that are important for the rest of the article. The first one is Named Entity Recognition (NER) which is a sub-task of information extraction that identifies and classifies named entities in text into predefined categories such as person names, organizations, locations, etc.

Daniel The NER model employed in this project follows a straightforward process. It begins by taking a sentence as input and proceeds with tokenization, breaking the text into individual words and phrases known as tokens. These tokens are then fed into the BERT model, which generates a series of hidden states representing the model's understanding of the input text. The subsequent step involves passing the hidden states through a classifier component, responsible for predicting the entity type for each token. This classifier is trained on a labeled dataset containing the correct entity types for each token. Taking advantage of its extensively trained classifier, the NER model accurately identifies named entities within the given text, demonstrating exceptional accuracy in classifying these entities.

In this project, after assigning each word to its respective POS, the model selectively extracts and saves the identified nouns for further analysis and processing. Importantly, the original sentences are retained alongside the extracted nouns, enabling a comprehensive understanding of the context in which the nouns are found. During the training phase, the NER model utilizes a technique known as fine-tuning, which involves customizing a pre-trained BERT model specifically for the task of named entity recognition. The fine-tuning procedure initiates by tokenizing the input text, breaking it into individual words and phrases. The BERT model processes the resulting tokens, generating hidden states that capture the model's understanding of the input text. These hidden states are then fed into a classifier, which predicts the entity type for each token. To train the classifier, a labeled dataset is used, containing the correct entity types for each to-

ken. Supervised learning is utilized to train the classifier. In this process, the model is presented with input-output pairs, where the input pairs represent the text tokens, and the output pairs indicate the correct entity types. The model is then trained to predict the accurate entity type for each input token.

The second main topic is Language-Agnostic BERT Sentence Embeddings (LaBSE)[5.1.2]. ¡replacement command¿

The unsupervised algorithm consists of several steps. In the initial stage, NER is utilized to extract nouns from the dataset. Next, a smaller-LaBSE model is employed to compute the proximity between sentences and pre-selected categories.

This stage involves three parts:

- Assessing the sentence's proximity to each category.
- Evaluating the proximity of the noun sequence to each category.
- Determining the average proximity of each noun in the sentence to each category.

Each part aims to determine the degree of sentence-category association. To achieve this, the sentences are encoded using a pre-trained model, generating vector representations. These vectors are then compared to the category vectors using cosine similarity, which measures their similarity based on orientation. A manually defined threshold is applied to determine the results, ensuring optimal performance of the unsupervised models. Note that the threshold plays a crucial role in optimizing the unsupervised models by controlling the level of association between sentences and categories. The category "other" was assigned to each sentence that did not receive any category during labeling.

The selection of an appropriate distance calculation technique is of utmost importance as it significantly impacts the outcomes of clustering. Depending on the nature of the data and the research objectives, different dissimilarity measures may be more suitable. For example, when dealing with numerical data and aiming to assess the distance between points in a coordinate system, Euclidean distance serves as a valuable tool, providing insights into the similarity or dissimilarity based on numerical attributes. However, Euclidean distance may not be well-suited for high-dimensional data, such as text data represented as high-dimensional vectors with each dimension corresponding to a unique word or feature. In contrast, cosine similarity measures similarity by considering the angle between vectors, which remains stable and meaningful in high-dimensional spaces.

Cosine similarity offers several advantages over Euclidean distance. Firstly, it is not influenced by the magnitude or length of vectors being compared, focusing solely on their direction or orientation. This property is particularly useful in text classification tasks where the frequency of words may vary widely, but their importance for classification is not necessarily related to their frequency. Cosine similarity enables effective comparisons based on the relevance of shared words, irrespective of their frequency.

Moreover, cosine similarity has been shown to capture semantic similarity better than Euclidean distance for text data. By focusing on the orientation of vectors, cosine similarity can identify similarities in terms of the meaning or context of words, even if their actual representations differ. This characteristic makes it well-suited for tasks such as document similarity, information retrieval, and recommendation systems.

Additionally, cosine similarity overcomes a challenge encountered in text data, which often exhibits sparsity where most dimensions or features have zero values. In such cases, Euclidean distance may exhibit bias towards documents with more non-zero values, whereas cosine similarity remains unaffected by sparsity. It allows for effective comparisons even when a majority of dimensions are zero, as it considers only the non-zero dimensions.

In contrast, Manhattan distance, another distance calculation technique, focuses on the absolute differences between coordinates rather than capturing the semantic relationship between text documents accurately. Hence, cosine similarity is preferred over Manhattan distance in text-related tasks. The cosine similarity metric is computed using the following equation:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Here, $(A \cdot B)$ represents the dot product of vectors A and B, while $||A||$ and $||B||$ represent the Euclidean norms or magnitudes of vectors A and B, respectively. Cosine similarity values range from -1 to 1, with 1 indicating perfect similarity, 0 indicating orthogonality or no similarity, and -1 indicating complete dissimilarity.

Eran In our implementation of the unsupervised method we have 3 approaches that we mention above. For each option, we will check to which category the sentence belongs (by calculating the computational distance between the vector representing the sentence and the vectors representing the categories). And In the end, we will get a csv file that will assign each sentence to a certain category or more.

After we get an output we need to check how close we were to the tag that the ChatGPT outputted using the supervised method.

In order to do this we will compare the results of the 2 methods. In other words, we will compare the categories we received in each option we mentioned

above using the metrics F1 score, Precision, Accuracy, and Recall.

4.2. Supervised Method

Taliya [Figure2] Catering to a multi-class [44] and multi-label [45] classification problem. Data, labeled using ChatGPT technology, served as the basis for the supervised solution. The primary challenge was assigning appropriate categories to an incoming, unseen sentence, a classic text classification problem. To manage this problem, an algorithm under supervised learning methods was employed. Data preprocessing was an essential first step. The rows representing sentences without any labeling were removed, as these null values could distort the learning process and impact the normalization level of machine learning algorithms. Further, to prevent the biasing of the learning process due to duplicate sentences, all duplicates were carefully identified and removed. A method to category analysis revealed 14 categories appearing only once and a total of 31 unique categories in the dataset. Instead of discarding these rare categories, which could have led to potential information loss, the categories were split into different rows, each becoming associated with a particular sentence. This augmented the dataset, addressing the challenge of an uneven category distribution. Following preprocessing, the refined dataset included 360 non-duplicate, well-prepared samples, each associated with their respective categories.

For the next stage of processing, the data was split into training and testing sets. The data was encoded to become compatible with deep learning models. The model used TensorFlow's StringLookup function to transform the categories into a multi-hot format - an ideal choice for a multi-label classification task. The process also included an inversion function to decode the multi-hot encoding back into the original categories, which aided model interpretability and validation. Sentence embedding is a crucial step in transforming human language into numerical representations that can be processed by machine learning models. In this framework, the transformation was achieved using three distinct text vectorization techniques: TextVectorization TF-IDF 4.2.1 and BERT Sentence4.2.2. The final stage involved constructing a deep learning model based on a Multi-Layer Perceptron (MLP) architecture. The MLP model, also known as a feed-forward neural network. In this model, four hidden layers were used, consisting of 512, 256, 128, and 64 neurons respectively. Each hidden layer used a Rectified Linear Unit (ReLU) activation function [46]. The ReLU activation function introduces non-linearity into the network, enabling the model to learn and solve complex problems, and is often a default choice for hidden layers. The output layer, on the other hand, used a sigmoid activation function [47]. The sigmoid function, also known as the logistic function, maps real-valued numbers into the range between 0 and 1, which can be used to represent probabilities for binary and multi-label classification problems. In the case of this multi-label text classification task, the sigmoid function allowed the model to assign a separate probability to each category, indicating the likelihood of an input sentence belonging to that specific category. The model's performance was evaluated during the training and testing phases using several metrics: Binary Accuracy, Precision, Recall, and F1 score. [48] The model was trained for 16 epochs, with a loss function set to "binary crossentropy" [49], considering the multi-label nature of the classification task. Adam Optimizer [50], a popular adaptive learning rate optimization algorithm, was applied.5

4.2.1. TF-IDF

The TextVectorization layer from TensorFlow was employed to transform text into a vector representation using the Term Frequency-Inverse Document Frequency (TF-IDF) method. In this process, each sentence in the dataset was tokenized into terms (words or 2-grams), and a TF-IDF score was calculated for each term within the sentence. The score evaluates the importance of a term within a sentence in the context of the whole dataset, thereby reflecting the significance of the term in representing the sentence's content.

