ICS 2406: COMPUTER SYSTEMS PROJECT

TEXT ANALYSIS USING BAG OF WORDS,LDA AND K-MEANS TO VISUALISE SERVICES THAT NEED IMPROVEMENT IN VARIOUS COUNTIES.

TEAM MEMBERS;

ALBERT KIPTOO:SCT 211-0277/2019

JEREMIAH NGULI MATHEKA:SCT 211-0012/2019

OPOLO ANDREW OMONDI: SCT211-0022/2019

FIRST SUPERVISOR:DR. AGNES MINDILA

SECOND SUPERVISOR: DR. KENNEDY OGADA

DATE OF SUBMISSION:19/6/2019

CLIENT NAME: IAN ONAI KIPROTICH, RESIDENT OF KIAMBU COUNTY

CLIENT EMAIL: lank @istlafrica.com

COLLABORATION ROLES:

MEMBER	ROLES
OPOLO ANDREW OMONDI	● Clustering using k-means model to determine similar services ● Geospatial data visualization techniques. ● UI/UX design
ALBERT KIPTOO	● Topic modelling using Latent Dirichlet allocation model to determine topics present in input ● Token-based authentication ● Api development
MATHEKA JEREMIAH NGULI	● Word frequency determination using Bag of Words algorithm ● SVM classifier model ● Database Management

Table of Contents

1. Introduction

Abstract

Local communities are rarely involved in deciding their living environment’s future. Our website will act as a bridge between residents and decision-makers, empowering voices to address concerns and implement positive changes.

The platform at its core gathers insights from residents about their county's services and infrastructure. By simplifying the opinion-sharing process and providing helpful data analysis tools, we aim to empower residents to have a meaningful impact on local development. We believe that every resident deserves the chance to contribute meaningfully to their community's betterment.

The data collected and analyzed helps the local authorities prioritize their resources, make informed decisions, and take actionable steps towards improving essential community services. This valuable information strives to benefit the community's most critical needs through evidence-based solutions.

Additionally, our website encourages community engagement by facilitating conversations and cooperative efforts between locals. It serves as a central platform for sharing experiences, discussing ideas, and building relationships with individuals who share the same vision of creating positive change in their county. By valuing collective opinions, we believe that we can collaborate towards attainable progress while keeping residents' needs and aspirations at heart.

Problem statement:

Many residents in various counties have services they feel should be improved by their respective county governments.

Wao.L(p.2) stated "*Kilifi county hit headlines because of prolonged drought and fears of looming famine making them rely on support from the county and national governments and humanitarian relief organizations. Therefore improving access to water can improve the lives of people in Kilifi*".

Shahow A.A (February 24th 2023 p.14) stated "*Poor service delivery The mismanagement, graft and elite capture of county resources has resulted in poor service delivery for the people of north-eastern counties. A major challenge is that the leadership has been unable to prioritize development that would transform and improve service delivery. Despite the billions in investment—cumulatively, the three counties received close to Shs100 billion in devolved funds over the last ten years—there is nothing much to show for it. Also, the leadership has simply been unwilling to prioritize and invest in areas of public need where the impact would be greatest. Instead, funds have been spent as they come in poorly thought-out contractor-driven "development" projects. As a consequence, crucial sectors such as livestock and water, healthcare and education provision, where the needs of the population lie, have been ignored and, in some instances, the quality of services has deteriorated compared to the period before devolution.*``

However, there is a lack of efficient platforms for residents to express their service-related concerns and visualize their needs leading to a disparity between a region's need and the decision making process of county government authorities.

Therefore, there is a need for a user-friendly application that enables residents to input data about services they feel need improvements in their county and utilize natural language processing techniques to analyze their input. The application therefore should generate a comprehensive visualization such as a pie chart to show the distribution of the services they need improved.By addressing this problem the solution provides a platform that fosters better communication between residents and government authorities therefore facilitating targeted improvements in the identified services

Background

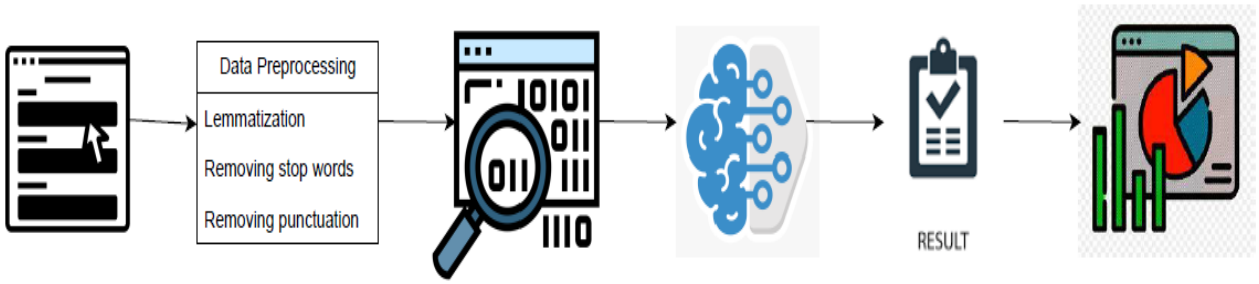
Due to the devolved government system, it is vital to enhance effective communication between residents of a county and government authorities regarding various issues that concern them. One area that this has fallen short is the identification and prioritization of the services that require improvement within specific counties. Many residents have pressing concerns about the quality of services in their community but finding a user-friendly platform to express these issues and visualize their distribution can be a huddle.

To bridge the gap, we propose the development of this application which enables a user to login, select their county of residence and input the services that they feel require improvements.Their input will then be analyzed using natural language processing techniques effectively. The application will then generate reports through visualization of the distribution of these services using a pie chart.

The visualization would be of help to various shareholders such as government authorities and non-governmental organizations who will be able to login and have an idea of what services that people see as a priority and act on their improvement and offer their helping hand respectively.