Mathematically, TF-IDF is calculated as follows:

Term Frequency (TF) is calculated as the count of a term 't' in a document 'd' (a sentence, in this case) divided by the total number of terms in the document 'd':

$$TF(t,d) = \frac{\text{count of term } t \text{ in document } d}{\text{number of terms in document } d}$$

Inverse Document Frequency (IDF) is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term 't':

$$IDF(t) = \log_e \left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}} \right)$$

The **TF-IDF** score is simply the product of the TF and IDF scores of the term:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

This TF-IDF score is assigned to each term in each sentence of the dataset, thereby converting each sentence into a vector of TF-IDF scores.

The strength of TF-IDF lies in its simplicity and its ability to highlight document-specific significant words. It excels in reducing the influence of common

words, thereby allowing the model to focus on more unique and category-specific words.

However, TF-IDF does have limitations. It doesn't consider semantic relationships and the context of words, which could impact the classification of complex texts where understanding sentence semantics is crucial. It treats all words equally and ignores the order of words in a sentence, potentially missing important language nuances.

Moreover, TF-IDF struggles with multilingual data as it is language-specific. This is a crucial limitation in this work with a multilingual dataset; it may not fully capture the complexities and variations of multilingual text data. Despite these, TF-IDF serves as a reliable foundational approach for text vectorization.

4.2.2. BERT Sentence

Another approach was adopted to generate sentence embeddings using the BERT-Base-NLI-Mean-Tokens model. This model, designed to yield robust sentence embeddings, offers vector representations of sentences that encapsulate their semantic meaning and structural characteristics.

The conversion of a sentence into its vector representation using BERT-Base-NLI-Mean-Tokens is a straightforward process. Firstly, the input sentence s is tokenized into a sequence of subword units utilizing the WordPiece tokenization strategy [51], denoted as $T(s)$. The BERT-Base-NLI-Mean-Tokens model, built on the transformer-based neural network architecture, then computes the average of the token embeddings to obtain the sentence embedding.

One advantage of BERT models, including BERT-Base-NLI-Mean-Tokens, is their ability to handle new words that were not seen during training. The contextual nature of BERT allows it to generate meaningful representations for words based on the surrounding context, even if it has not encountered those specific words before.

The size of the output vector in BERT-Base-NLI-Mean-Tokens is a fixed length of 768 dimensions. This vector representation captures the syntactic and semantic properties of the sentence, making it valuable for diverse Natural Language Processing (NLP) tasks including text classification, sentiment analysis, and information retrieval.

After the sentence embeddings are generated, they can be further normalized using techniques such as L2-normalization [52]. This normalization ensures uniformity of scale across all embeddings, enhancing their effectiveness during model training.

These generated sentence embeddings, enriched with semantic and contextual information, can be employed to measure semantic similarity between sentences, facilitate multilingual translations, or assist in any NLP task that requires a comprehensive representation of the input text.

$$N(v) = \frac{v}{\|v\|}$$

Here, N denotes the normalization function, and $\|v\|$ represents the L2 norm of the vector v. This normalized sentence representation $N(v)$ ensures uniformity of scale across all embeddings, thereby enhancing their effectiveness during model training.

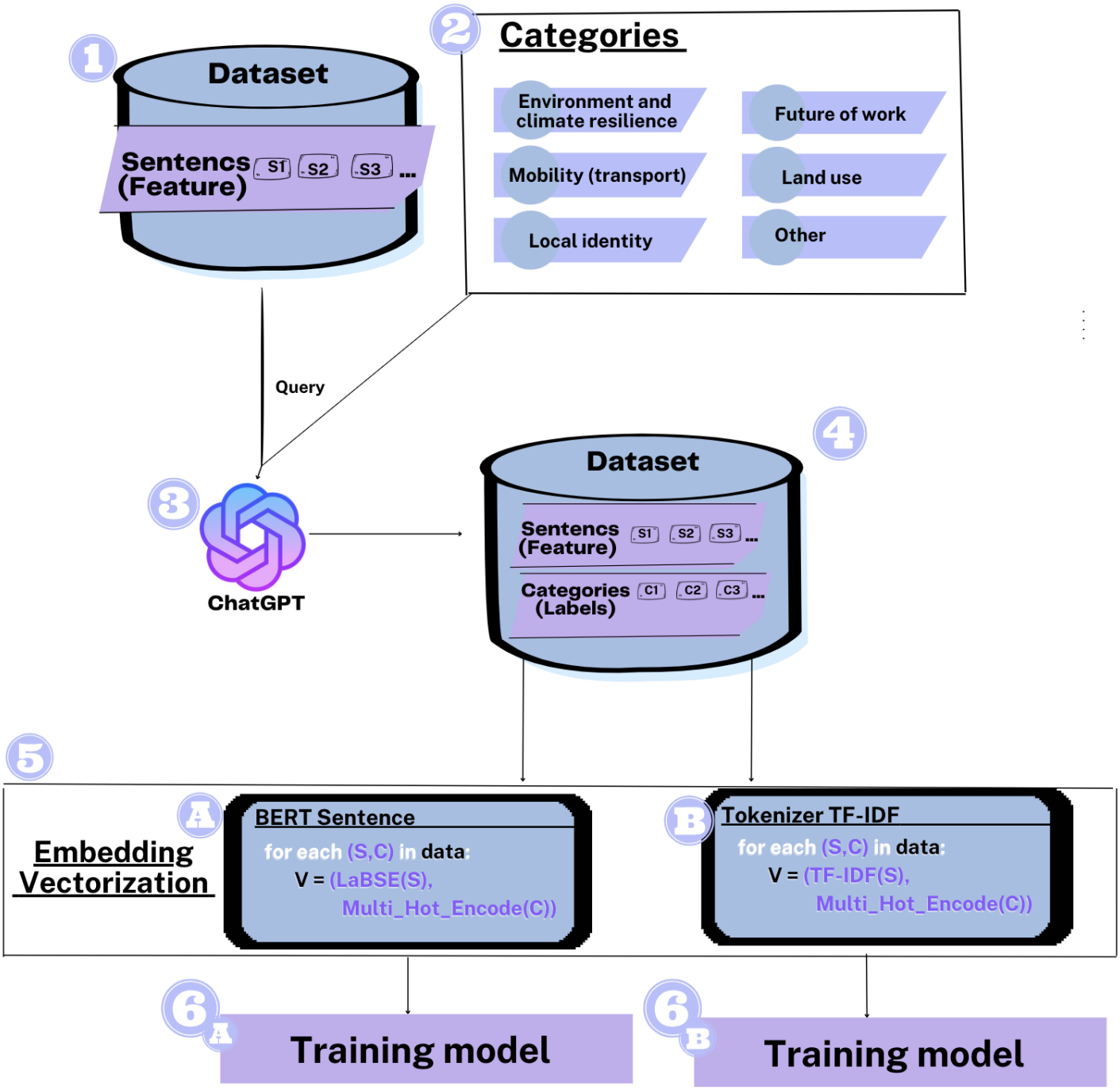

Figure 2: *This figure provides an illustration of the* ***Supervised learning process****. Each original sentence 's' from the dataset (sourced from chat conversations by Pro-Bogota) undergoes a query with six categories for tagging by ChatGPT, resulting in a tagged dataset. This dataset, consisting of a feature column of sentences and a corresponding multi-label category column, undergoes embedding via two distinct methods: Language-agnostic BERT Sentence Embedding (LaBSE), which performs multilingual sentence vectorization considering the sentence's context, and Term Frequency-Inverse Document Frequency (TF-IDF), which performs sentence vectorization considering word frequency, without reference to context. In both methods, the labels are vectorized using a multi-hot encoding scheme. Once the numerical data is available, it serves as input for a fully-connected neural network, which undergoes a learning process. The network output consists of six probabilities, each representing the likelihood of a sentence being associated with a specific category.in [Figure3] you can see the process a sentence go through this method.*

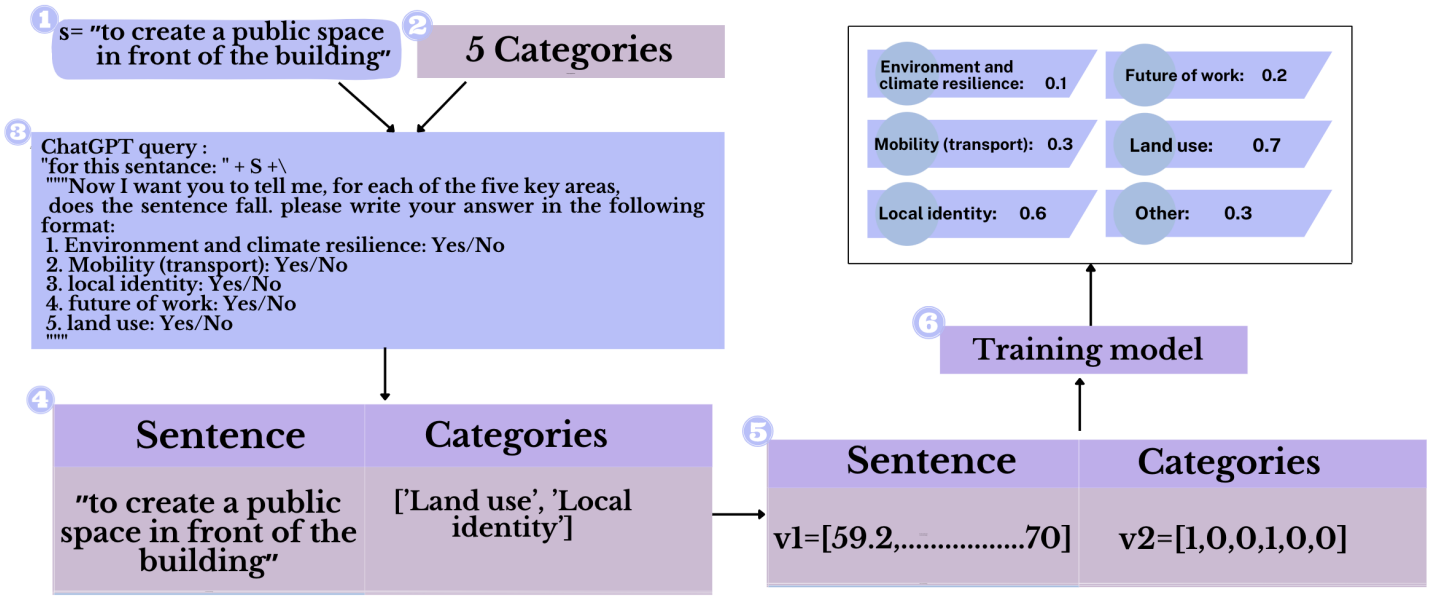

Figure 3: *An illustration of a process where a specific sentence input passes through the supervised learning model, and the output yields six different probabilities for classifying the sentence into each category.*

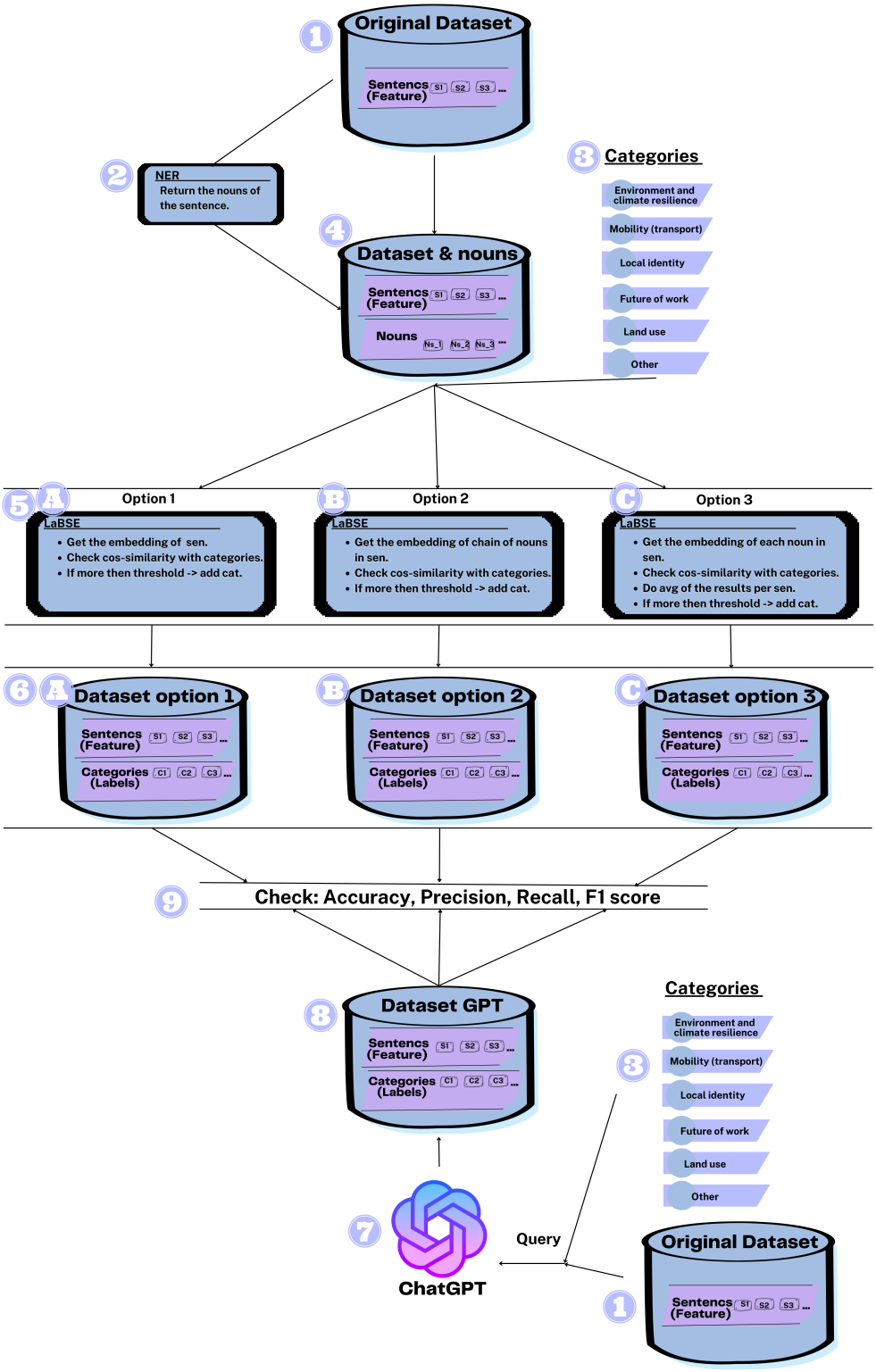

Figure 4: *This figure provides an illustration of the* ***Unsupervised learning process****. In this process, original sentences from a data set are passed through a Named Entity Recognition (NER) model to extract the nouns within each sentence. The LaBSE model is used for the embedding process. The unsupervised learning process offers three options for performing embedding: Option 1: Embedding the complete sentence. Option 2: Embedding a chain of nouns within the sentence. Option 3: Embedding each noun separately and averaging the results. After the embedding step, the cosine similarity is calculated between the vector representing the sentence and the vector representing each category. If the similarity result exceeds a certain threshold, the sentence is associated with that particular category. Consequently, for each option, a data set is generated, comprising sentences as features and their corresponding categories as labels. These data sets can then be compared against a data set labeled by ChatGPT (truth ground) using evaluation metrics such as accuracy, precision, recall, and F1 score.*