Proposed solution

We are proposing the development of a solution that will incorporate aspects of natural language processing such as tokenization, lemmatization, data preprocessing and visualization which we feel is a field that needs exploration because of its vast nature.
In addition, using a data-driven approach because text data is the main aspect of our solution that will be used to make decisions on visualisation after analysis.
The proposed solution will be a web application which will enable various categories of users to login. The categories of users include:county residents, government officials or organisations,non-governmental organisations and bloggers.
Users will be able to create accounts by providing basic details such as names, email addresses and passwords. The registration process will ensure that users have their privacy secured by enhancing authentication measures such as encryption.
i)County residents upon logging in, are provided with an interface where they are able to select their county of residence from a predefined list which we are going to fetch from an API that contains the list of the 47 counties.The application utilises natural language processing techniques to analyse the input from residents and extract relevant information.This includes data preprocessing tasks such as tokenization and lemmatization.Word frequency analysis and sentiment analysis is carried out on the preprocessed data.From the analysed data, a pie chart will be constructed to visualise the distribution of the services. Each category of the service improvement will be represented by the segments of the pie-chart.
ii)Bloggers upon logging in share their work on various topics such as transport and infrastructure, food security,transport e.t.c. Their work is then made available to the public who can then comment on the said topics. The blogs and comments are also analysed by the Latent Dirichlet algorithm to identify the topics.Sentiment analysis is also carried out to determine how people feel about various topics.
iii)Government organisations-upon logging in, are able to view the insights from residents of a particular county which will assist in the decision making process concerning distribution of services and resources.They will also have a platform to create awareness on the available services and how to access them.
iv)Non-governmental organisations-will access the visualisation from the services,identify the areas that feel need assistance and take relevant actions based on these visualizations.They will have a platform to showcase the services they provide, contacts on how to access them.o


Data Preprocessing
Lemmatization
Removing stop words
Removing punctuation
RESULT
User input
Data Cleaning
Feature Extraction
Classified by Model
Model outputs result
Compiled Results are graphed

Objectives

1. To develop a system that visualises services that need improvement in various counties.
2. To investigate the use of Bag of Words algorithm in service frequency determination.
3. To investigate the use of the Latent Dirichlet allocation algorithm identify the topics present in the input.
4. To incorporate an SVM model to facilitate service improvement category prediction based on topic modelling outcomes.
5. To apply the K-Means algorithm to group similar services that need improvement in a county.

2. State of art/Review of similar related works

Clustering Algorithms

1.Comparison between Data Clustering Algorithms by Osama Mahmoud

This paper is intended to study and compare different data clustering algorithms.
The algorithms under investigations:

- K-means algorithm
- Hierarchical clustering algorithms
- Self-organizing Maps algorithms
- Expectation Maximization algorithm.

The factors considered when evaluating these algorithms are:

- Size of the dataset
- Number of clusters
- Type of dataset
- Type of software used

The metrics of evaluation for the algorithms was:

- Performance
- Quality
- Accuracy

The general reasons for selecting the 4 algorithms was cited as the following:

- Popularity
- Flexibility
- Applicability
- Handling high dimensionality

However, specific reasons for choosing each algorithm was outlined as follows:

K-Means Algorithm

- ➢ Its time complexity is O(nkl), where n is the number of patterns, k is the number of clusters, and l is the number of iterations taken by the algorithm to converge.
- ➢ Its space complexity is O(k + n). It requires additional space to store the data matrix.
- ➢ It is order-independent; for a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the patterns are presented to the algorithm

Hierarchical Clustering Algorithm

- ➢ Embedded flexibility regarding a level of granularity.
- ➢ Ease of handling any forms of similarity or distance.
- ➢ Consequently applicability to any attributes types.
- ➢ Hierarchical clustering algorithms are more versatile

Self-Organizing Maps Algorithm

- ➢ While the Voronoi regions of the map units are convex, the combination of several map units allows the construction of non-convex clusters.
- ➢ Different kinds of distance measures and joining criteria can be utilized to form the big clusters.
- ➢ It has been successfully used for vector quantization and speech recognition.
- ➢ The SOM generates a sub-optimal partition if the initial weights are not chosen properly.

Expectation Maximization Algorithm

- ➢ It has a strong statistical basis.
- ➢ It is linear in database size.
- ➢ It is robust to noisy data.
- ➢ It can accept the desired number of clusters as input.
- ➢ It can handle high dimensionality.
- ➢ It converges fast given a good initialization.

Comparison of the 4 Algorithms

The four clustering algorithms are compared according to the following factors: the size of the dataset, number of the clusters, type of dataset, and type of software. For each factor, four tests are made, one for each algorithm. For example, according to the size of data, each of the four algorithms: K-means, Hierarchical Clustering, SOM, and EM is executed twice; first by trying a huge dataset and then by trying a small dataset. This is repeated for every other factor and conclusions made at the end.

The following table shows how the algorithms are compared:

	Size of dataset	Number of clusters	Type of Dataset	Type of software
K-means Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
HC Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
SOM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
EM Alg.	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package

This is how the algorithms perform under the different constraints. We measure the performance, quality and accuracy of the algorithm:

Number of Clusters and the Performance of the Algorithm				
	Performance			
N.o Of Clusters	SOM	K-Means	E.M	HCA
8	59	63	62	65
16	67	71	69	74
32	78	84	84	87
64	85	89	89	92

Number of Clusters and the Quality of the Algorithm				
	Quality			
N.o Of Clusters	SOM	K-Means	E.M	HCA
8	1001	112	1101	1090
16	920	1089	1076	960
32	830	910	898	850
64	750	840	820	760

Effect of Data Size				
	K=32			
Data Size	SOM	K-Means	E.M	HCA
36000	830	910	898	850
4000	89	95	93	91

Effect of Data Type				
	K=32			
Data Type	SOM	K-Means	E.M	HCA
Random	830	910	898	850
Ideal	798	810	808	829

Conclusion

After analyzing the results of testing the clustering algorithms and running them under different factors and situation, the following conclusions are obtained:

- As the number of clusters, k, becomes greater, the performance of SOM algorithm becomes lower.
- The performance of K-means and EM algorithms is better than hierarchical clustering algorithm.
- SOM algorithm shows more accuracy in classifying most the objects into their suitable clusters than other algorithms.
- As the value of k becomes greater, the accuracy of hierarchical clustering becomes better until it reaches the accuracy of SOM algorithm.
- K-means and EM algorithms have less quality (accuracy) than the others.
- All the algorithms have some ambiguity in some (noisy) data when clustered.
- The quality of EM and K-means algorithms become very good when using huge dataset.
- Hierarchical clustering and SOM algorithms show good results when using small dataset.
- As a general conclusion, partitioning algorithms (like K-means and EM) are recommended for huge dataset while hierarchical clustering algorithms are recommended for small dataset.
- Hierarchical clustering and SOM algorithms give better results compared to K-means and EM algorithms when using random dataset and the vice versa.
- K-means and EM algorithms are very sensitive for noise in dataset. This noise makes it difficult for the algorithm to cluster an object into its suitable cluster. This will affect the results of the algorithm.
- Hierarchical clustering algorithm is more sensitive for noisy dataset than SOM algorithm.