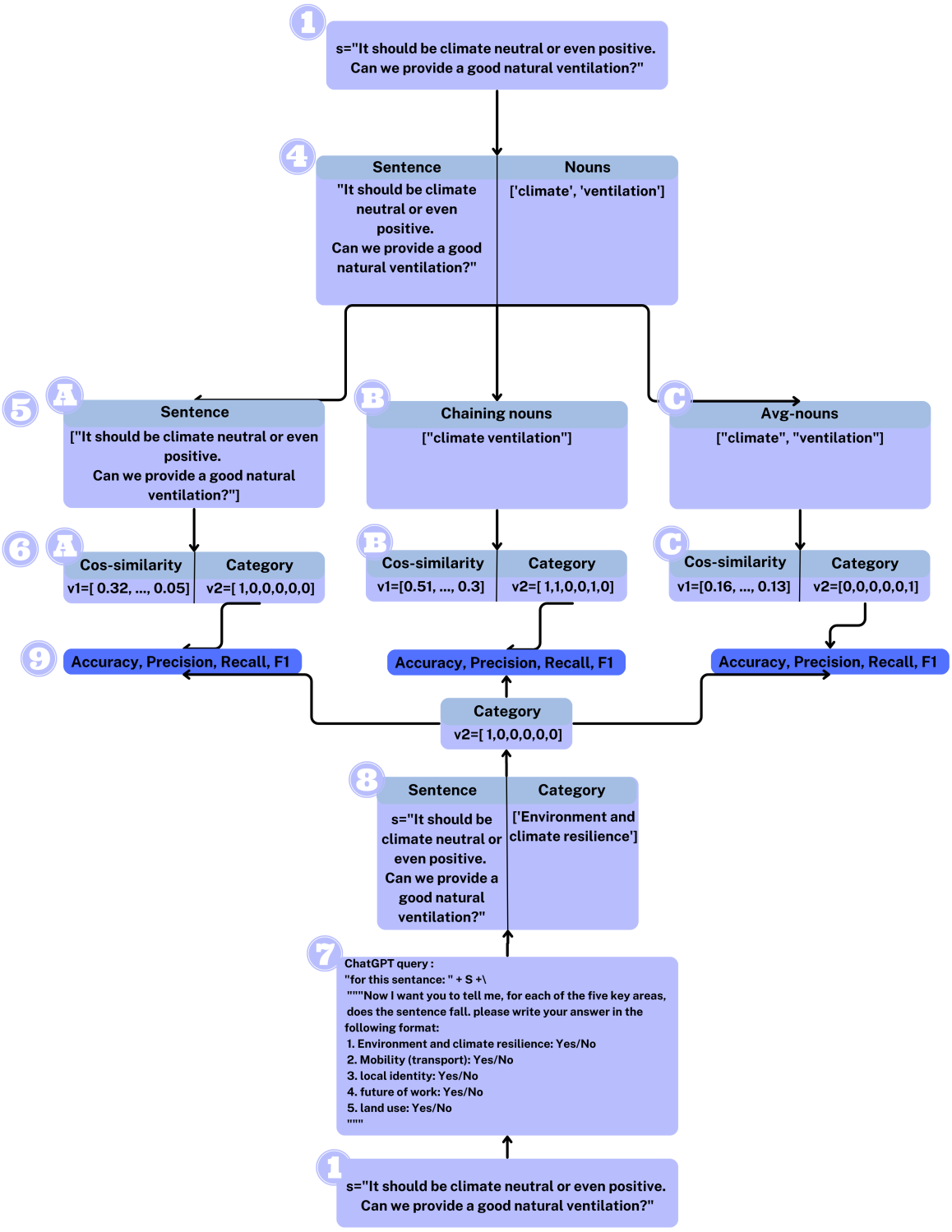

Figure 5: *An illustration of a process where a specific sentence input passes through the unsupervised learning model, and the output yields 3 performance tables (Accuracy, precision, recall, F1 score).*

5. Experimental Evaluation

Eran There are a few ways to evaluate the metrics, which are the results of the model, that are commonly used:

- **Accuracy** Is a measure of how often the model makes correct predictions. It tells us the percentage of instances that the model classified correctly out of all the instances it predicted. For example, if a model has an accuracy of 85%, it means that it correctly predicted the outcome for 85 out of every 100 instances.

  The accuracy is given by the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

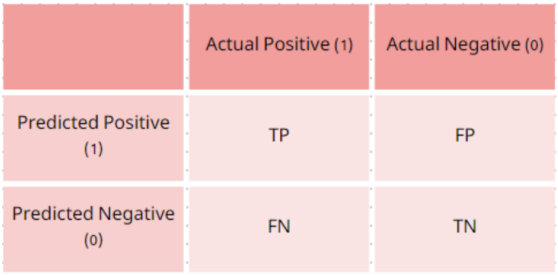

Figure 6: *Confusion matrix*

  where TP (True Positives) is the number of instances correctly classified as positive, TN (True Negatives) is the number of instances correctly classified as negative, FP (False Positives) is the number of instances incorrectly classified as positive, and FN (False Negatives) is the number of instances incorrectly classified as negative.

- **Precision** Precision is a measure of how accurate a classifier is when it predicts positive instances. It focuses on the proportion of correctly predicted positive instances out of all instances that the classifier labeled as positive. It is given by the formula:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

  For example, if the classifier predicts that 100 emails are spam, and it correctly identifies 90 of them as spam while misclassifying 10 non-spam emails as spam, the precision would be 90%. It means that out of all the emails the classifier labeled as spam, 90% of them were indeed spam.

- **Recall** Recall is a metric that measures the ability of a classifier to find and correctly identify all positive instances. It tells us the proportion of actual positive instances that the classifier correctly detects.

  A high recall means that the classifier is able to find most of the positive instances, while a low recall indicates that the classifier is missing a significant number of positive instances.

  The formula for the recall is:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- **F1 score** The F1 score takes both precision and recall into account and gives you a balanced measure of the model's performance. It's like finding a middle ground between precision and recall.

  A higher F1 score means the model is doing well in both correctly identifying positive cases and finding all the positive cases. A lower F1 score indicates that the model may be lacking in one or both of these areas.