Verdict

We select K-Means clustering algorithm due to its high performance and accuracy with huge datasets since it is relevant to our project.

Our project entails very large data sets of randomized data which are the strengths of K-means algorithm

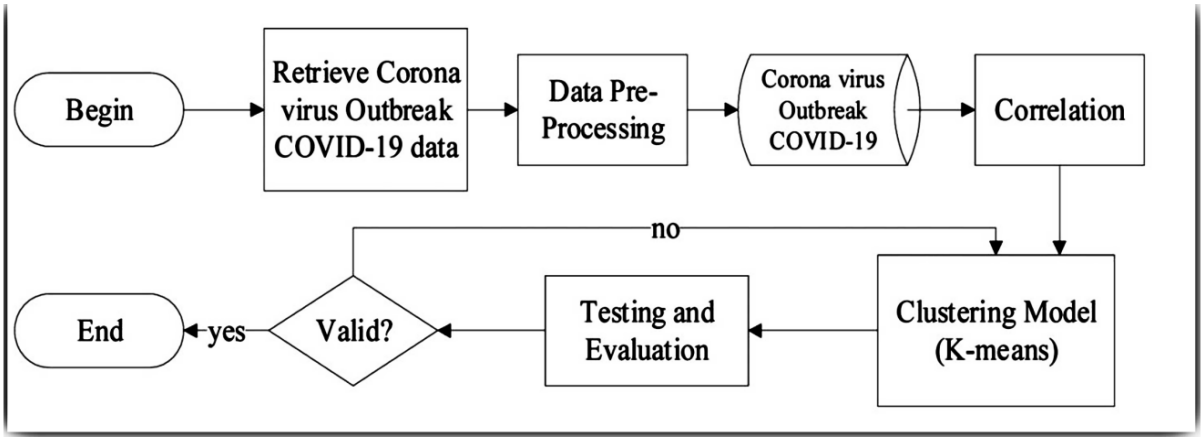
We now further explore areas where K-means algorithm has been used for clustering.

2. Muhtasim, Masud, M.A. (2023) Clustering Countries on COVID-19 Data among Different Waves Using K-Means Clustering.

In this work, a set of data from the COVID-19 coronavirus outbreak has been subjected to two well-known unsupervised learning techniques: K-means clustering and correlation. The COVID-19 virus infected several nations, and K-means automatically looks for undiscovered clusters of those infections. To examine the spread of COVID-19 before a vaccine becomes widely available, this work has used unsupervised approaches to identify the crucial county-level confirmed cases, death cases, recover cases, total_cases_per_million, and to-tal_deaths_per_million aspects of county-level variables. The researcher combined countries into significant clusters using this feature subspace to assist more in-depth disease analysis efforts. They used a clustering technique to examine various trends in COVID-19 incidence and mortality across nations.

The COVID-19 pandemic started on December 29, 2021, and as of December 29, 2022, there have been 619,391,055 confirmed cases, including 6,537,201. This data pertains to over 230 countries, regions, or territories that are affected by COVID-19 [2]. The illness pattern was not consistent between these locations, and understanding this heterogeneity is an essential source of knowledge for academics and policymakers. Unsupervised machine learning is used by Carrillo et al. to categorize 155 nations that have a similar COVID-19 profile. Clustering is done for cases that have COVID-19 confirmation. As feature variables, the following are used: disease prevalence, male population, air quality index, socioeconomic metrics, and health system indicators [3]. The clusters created to provide light on the similarities and contrasts between nations in terms of how COVID-19 has affected them.

Methodology



Conclusion

With K-means clustering, the researcher was able to quickly search for hidden or unidentified clusters across numerous COVID-19-infected nations, and we were also able to discover the correlations between various variables.

In order to study the factors directly associated with the spread of disease, 230 countries were clustered using an unsupervised K-Means algorithm based on socioeconomic, disease prevalence, and health system indicators. COVID-19 confirmed cases and COVID-19 death

cases were used as evaluation parameters. To determine the ideal number of clusters, the elbow approach was utilized. The incidence of COVID-19 verified cases is significantly positively correlated with the prevalence of asthma, diabetes mellitus, cardiovascular illness, dietary inadequacies, and health expenditure.
When using K-Means on COVID-19 confirmed cases and COVID-19 death cases, three clusters are produced. Cluster 0 consists of nations with high rates of recovery and low rates of mortality. These are the nations that have successfully contained the COVID-19 by strictly adhering to pandemic control procedures. Those nations in Cluster 1 have low mortality and low recovery rates. Some of these countries have an extremely large number of infected cases, but low mortality is a good indicator as a result, therefore these countries need to speed up their recovery rate to get out of it. Cluster 2 is a group of nations with a high mortality rate and a very high rate of recovery. In essence, a small number of these clusters' member nations have already experienced the worst of the epidemic, but they are currently recovering with a strong Recovery Rate.

Verdict
The research aligns with our project objectives. It clearly shows the benefits of using the K-means algorithm to cluster related data. This can be extrapolated to our project where we group similar services in each county. The grouped data can then be used by policymakers for analysis.

3.Spatial stratification and socio-spatial inequalities: the case of Seoul and Busan in South Korea
This study approaches the spatial stratification phenomenon through a data-based social stratification approach. In addition, by applying a dissimilarity-based clustering algorithm, this study analyzes how regions cluster as well as their disparities, thereby analyzing socio-spatial inequalities. Ultimately, through map visualization, this study seeks to visually identify spatial forms of social inequality and gain insight into the social structure for policy implications. The results determine how the regions are socioeconomically structured and identify the social inequalities between the spaces.

4. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June 1). *An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges*. IEEE Xplore. https://doi.org/10.1109/IEC47844.2019.8950616
The Bag-of-Words (BoW) model is a fundamental technique used in natural language processing (NLP) for text representation. It is a simple and effective way to convert text documents into numerical vectors that can be understood by machine learning algorithms. Here's an in-depth review of the Bag-of-Words model:

1. Overview: The BoW model represents a text document as a collection of unique words, disregarding the order and structure of the sentences. It treats each document as a "bag" of words, hence the name. The model focuses solely on the presence or absence of words in a document and their frequency.
2. Vocabulary Construction: To create the BoW representation, the first step is to construct a vocabulary or a dictionary. The vocabulary consists of all the unique words found in the corpus of documents being analyzed. Each word is assigned a unique index or identifier.
3. Text Preprocessing: Before constructing the vocabulary, text preprocessing steps are typically applied. This involves converting text to lowercase, removing punctuation, stop words (common words like "and," "the," "is," etc.), and performing stemming or lemmatization to reduce words to their base forms.
4. Document-Term Matrix: The BoW model represents each document as a vector known as a Document-Term Matrix (DTM). The DTM has one row per document and one column per word in the vocabulary. Each element in the matrix represents the frequency or occurrence of a word in a document. Alternatively, it can represent binary values (0/1) indicating the presence or absence of a word.
5. Vectorization: Once the DTM is constructed, it serves as the numerical representation of the text data. Machine learning algorithms can then operate on these vectors. The vectorization process involves converting the text data into numerical feature vectors. The DTM is often transformed into a sparse matrix to efficiently represent the high-dimensional space.
6. Feature Selection: Depending on the application, additional feature selection techniques may be applied to reduce the dimensionality of the DTM. This can include filtering out infrequent words, removing highly frequent words (stop words), or using more advanced techniques like term frequency-inverse document frequency (TF-IDF) to assign weights to words based on their importance.
7. Classification or Clustering: Once the text data is represented as vectors, various machine learning algorithms can be applied for classification or clustering tasks. Common algorithms used with BoW include Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and more.

Advantages of the Bag-of-Words Model:

- Simplicity: The BoW model is easy to understand and implement, making it a good starting point for text analysis tasks.
- Versatility: It can be applied to a wide range of NLP tasks, such as sentiment analysis, document classification, information retrieval, and more.
- Efficiency: The BoW representation can handle large amounts of text data efficiently, particularly when using sparse matrix representations.
- Language Independence: The model does not rely on language-specific rules or grammar, making it language-independent.

Limitations of the Bag-of-Words Model:

- Loss of Word Order: The BoW model does not consider the order or sequence of words in the text, leading to the loss of important contextual information.
- Lack of Semantic Understanding: It treats each word as an independent feature, without capturing the semantic relationships between words.
- Vocabulary Size: The size of the vocabulary can become very large, especially for large corpora, which can result in high-dimensional representations and computational challenges.
- Rare Words: Rare or unique words that are not present in the training data may not be adequately represented in the BoW model.

However, the BoW model suits our approach for the following reasons:

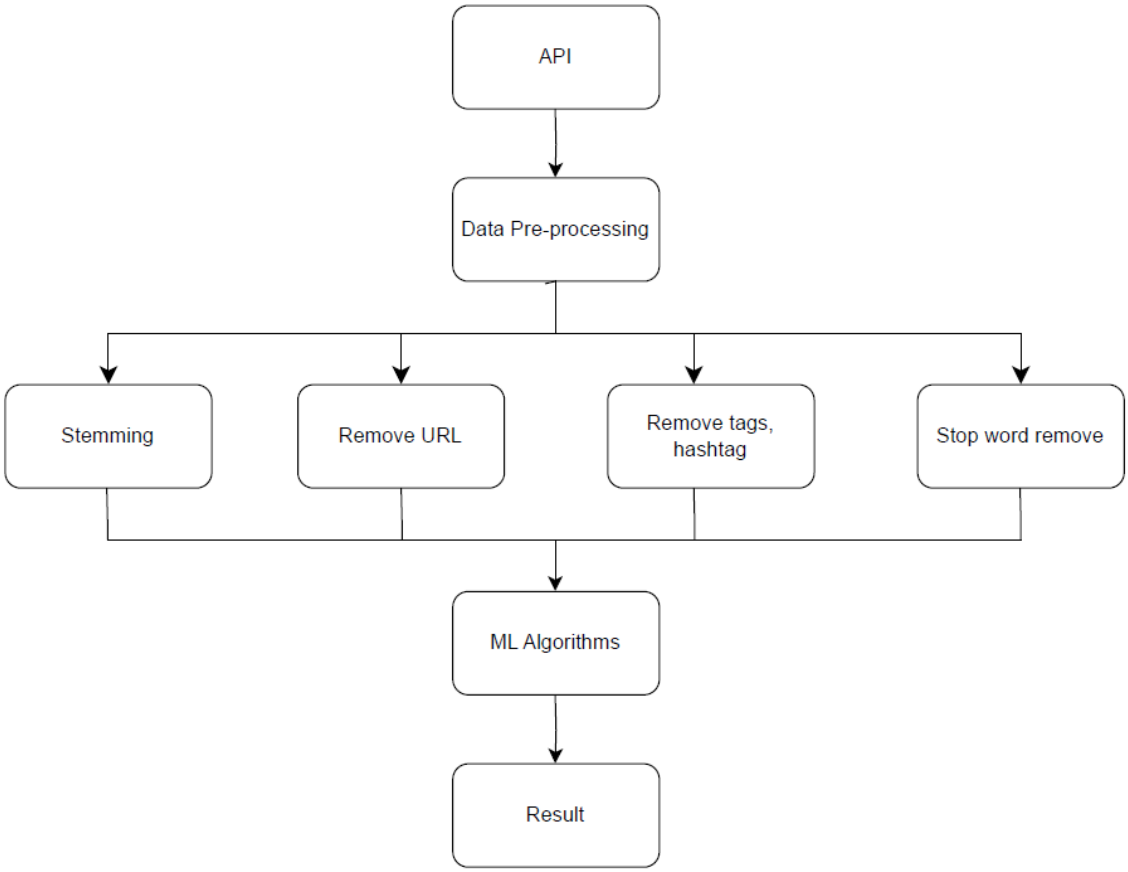
i) We only require the service a user wants improved, which can be broken down into an independent feature with no relationship with the other words.

ii) The word order is of no consequence as we are dealing with an individual independent feature.

iii) Availability of Labeled Data: The BoW model works well when there is a sufficient amount of labeled training data available for sentiment analysis. By representing text documents as numerical vectors, the BoW model allows for the application of traditional machine learning algorithms, which typically require labeled data for training.

5. Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019, November 1). *Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset*. IEEE Xplore.
https://doi.org/10.1109/SMART46866.2019.9117512

Sentiment analysis and opinion mining in social networks are concentrated on classification of texts collected from social networks into binary or ternary classification. Authors in (Rahat et al., 2019) collected a dataset using Twitter API. They then preprocessed the reviewed text and trained a model. They used this opportunity to pit two of the most common algorithms used in Opinion mining against each other i.e. Naive-Bayes theorem and Support Vector Machine.