  F1 is calculated using the harmonic mean of precision and recall. The formula for F1 is as follows:

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

5.1. Unsupervised Experiment and Conclusion

In option 1 [5.2] with a threshold of 0.25 the result are:

Table 5: *Performance Metrics for Different Categories presents the performance metrics of option 1 when we send to the model all the sentences, including accuracy, precision, recall, and F1 score, for our 6 categories. The data highlights variations in these metrics across the categories.*

Category	Accuracy (%)	Precision (%)	Re
Environment and climate resilience	71.71	73.68	
Mobility (transportation)	40	27	
Local identity	80.85	63.63	
Future of work	80.5	100	
Land use	82.57	20	
Other	92.28	10.52	

When looking at category number 1 the category has a relatively high accuracy of 71.71% and a precision of 73.68%, meaning that the model is fairly good at correctly predicting this category. However, the recall is quite low (13%), implying that the model is not

capturing all relevant sentences that should be classified under this category. The low F1 score (22.04%) also suggests an imbalance between precision and recall, primarily due to the low recall.

In category 2 the model's performance is mixed. It has a lower accuracy (40%) and precision (27%) but very high recall (80%). This suggests that while the model is good at identifying sentences that relate to mobility, it's likely classifying too many sentences into this category, resulting in lower precision. The balanced F1 score (40.3%) reflects this trade-off.

Category 3 shows high accuracy (80.85%) but lower precision (63.63%) and recall (27.6%). The high accuracy suggests that the model is generally good at identifying whether a sentence is about local identity or not, but it's not as successful in correctly classifying these sentences (reflected by the lower precision) or identifying all relevant sentences (reflected by the lower recall). This is also mirrored in the moderate F1 score (38.53%).

In 4, the category shows high accuracy (80.5%) and perfect precision (100%), but extremely low recall (8.10%), leading to a very low F1 score (15%). This suggests the model is highly precise when it classifies a sentence as relating to the future of work - it's always right - but it's missing a lot of sentences that should be categorized as such.

In 5, the model has high accuracy (82.57%) but low precision (20%), recall (8.17%), and F1 score (11.6%) in this category. This means that while the model can accurately predict whether a sentence is about land use or not, it struggles to correctly classify these sentences (low precision) and capture all relevant sentences (low recall). The last one, category number 6 has the highest accuracy (92.28%) but low precision (10.52%) and moderate recall (16.66%), leading to a low F1 score (12.9%). This suggests that the model is good at determining if a sentence does not fit into the other five categories, but it may be over-classifying sentences into this "other" category, reducing precision.

In conclusion, the model seems generally accurate but struggles with precision and recall in most categories, particularly recall. This indicates that it's often able to correctly identify when sentences are not related to a specific category (high accuracy) but struggles to identify all the relevant sentences that should be classified under that category (low recall), and often misclassified sentences (low precision).

For option 2 [5.2] with a threshold of 0.27 the results are:

For category 1 **Environment and climate resilience** the accuracy dropped from 71.71% to 65.71%, and precision dropped from 73.68% to 42.85% compared to option 1. However, the recall has increased from 13% to 33.33%, resulting in an increased F1 score of 37.5%. This shows that the model identified more relevant sentences for this category but also made more mistakes in classifying sentences.

Category	**Accuracy (%)**	**Precision (%)**	**Re**
Environment and climate resilience	65.71	42.85	
Mobility (transportation)	53.71	32.33	
Local identity	81.14	70.83	
Future of work	76	32.14	
Land use	75.42	21.5	
Other	87.14	7.7	

Table 6: *Performance Metrics*
presents the performance metrics of option 2 when we send to the model all nouns in a concatenated string. The table includes accuracy, precision, recall, and F1 score, for our 6 categories. The data highlights variations in these metrics across the categories.

For **Mobility (transportation)** there's a significant increase in accuracy (from 40% to 53.71%) and precision (from 27% to 32.33%). Recall is slightly lower (from 80% to 73.33%), but the F1 score increased from 40.3% to 44.89%. This indicates improved model performance for this category.

For **Local identity** accuracy increased slightly from 80.85% to 81.14%. However, precision increased (from 63.63% to 70.83%), while recall dropped significantly (from 27.6% to 22.36%), resulting in a decrease in F1 score to 33%. The model seems to be better at identifying non-relevant sentences but worse at identifying all relevant sentences.

For **Future of work** accuracy decreased slightly (from 80.5% to 76%), and precision significantly dropped from 100% to 32.14%. Recall also decreased from 8.1% to 12.16%. Consequently, the F1 score increased to 17.64%. Despite lower precision, the model is now better at identifying relevant sentences.

For **Land use** accuracy dropped from 82.57% to 75.42%. Precision slightly increased from 20% to 21.5%, but recall saw a significant increase from 8.17% to 28.57%, leading to an increased F1 score of 24.45%. The model is better at identifying relevant sentences but makes more mistakes in classifying sentences.

And for the last category, **Other** accuracy dropped from 92.28% to 87.14%. Precision significantly decreased from 10.52% to 7.7%, and recall increased from 16.66% to 25%, resulting in a slight decrease in the F1 score to 11.76%. The model is better at identifying relevant sentences but makes more mistakes in classifying sentences.

In conclusion, option 2 seems to have improved recall in most categories but often at the expense of precision, except for Category 2 (Mobility), where

all metrics improved. This suggests that focusing on nouns in sentences allows the model to capture more instances of each category, but it also leads to more misclassifications. The overall trade-off depends on the specific importance of precision and recalls for your task. If finding all relevant sentences (even at the risk of including irrelevant ones) is the goal, then option 2 may be a good choice.

For option 3 [5.2] with a threshold of 0.29 the results are:

for **Environment and climate resilience (Category 1)**: The accuracy has dropped significantly to 40.86% and precision to 32.99%. However, the recall has significantly increased to 88.89%, resulting in a higher F1 score of 48.12%. This shows the model captures a higher number of relevant sentences but makes more mistakes in classifying sentences.

And for **Mobility (transportation) (Category 2)**: The accuracy (76.57%) and precision (57.14%) have increased while recall decreased to 35.56%. The F1 score slightly decreased to 43.84%. The model shows improved performance in predicting and classifying sentences, but it's missing out on capturing more relevant sentences.

Local identity (Category 3): Accuracy dropped to 62.86%, precision significantly decreased to 21.88% but recall increased to 27.63%. The F1 score decreased to 24.42%. The model is better at identifying relevant sentences but is making a lot more classification mistakes.

Future of work (Category 4): Accuracy decreased to 39.71%, precision dropped to 20.35% but recall significantly increased to 63.51%. The F1 score almost doubled to 30.82%. Despite the lower precision, the model is much better at identifying relevant sentences.

Land use (Category 5): Accuracy decreased significantly to 33.71%, precision dropped to 15.99% but recall greatly increased to 87.76%. The F1 score more than doubled to 27.04%. The model is much better at identifying relevant sentences but has a lot of false positives.

Other (Category 6): Accuracy significantly dropped to 20.86%, precision is extremely low at 3.83¡replacement command¿¡replacement command¿¡replacement command¿% but recall is high at 91.67%. The F1 score decreased to 7.36%. This shows the model identifies almost all relevant sentences but also classifies many irrelevant sentences into this category.

In conclusion, option 3 seems to have drastically improved recall in most categories but at a significant cost to accuracy and precision. This suggests that breaking sentences down to individual nouns allows the model to capture more instances of each category but also leads to a lot more misclassifications.

Overall, if a high recall is of utmost importance, option 3 appears to be the best choice. However, if a balance between precision and recall (F1 score) is required, option 2 might be the best. If high accuracy and precision are required, then option 1 might be better. Each option has its trade-offs, and the best choice depends on the specific context and requirements.

5.2. Supervised Experiment and Conclusion

Taliya After training the model, predictions were made on the test data, which the model had not seen before. In order to evaluate the predictions, each sentence labeled by the model was reviewed to determine if it was correctly classified. In cases where the model made an incorrect classification, the label assigned by the model was replaced with the manually annotated label from our tagged ChatGPT data. This provided a stronger ground truth since the test data received manual and human evaluation compared to the ChatGPT, which had to classify sentences it had not necessarily been trained on, making it potentially less accurate and reliable.