1). Naïve Bayes Approach

Naive Bayes is a collection of classification algorithms which are based on Bayes Theorem. Naive Bayes classifier gives us an excellent result when one uses it for text data analysis. Such as Natural Language Processing. Naive Bayes algorithm gives us a probability analyzing the data set we have given. Naïve Bayes classifier is used as a probabilistic classifier. To perform the classifier, it uses the concepts of mixture models. A mixture model is capable of establishing the probability of the component that it consists of Bayes theorem to perform as a probabilistic classifier (Rahat et al., 2019)

2). Support Vector Machine

Support Vector Machine is a universal learner. Support Vector Machine has defined both input and output format. The output is either positive or negative and input is vector space. The text document is not suitable for learning. Those texts are transformed into a structured format. The text is transformed into a format which matches the input of the machine learning algorithm. The score of the texts are calculated and then the score is given as input to Support Vector Machine. Support Vector Machine has been proved one of the most powerful learning algorithms for text Categorization. But text categorization sometimes may produce errors. To decide which one is better between texts a comparison of text classifiers is required. The performance measure is used in this case.(Rahat et al., 2019)

Experiment and Result

Before the experiment result, set the train sample size 67% of the whole dataset and set the random state to 40. Then used the SVM and Naive Bayes algorithm for result prediction. For SVM we used SVC linear kernel and for Naïve Bayes used multinomial Naive Bayes. In our analysis Support Vector Machine gives a more accurate result than Naïve Bayes algorithm. After the train, we test both algorithm prediction by the review twitter and define the best performer(Rahat et al., 2019)

	SVM	**Naive Bayes**
Accuracy	82.498	76.56
Precision	90.33	89.00
Recall	81.79	83.75
F1 Measure	85.85	86.37

Based on the results, SVM proves to be more accurate and precise than Naive Bayes theorem. We therefore choose to use the SVM algorithm to create our model that will classify the data we obtain from users' input.

6. Rameshbhai C.J & Paulose.J (2019) Opinion mining on newspaper headlines using SVM and NLP

https://pdfs.semanticscholar.org/5394/85ab1dee53752e2541f2906535ba541241f2.pdf

In this paper, the authors focus on performing Opinion Mining based solely on news headlines using an SVM model, without examining the entire articles. This approach allows for a quicker and more efficient analysis of public sentiments and opinions. By analyzing headlines alone, valuable insights can be gained, shedding light on the prevailing sentiments towards various topics, products, or entities without the need to process the entire articles.

Approach

Figure 2 depicts the basic flowchart of the data preprocessing and model building. Stop words are removed from news headlines followed by converting uppercase texts to lowercase. The semi-processed headlines are fed to coreNLP. The output of coreNLP with sentiment scores are set as input for process II. The input is received from process I and converted into unigram and bi-gram representation. Model A is generated from the unigram and bi-gram representation. Model A employs Linear SVM. Data representation is further converted into Tf-idf resulting in Model B and C. Model B and C employ Linear SVM. However, unlike model B, model C uses SGD classifer to train the data.

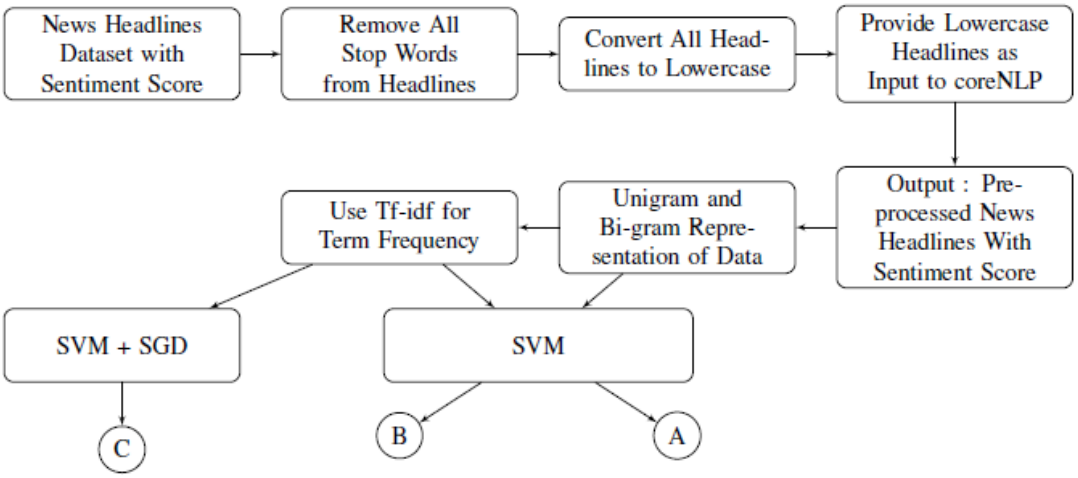


Figure 2. Flow diagram of data pre-processing and model building

In this section, three different models are described for performing Opinion Mining or Sentiment Analysis on newspaper headlines. These models utilize different approaches to build the classifiers and process the textual data. Here is a description of all 3 models:

Model A. Linear SVM:

- The data used to build the model is numeric, and it is represented in matrices.
- Two types of matrices are generated using unigram and bi-gram approaches.
- Unigram model has 1472 samples (news headlines) and 4497 features (total unique words in the dataset).
- Bi-gram model has 1472 samples and 13832 features.
- 80% of the data is used for training the model, and 20% is used for evaluating the model.
- A linear kernel is used because there are two class labels, and SVM generates a linear hyperplane to separate words into positive and negative news headlines.

Model B. Tf-idf and Linear SVM:

- Linear SVM is used again for building the model.
- The dataset is transformed into document frequency using Tf-idf (Term-frequency inverse document frequency).
- Tf-idf is used to calculate the importance of a word in a headline relative to its overall occurrence in the dataset.
- The resulting Tf-idf vectors are normalized using the Euclidean norm.
- The total number of sample data size and feature data size remain the same as in Model A.

- The frequency of each word is changed according to Tf-idf, giving higher importance to words that are frequent in a headline but less frequent overall.
- 80% of the data is used for training, and 20% is used for testing.

Model C. Stochastic Gradient Descent (SGD) Classifier:

- The SGD Classifier is used to train the data for Linear SVM.
- SGD is a discriminative learning algorithm used for linear classifiers like SVM and Logistic Regression, particularly effective in NLP and text categorization problems.
- The data provided is sparse, and the classifiers in SGD are efficient in handling large datasets with many training samples and attributes.
- In this research problem, a small dataset is used, but the approach can be extended to handle datasets with up to 105 features.