Two different methods were employed for the categorization mission: TF-IDF and sentence BERT. For each method, the results were evaluated separately for each category using the four metrics (accuracy, precision, recall, F1 score).

5.3. TF-IDF

Taliya For the TF-IDF method, during the training of the model, we observed significant and high results in terms of precision, recall, and F1 score after 60 epochs (refer to figures). This indicates that the model learned well from the training data [Loss7],[Precision8], [Recall9] . However, the results on the test data were generally not high.

[Table 8] For the category "Environment and climate resilience," we observed an accuracy of 76.36%, precision of 35%, and recall of 38%. These results were poor primarily because this category was underrepresented in the data, and the model struggled to identify the words that associated a sentence with this category.

The remaining categories achieved an accuracy of over 80%. It can be seen that for "mobility (transportation)," we obtained an accuracy of 97.2%, but precision, recall, and F1 score were zero. This is due to the presence of only two such examples in the test data, which is very limited. The model failed to capture these examples and learn from this category.

For the "Future of work" category, the recall percentage was low at 33%. This can be attributed to several possible reasons, such as a lack of representative

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Environment and climate resilience	40.86	32.99	88.89	48.12
Mobility (transportation)	76.57	57.14	35.56	43.84
Local identity	62.86	21.88	27.63	24.42
Future of work	39.71	20.35	63.51	30.82
Land use	33.71	15.99	87.76	27.04
Other	20.86	3.83	91.67	7.36

Table 7: *Unsupervised Performance Metrics*
Presents the performance metrics of option 3 when we send to the model each noun from the sentence and calculate the avg for each category. The table includes accuracy, precision, recall, and F1 score, for our 6 categories. The data highlights variations in these metrics across the categories.

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Environment and climate resilience	76.36	35.71	38.46	37
Mobility (transportation)	97.2	0	0	0
Local identity	84.7	59	86.6	70.2
Future of work	91.6	50	33.3	40
Land use	80.5	72.2	59	65
Other	84.7	73.6	70	71.7

Table 8: *TF-IDF Performance Metrics*
presents the performance metrics when we use the TF-IDF. The table includes accuracy, precision, recall, and F1 score, for the 6 categories. The data highlights variations in these metrics across the categories.

examples in the training data or the complexity of correctly identifying sentences related to this category.

The ”Local identity” and ”land use” categories achieved relatively high accuracy rates, mainly because they had a larger number of examples in the data, allowing the model to better identify sentences belonging to these categories.

The category ”other” also achieved relatively good accuracy rates, which is positive as it indicates that the model is capable of identifying sentences that are not related to urban planning in collaboration.

5.4. BERT

Renana For the Sentence-BERT method, during the training of the model, we observed significant and high results in terms of precision, recall, and F1 score after 200 epochs (refer to figures). The number of epochs was set to 200, even though the accuracy stabilizes after around 50 epochs, as the recall and precision metrics continue to improve and reach their peak after approximately 200 epochs. Success is measured by all four metrics, and we cannot solely rely on the accuracy metric. [Loss10],[Precision11], [Recall12].

As can be seen in [Table 9], The results of the model for sentences embedded with the sentence-BERT method were impressive as expected. The accuracy was particularly high, with accuracy above 90%

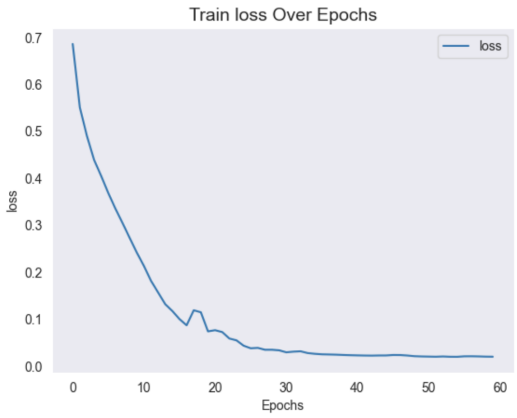

Figure 7: *Loss of TF-IDF training is a way of measuring mistakes, the lower it is, the fewer mistakes the model makes. Loss going down till 0 in the end of the training.*

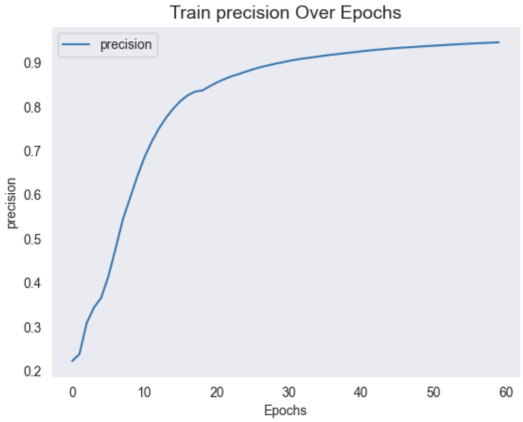

Figure 8: *The graph shows the relationship between epochs and precision for the TF-IDF model. Precision improves steadily with more epochs, reducing false positives. The model achieves 95% precision. Increasing epochs result in more accurate predictions.*

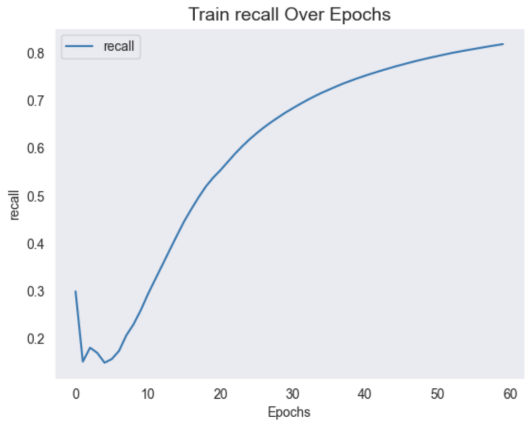

Figure 9: *The graph depicts the relationship between epochs and recall for the TF-IDF model. Recall steadily improves with more epochs, resulting in a higher rate of correctly identifying positive cases. At the end of the training process, the model achieves a recall rate of 85%, indicating its ability to correctly identify 85% of positive cases in the training data.*

for most categories. Additionally, the results of other metrics were also satisfactory. The F1 score, which includes recall and precision, exceeded 64% for all categories.

Upon further analysis of the results for each category, it was noticeable that the lowest accuracy was significantly lower for two categories: ”land use” and ”other.” The lower accuracy for these categories can be explained by the fact that they are similar to each other and very generic. In other words, the ”other” category represents sentences that do not belong to any other category, and it does not have a specific definition of its own. Furthermore, the ”land use” category encompasses various and diverse concepts, including surfaces, places, materials, and so on. The three standout categories in terms of accuracy are ”Environment and climate resilience,” ”Mobility (transportation),” and ”Future of work,” which have more specific definitions. The ”local identity” category is more defined than ”land use” and ”other,” but still not very specific, which is why its accuracy is not the highest.

In terms of recall, it can be observed that the ”Mobility (transportation)” category achieved a result of 100%. Since there is limited data available, especially a small number of examples from this category, the model managed to capture all the examples from the class. Impressively, the model learned to distinguish between examples belonging to this category and examples that do not, despite the lack of data.

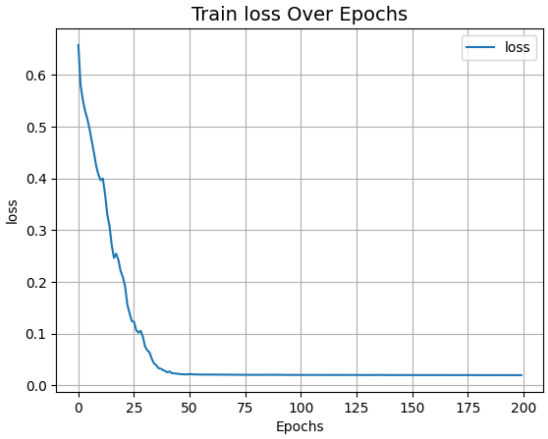

Figure 10: *BERT: Loss going down as the model trains, after about 50 epochs stabilizes.*

For the purpose of comparing the TF-IDF-based embedding model to the BERT-based embedding model, it is worth mentioning that overtly, the BERT-based model achieved better results in all metrics. The main explanation for this stems from the fact that the

Category	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Environment and climate resilience	93.06	85.71	60.00	70.59
Mobility (transportation)	98.61	66.67	100.00	80.00
Local identity	90.28	60.00	90.00	72.00
Future of work	94.44	66.67	85.71	75.00
Land use	86.11	60.00	69.23	64.29
Other	79.17	82.61	63.33	71.70

Table 9: *BERT Performance Metrics*
presents the performance metrics with Sentence-BERT embedding. The table includes accuracy, precision, recall, and F1 score, for the 6 categories. The data highlights variations in these metrics across the categories.

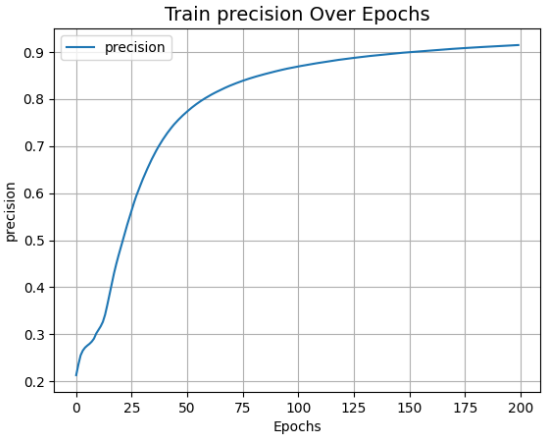

Figure 11: *BERT: Precision - As the model is trained over more epochs, its ability to make accurate and relevant predictions becomes more pronounced. Continues to grow up to about 200 epochs.*

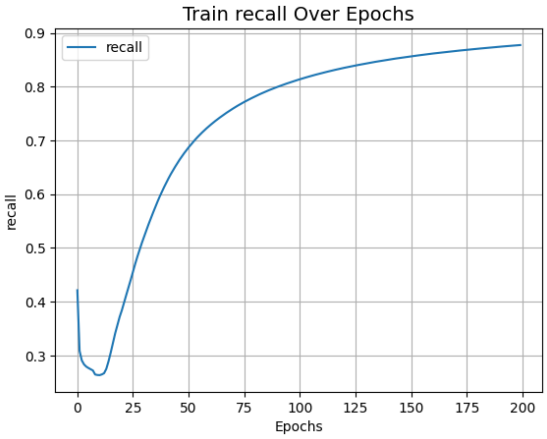

Figure 12: *BERT: Recall - As the model is trained over more epochs, its ability to make accurate and relevant predictions becomes more pronounced. Continues to grow up to about 200 epochs*

TF-IDF calculation method is based solely on tokens and does not consider the contextual information of the sentences, resulting in a lack of understanding sentence meaning. On the other hand, the BERT model embeds sentences with semantic understanding and a broader context perspective.

Furthermore, the BERT model was pretrained on a large dataset and fine-tuned on our smaller dataset, unlike TF-IDF, which was solely trained on our dataset.

BERT's understanding of contextual information and semantics helped it achieve better performance, especially in categories with fewer examples, demonstrating the strength of transformer-based models in handling small data sets.

6. Conclusion and Future Work

Eran The research aimed to explore and compare different supervised and unsupervised techniques for categorizing text data into predefined categories. The text data in this study were separated into pre-defined six categories. The study focused on two types of word embeddings – Term Frequency-Inverse Document Frequency (TF-IDF) and BERT bert-base-nli-mean-tokens. Applying these embeddings in the supervised models and in the unsupervised method using the smaller-LaBSE.

In the initial phase of the research, an unsupervised approach was implemented with a clustering algorithm. Here, the data was not labeled. Instead, the algorithm was used to cluster the data based on similarity in word embeddings. The primary aim was to see how well the natural structure of the data could be discovered by the algorithm. The performances of the algorithm with BERT embeddings were compared between 3 options. [5.2]

The results indicated that the unsupervised learning method had varying levels of success, largely de-

pendent on the chosen word embeddings and the specific category. In general, it was found that while the unsupervised technique could cluster the data reasonably well, it struggled to capture some of the nuances of certain categories, particularly when the categories were more closely related or overlapping.

Next, the research delved into a supervised learning approach. This time, the models had access to labeled data. They were trained to predict the predefined categories using the TF-IDF and bert-base-nli-mean-tokens embeddings. To evaluate the model performance metrics like accuracy, precision, recall, and F1-score were calculated for each category.

The TF-IDF model demonstrated strong precision for categories with distinct vocabulary sets, like 'Local identity' and 'Other'. Its high performance, particularly in precision, is an indication of how well it can handle explicit and distinct textual features when the dataset is manually tagged, but it also reflects its struggle with smaller categories due to its limited contextual understanding.

On the other hand, the BERT model showed strong resilience to small sample sizes, particularly evident in its ability to handle the 'Mobility (transport)' category. Its robust performance, even with smaller and imbalanced datasets, reveals the benefits of transformer-based models. BERT's strength lies in its understanding of contextual information, semantics, and the syntactic structure of language, allowing it to effectively capture subtle nuances across various categories.

However, both models exhibit distinct limitations. TF-IDF's lack of contextual understanding resulted in its failure to correctly predict underrepresented categories. BERT, despite its advanced architecture, showed some inconsistencies in precision and recall, suggesting it might be sensitive to imbalanced data.

Looking forward, we see a few potential directions for improving the performance of these models and the overall text classification process. First, addressing the data imbalance issue might improve the performance of the BERT model.

For BERT, we can explore fine-tuning techniques or use more recent transformer models, like RoBERTa or ELECTRA, which have shown improved performance on various natural language processing tasks.

Lastly, it would be valuable to collect more data, especially for underrepresented categories. Not only would this provide more examples for the models to learn from, but it would also enhance the representativeness and generalizability of the model's performance.

In conclusion, this research serves as a stepping stone for better understanding the challenges and potential solutions in text classification tasks, especially when dealing with smaller and imbalanced datasets.

7. References

[1] C. Calderon, "Unearthing the political: differences, conflicts and power in participatory urban design," *Journal of Urban Design*, vol. 25, no. 1, pp. 50–64, 2020. [Online]. Available: https://doi.org/10.1080/13574809.2019.1677146

[2] S. Münster, C. Georgi, K. Heijne, K. Klamert, J. Rainer Noennig, M. Pump, B. Stelzle, and H. van der Meer, "How to involve inhabitants in urban design planning by using digital tools? an overview on a state of the art, key challenges and promising approaches," *Procedia Computer Science*, vol. 112, pp. 2391–2405, 2017, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917314461

[3] M. Krüger, A. B. Duarte, A. Weibert, K. Aal, R. Talhouk, and O. Metatla, "What is participation? emerging challenges for participatory design in globalized conditions," *Interactions*, vol. 26, no. 3, p. 50–54, apr 2019. [Online]. Available: https://doi.org/10.1145/3319376

[4] D. C. Brabham, "Crowdsourcing the public participation process for planning projects," *Planning Theory*, vol. 8, no. 3, pp. 242–262, 2009. [Online]. Available: https://doi.org/10.1177/1473095209104824

[5] A. Wilson, M. Tewdwr-Jones, and R. Comber, "Urban planning, public participation and digital technology: App development as a method of generating citizen involvement in local planning processes," *Environment and Planning B: Urban Analytics and City Science*, vol. 46, no. 2, pp. 286–302, 2019.