Results

	Model A (Linear SVM)	Model B (Tf-idf + Linear SVM)	Model C (SGD)
Unigram	87.11 %	90.84 %	83.72 %
Bi-gram	89.49 %	91.52 %	88.13 %

This table implies that the bi-gram will give more accurate result than unigram. However, in unigrammodel, number of feature is less than bi-gram model, due to which time in building the model in unigram is less than bi-gram. Here, the accuracy of Model B is higher than Model A because it is trained with Tf-idf. With the increase in feature size (>20000), Model A and B will not provide feasible solutions. To overcome such issues Model C is introduced in this paper and it is trained using SGD, it supports up to 105 features (3) for building a model. Thus Model C can be used when the feature size is high, otherwise Model B works well when the feature size is less.

Relevance to our project

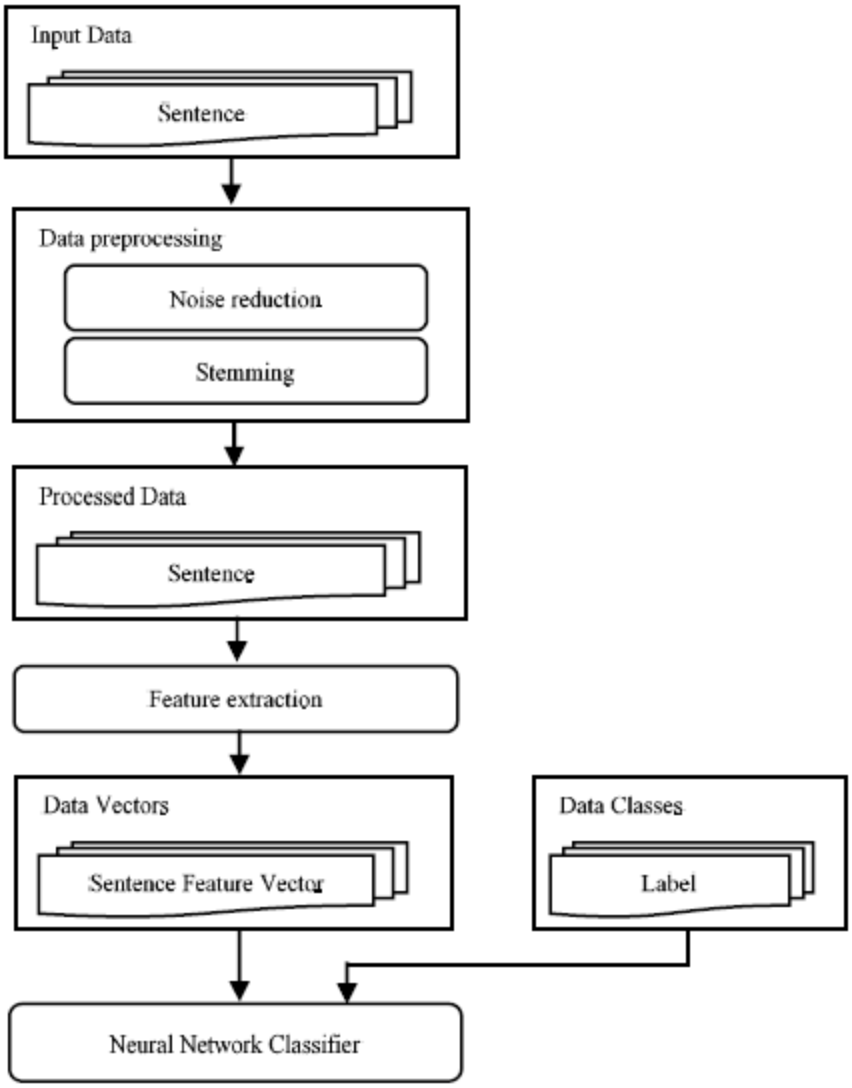
- The paper further supports our choice of SVM as the algorithm to be used in training our classification model, as the results are very favourable with the least accurate model being 83.72% accurate which is impressive.
- As we intend to pair SVM with Bag of Words, an alternative method to TF-IDF of representing text as numerical vectors, we are confident of producing a model with similarly relatively high accuracy.
- Based on the results of comparing Unigram and Bi-gram models, we chose to use a Unigram model as it is simpler and the slight difference in accuracy does not justify the increased complexity of a Bi-gram model.

7. Yan, Y., & Zheng, K. (2020, December 1). Text Classification Model Based on Multi-level Topic Feature Extraction. IEEE Xplore. https://doi.org/10.1109/ICCC51575.2020.9344894

Text classification is an important part of natural language processing applications like email filtering, sentiment analysis, and search engines. It involves assigning text documents to predefined categories based on their content. Before applying machine learning models, information needs to be extracted from the text message, which is called text feature extraction. This step removes irrelevant features and can improve the accuracy of the learning model.

This paper explores different approaches to text feature extraction and evaluates the performance of a neural network classifier. The neural network is commonly used in natural language processing tasks. Feature extraction is crucial in text classification as it directly affects classification accuracy. It is based on the vector space model, where each dimension represents a feature of the text. One approach is the Term Frequency Inverse Document Frequency (TF-IDF) weighing scheme, but it can generate high-dimensional feature vectors in large text corpora, increasing the risk of overfitting. Dimensionality reduction techniques can address this issue by representing the document concepts in reduced dimensions

TF-IDF has a large feature set equal to the vocabulary size of the corpus, leading to high computational requirements. LSA generates concepts based on word-document relationships using a TF-IDF weights matrix and Singular Value Decomposition. LDA finds a linear combination of features for characterizing or separating classes, involving steps such as calculating mean vectors, scatter matrices, and eigenvectors. Finally, the transformation equation $Y = X \times W$ is used to transform samples onto a new subspace in LDA.


Input Data
Sentence
Data preprocessing
Noise reduction
Stemming
Processed Data
Sentence
Feature extraction
Data Vectors
Sentence Feature Vector
Data Classes
Label
Neural Network Classifier

Literature Review on Topic Modelling using Latent Dirichlet allocation algorithm

Introduction

The Latent Dirichlet Allocation Algorithm has been used often in topic modelling whereby useful themes and or topics are extracted from a corpus of words or documents. Research has been done to investigate the its' application to the field of Topic modelling. In this section we are going to discuss the various steps the researchers took to derive topics from the corpus, emerging trends in topic modelling, gaps and opportunities for future improvements.

Related works on Latent Dirichlet Allocation algorithm for topic modelling

Tile	_Objectives	Methodology	Main Takeaways
(1) Hu, R., Wencong, M., Lin, W., Chen, X., Zu-Chang, Z., & Chu-Hong, Z. (2022). Technology Topic Identification and Trend Prediction of new energy Vehicle using LDA modelling. *Complexity*, *2022*, 1–20. https://doi.org/10.1155/2022/9373911	To identify topics and trends in New Energy vehicle using LDA algorithm	**1.Patent database of CNKI was used to search the data related to new energy vehicles** **2. New energy vehicle technology topic identification using and evolution analysis based on LDA topic model.** **3.Text preprocessing** **4.Calculation of TF-IDF Assignment Weights and Optimal number of Topics**	-Topic distribution of a document can be obtained by sampling from the Dirichlet distribution 1) The selection of a number of topics is directly related to the training effect of the LDA topic model. 2) Three common methods of setting number of topics- -Using the size of confusion to evaluate the goodness of a model-smaller perplexity indicating better training result of training set. - Using Hierarchical Dirichlet Process method which assumes

			that the documents share the same topics before training but exact number of topics identified by derivation of Dirichlet parameters -Application of Bayesian model approach to determine optimal number of topics Using TF-IDF to for feature extraction using the count vectorizer function

Title	**Objective**	Methodology	Main Takeaway
(2) García-Méndez, S., De Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with Latent Dirichlet Allocation. *Applied Intelligence*.	**<u>To investigate relevant topics in financial news using LDA algorithm</u>**	1) **Multi-paragraph topic segmentation and co-reference resolution to separate author expression patterns.** 2) **Detection of relevant text through topic modelling with Latent Dirichlet Allocation (LDA), outperforming a rule-based system.** **3)Identification of forecasts and predictions within relevant text using discursive temporality analysis and Machine Learning**	1)Enhanced Textiling algorithm is used to segment topics into subtopics The algorithm uses lexical co-occurrence and discourse distribution patterns to identify 2) Coreference resolution step is important as references are converted to meaningful words for better efficiency when using the LDA algorithm. Consideration of temporal, textual and numeric features Using Freeling library to detect dependencies between terms

Https://doi.org/10.1007/s10489-023-04452-4			

Title	**Objective**	**Methodology**	MainTakeaways
(3) Sutherland, I., & Kiatkawsin, K. (2020). Determinants of Guest Experience in Airbnb: A Topic Modeling Approach using LDA. *Sustainability*, *12*(8), 3402. https://doi.org/10.3390/su12083402	**To identify guest experience in Airbnb using LDA topic modelling.**	**1. Research context. Research was carried out in the various cities to compare the behavior of tourists.** **2. Data was collected inside via (insideairbnb.com)** **3. Preprocessing analysis, modelling and visualisation using the R-language** **4. The number of latent topics is identified via the maximisation of the information divergence of all topic pairs** **5. Evaluation of the topics**	-Phi-values which is the conditional probability of words given a specific topic in the topic model -The top n words in each topic ranked by phi-values are used in order for the human-interpretation of the words to understand what the underlying concept that the words of highest phi-values in a particular topic are depicting -Ward hierarchical clustering is used to show complex relationships that exist between the topics

Title	**Objective**	**Methodology**	MainTakeaway
(4).Brzustewicz, P., & Singh, A. (2021). Sustainable Consumption in Consumer Behavior in the time of COVID-19: Topic Modeling on Twitter Data using LDA. *Energies*, *14*(18), 5787. https://doi.org/10.3390/en14185787	**To identify sustainable consumption in consumer behaviour during COVID 19 using LDA algorithm**	**1. Data collection and preprocessing. Twitter posts were extracted using the Twitter API. Streaming data and analysis using the R programming language.** **2Topic mining using approaches such as topic modelling, semantic network analysis and sentiment analysis.** **3 Evaluation and modelling. The model evaluated using coherence and perplexity measures of evaluation.** **4 Visualization of the topic distribution.**	-Perplexity measures how well a model describes a document according to a generative process based on the learned set of topics. A lower perplexity implies a greater fit -Topic coherence captures the optimal number of topics based on the degree of semantic similarity between high scoring words within the topic; thereby giving the human interpretable topics -Quality of a topic model depends not only on its performance in statistical metrics but also on its reasonability and the interpretability of each topic

			-The Louvain algorithm was implemented to identify semantic clusters of latent topics during the semantic analysis

Title	Objective	Methodology	MainTakeaway
(5) Ebrahimi, F., Dehghani, M. H., & Makkizadeh, F. (2023). Analysis of Persian Bioinformatics Research with Topic Modeling. *BioMed Research International*, *2023*, 1–8. https://doi.org/10.1155/2023/3728131	**To identify topics and words present in bioinformatics using the LDA algorithm**	**1)Data preprocessing by by removal of stop words and punctuation and empty records; Performing lemmatization to reduce words to their base form and tokenization to break texts into smaller units** **2 Text conversion to numeric form that is suitable for analysis.** **3)Topic modelling using the LDA algorithm**	-The TF-IDF algorithm is useful in conversion of text into numeric form which is an important step in feature extraction -The clustering of texts requires the extraction and preservation of semantic information -A topic model enables one to obtain data from multiple clusters rather than just one as opposed to conventional clustering.
Title 6)Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021, November 23). LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. *Applied Sciences*, *11*(23), 11091.	**Objective** **To identify influence of text preprocessing on model interpretability**	**Methodology** **1)Data collection from Lib.ru and corpus of russian short stories** **2)During data preprocessing punctuation,digits and function words were removed.Lemmatization**	Main Takeaways -Corpus of text is considered as mixture of topics -Each document in the collection is described by a set of hidden semantic structure -topics which in turn are

https://doi.org/10.3390/app112311091		**was done using ru_core_news_sm model of the spaCy library** **3)Analysis using LDA algorithm to obtain topics present in the corpora.** **4)Topic model evaluation to determine the interpretability of given topics.**	composed of words that contribute to it with a certain degree of probability. -Well interpretable data is obtained regardless of length of text because it does not require any preliminary training of data During topic evaluation factors that are put into consideration-how interpretable the resulting topics are and how well the topics describe the document in question. Perplexity and coherence are the statistical measures that gauge the quality of topic models. Perplexity measures how evenly distributed words are in text. Coherence measures how often words that occur together in one

			text fall in the same topic. Coherence is mostly used as it correlates with expert assessments. The higher the coherence the better.