[6] J. Dortheimer and T. Margalit, "Open-source architecture and questions of intellectual property, tacit knowledge, and liability," *The Journal of Architecture*, vol. 25, no. 3, pp. 276–294, 2020. [Online]. Available: https://doi.org/10.1080/13602365.2020.1758950

[7] S. Giering, *Public participation strategies for transit*. Transportation Research Board, 2011, vol. 89.

[8] S. L. Muehlhaus, C. Eghtebas, N. Seifert, G. Schubert, F. Petzold, and G. Klinker, "Game.up: Gamified urban planning participation enhancing exploration, motivation, and interactions," *International Journal of Human–Computer Interaction*, vol. 39, no. 2, pp. 331–347, 2023. [Online]. Available: https://doi.org/10.1080/10447318.2021.2012379

[9] N. Förster, I. Bratoev, J. Fellner, G. Schubert, and F. Petzold, "Collaborating with the crowd," *International Journal of Architectural Computing*, vol. 20, no. 1, pp. 76–95, 2022.

[10] F. Delgado, S. Yang, M. Madaio, and Q. Yang, "Stakeholder participation in ai: Beyond "add diverse stakeholders and stir"," 2021.

[11] A. Sears, J. Jacko, and J. Jacko, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Second Edition*, 2nd ed. CRC Press, 2007.

[12] A. Drutsa, V. Farafonova, V. Fedorova, O. Megorskaya, E. Zerminova, and O. Zhilinskaya, "Practice of

efficient data collection via crowdsourcing at large-scale," *CoRR*, vol. abs/1912.04444, 2019. [Online]. Available: http://arxiv.org/abs/1912.04444

[13] A. Gordon, "Crowdsourcing and its relationship to wisdom of the crowd and insight building: a bibliometric study," *Scientometrics*, vol. 126, no. 5, pp. 4373–4382, May 2021. [Online]. Available: https://doi.org/10.1007/s11192-021-03932-z

[14] H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.

[15] Z. Cao, X. Li, Y. Feng, S. Chen, C. Xia, and L. Zhao, "Contrastnet: Unsupervised feature learning by autoencoder and prototypical contrastive learning for hyperspectral imagery classification," *Neurocomputing*, vol. 460, pp. 71–83, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221010493

[16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," 2022.

[17] ——, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.

[18] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Overview of supervised learning," *The elements of statistical learning: Data mining, inference, and prediction*, pp. 9–41, 2009.

[19] B. Lavine and T. Blank, "3.18 - feed-forward neural networks," in *Comprehensive Chemometrics*, S. D. Brown, R. Tauler, and B. Walczak, Eds. Oxford: Elsevier, 2009, pp. 571–586. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444527011000260

[20] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence bert for answering online student faqs," *Results in Engineering*, vol. 17, p. 100856, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590123022005266

[21] I. Arroyo-Fernández, C.-F. Méndez-Cruz, G. Sierra, J.-M. Torres-Moreno, and G. Sidorov, "Unsupervised sentence representations as word information series: Revisiting tf–idf," *Computer Speech &Language*, vol. 56, pp. 107–129, 2019.

[22] S. Ullah, M. Liaqat, A. Asif, A. Khan, U. Aslam, and H. Asif, "Deep auto encoder based chatbot for discrete math course," in *2022 International Conference on Recent Advances in Electrical Engineering Computer Sciences (RAEE CS)*, 2022, pp. 1–7.

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] A. Mohammed and R. Kora, "An effective ensemble deep learning framework for text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Part A, pp. 8825–8837, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157821003013

[25] C. Park, S. Jeong, and J. Kim, "Admit: Improving ner in automotive domain with domain adversarial training and multi-task learning," *Expert Systems with Applications*, vol. 225, p. 120007, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423005092

[26] A. Jiao, "An intelligent chatbot system based on entity extraction using rasa nlu and neural network," *Journal of Physics*, vol. 1487, 2020.

[27] Y. Mohammadi, F. Ghasemian, J. Varshosaz, and M. Sattari, "Classifying referring/non-referring adr in biomedical text using deep learning," *Informatics in Medicine Unlocked*, vol. 39, p. 101246, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914823000886

[28] K. Yalcin, I. Cicekli, and G. Ercan, "An external plagiarism detection system based on part-of-speech (pos) tag n-grams and word embedding," *Expert Systems with Applications*, vol. 197, p. 116677, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422001610

[29] L. R. Lukas Tommy, Chandra Kirana, "The combination of natural language processing and entity extraction for academic chatbot," *CITSM*, 2020.

[30] C. Zhang, P. Mayr, W. Lu, and Y. Zhang, "Knowledge entity extraction and text mining in the era of big data," *Data and Information Management*, 2021.

[31] X. Kong, X. Liu, B. Jedari, M. Li, L. Wan, and F. Xia, "Mobile crowdsourcing in smart cities: Technologies, applications, and future challenges," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8095–8113, 2019.

[32] J. Mueller, H. Lu, A. Chirkin, B. Klein, and G. Schmitt, "Citizen design science: A strategy for crowd-creative urban design," *Cities*, vol. 72, pp. 181–188, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0264275117304365

[33] JingWei, SungdongKim, HyunhoonJung, and YoungHoKim, "Leveraging large language models to power chatbots for collecting user self-reported data," ", 2023.

[34] M. V. Dominika Krasňanská, Silvia Komara, "Keyword categorization using statistical methods," *TEM Journal*, vol. 10, p. 1377-1384, 2021.

[35] B. Das and S. Chakraborty, "An improved text sentiment classification model using tf-idf and next word negation," 2018.

[36] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[37] M. A. Rashid and H. Amirkhani, "Improving edit-based unsupervised sentence simplification using fine-tuned bert," *Pattern Recognition Letters*, vol. 166, pp. 112–118, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865523000168

[38] K. Kaur and P. Kaur, "Improving bert model for requirements classification by bidirectional lstm-cnn deep model," *Computers and Electrical Engineering*, vol. 108, p. 108699, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790623001234

[39] A. Conneau and D. Kiela, "Senteval: An evaluation toolkit for universal sentence representations," *arXiv preprint arXiv:1803.05449*, 2018.

[40] H. Li, W. Wang, Z. Liu, Y. Niu, H. Wang, S. Zhao, Y. Liao, W. Yang, and X. Liu, "A novel locality-sensitive hashing relational graph matching network for semantic textual similarity measurement," *Expert Systems with Applications*, vol. 207, p. 117832, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422010910

[41] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bing, "An unsupervised sentence embedding method by mutual information maximization," *ACL Anthology*, 2021.

[42] J. M. Leimeister, "Collective intelligence," *Business & Information Systems Engineering*, vol. 2, 2010.

[43] J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva, "The groningen meaning bank," *Handbook of linguistic annotation*, pp. 463–496, 2017.

[44] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," 2020.

[45] J. Read and F. Perez-Cruz, "Deep learning for multi-label classification," 2014.

[46] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018. [Online]. Available: http://arxiv.org/abs/1803.08375

[47] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," 2022.

[48] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance analysis and comparison of machine and deep learning algorithms for iot data classification," 2020.

[49] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[51] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2089–2103. [Online]. Available: https://aclanthology.org/2021.emnlp-main.160

[52] M. Yang, M. K. Lim, Y. Qu, X. Li, and D. Ni, "Deep neural networks with l1 and l2 regularization for high dimensional corporate credit risk prediction," *Expert Systems with Applications*, vol. 213, p. 118873, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422018917