Title **(7)**Wu, Z., Xie, P., Zhang, J., Zhan, B., & He, Q. (2022, May 25). Tracing the Trends of General Construction and Demolition Waste Research Using LDA Modeling Combined With Topic Intensity. *Frontiers in Public Health*, *10*. https://doi.org/10.3389/fpubh.2022.899705	**Objective** **To identify how to extract existing topics about construction and demolition waste from text using LDA algorithm** **To identify popularity of topics using topic intensity**	**Methodology** **1.Data collection from WoS database** **2 Data preprocessing by stopwords removal using the NLTK Toolkit.** **3.LDA-based topic modelling implementation.The choice of number of topics set in advance.**	Main Takeaway The probability of co-occurrence of two words in a sentence is the coherence score. A higher coherence score implies a better result. The number of topics that are set that gives the highest coherence score is the right measure. Topic intensity which is a measure of how often the topic in question is being discussed.Upward trend of topics can provide more valuable information for real-time analysis.

Title	**Objectives**	**Methodology**	Main Takeaways
(8)Hidayatullah, A. F., Aditya, S. K., Karimah, & Gardini, S. T. (2019, March 11). Topic modelling of weather and climate condition on twitter using latent dirichlet allocation (LDA). *IOP Conference Series: Materials Science and Engineering*, *482*, 012033. https://doi.org/10.1088/1757-899x/482/1/012033	**To investigate the use of LDA in determining topics on weather information and forecasting**	**1 Data collection from original twitter account of BMKG** **2 Data preprocessing by tokenization and using bigrams library in NLTK to join words that appear together into one.** **3.Topic modelling using the LDA algorithm.** **4.Data visualization.**	Key steps taking in topic modeling -Building term dictionary and corpus -Building the LDA object -Visualization

Current trends in LDA topic modelling

Multilingual Topic Modelling

Farkhod et al.,(2021). Multilingual Topic modelling for tracking COVID-19 trends based on facebook data analysis which covered 7 languages; English, Arabic, Spanish, Italian, German,French and Japanese.

Dynamic Topic models

Glyn et al.,(2019). Bayesian analysis of dynamic linear topic models

Discovering temporal evolution of themes from time-stamped collection of texts poses a challenging statistical learning process.Dynamic topic modelling offers a probabilistic modelling framework to decompose a corpus of text documents into topics while simultaneously learning the temporal dynamics of relative prevalence of these topics

Applying deep learning techniques in LDA

Bhat et al (2019). A new way to topic model

Probabilistic topic models like Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) and Biterm Topic Model (BTM) have been successfully implemented and used in many areas like movie reviews, recommender systems and text summarization etc. These models however become computationally heavy if tested on humongous corpus. Keeping in view the vide acceptability of Deep Neural network based machine learning, this research proposes two deep neural network variants (2NN DeepLDA and 3NN DeepLDA) of existing topic modelling technique Latent Dirichlet Allocation (LDA) with specific aim to handle large corpuses with less computational efforts. Two proposed models (2NN DeepLDA and 3NN DeepLDA) are used to mimic the statistical process of Latent Dirichlet Allocation. Reuters-21578 dataset has been used in the study. Results computed from LDA are compared with the proposed models (2NNDeepLDA and 3NNDeepLDA) using Support Vector Machine (SVM) classifier. Proposed models have shown significant accuracy besides computational effectiveness in comparison to traditional LDA.

Gaps and limitations of the literature review

Metrics of evaluation of topics

Metrics of evaluation of interpretability of topics such as coherence score and perplexity measure are not adequate to determine the accuracy of the models.
We are proposing the use of other metrics of evaluation such as semantic coherence which measures semantic similarity between topics,normal mutual information which measures degree of similarity between topics in the model and ground truths. Using these metrics in addition to coherence and perplexity will improve the process of evaluation.

Visualisation

This is a very vital process since it gives insights on the analysed data.The literature we studied above had not emphasized on the visualization aspect. Most provide a single type of chart after analysis. In our project we are going to use a combination of bar charts,pie charts and scatter charts whereby the user is going to select the type of visual chart according to their preference

Methodology

1) **Data collection**

2) **Data preprocessing**

1) Importing following python libraries commonly used for topic modelling
   Gensim: This library provides implementations of various topic modelling algorithms, including LDA.

NumPy: This library provides support for mathematical operations on arrays.

Pandas: This library provides support for data manipulation and analysis.

Matplotlib: This library provides support for plotting data.

2)Loading the data.The data to be used will be in a csv file after performing data preprocessing

3)Choosing the number of topics to be identified

4) Training the models using genism.models.ldamodels.LdaModels() class.

The `LdaModel()` class takes the following parameters:

- `corpus`: The corpus of documents to be used for training the model.
- `num_topics`: The number of topics to be extracted from the corpus.
- `alpha`: A hyperparameter that controls the distribution of documents over topics.
- `beta`: A hyperparameter that controls the distribution of words over topics.

Evaluating the model using coherence score. A high coherence score indicating that the model has performed well.

3. Approach/Methodology

3.1 Description

We are planning to use an iterative and incremental approach.
We will first conduct a comprehensive research on the stakeholders of the system to gain insights about project requirements by conducting interviews and using questionnaires to refine our understanding about the user requirements.
Thereafter,build the application in increments, first starting with the user interface to allow feedback by the users.The user interface will incorporate the various stakeholders such as citizens, government institutions, non-governmental organizations and bloggers We will then build the machine learning models that will be used to analyze data from the user input.
We are going to implement the backend logic of the application concurrently with the frontend .

3.2 Technology

We are planning on using Reactjs which is a JavaScript framework and chakraUI for the frontend of the application and Django python framework for backend implementation. The model for analysis will be built using the Natural Language Processing Toolkit (NLTK) Python library through a Jupyter notebook. Github will be used for version control and ClickUp for project management.
Bag of words algorithm will be used where there is a dictionary that represents occurrences of certain words. This algorithm will be used for feature extraction. We have decided to use the algorithm because it can extract frequency based on text input.
We intend to use the Python library Natural Language Toolkit (NLTK) before constructing the BoW model, to preprocess the text by tokenizing it into individual words, removing stopwords, and performing any additional text cleaning or normalization steps and tokenizing the users' sentiments.
For classification of text based on word frequencies, we will use the Support Vector Machine algorithm after extracting the features using the bag of words algorithm.

3.3 Data

Our project will use both primary and secondary sources of data with the primary data being the sentiments from the users and the secondary data being data from the Humanitarian Data Exchange (humdata).
We are using textual data from the text input field from the forms. Residents will be able to type what services they feel need improvement and submit their responses.Additional data about the 47 counties will be retrieved from an API.
The following represents how data will flow:
i)Data collection-Forms will be used majorly for the collection of data. We will use ReactJS forms.

ii)Data preprocessing- Involves removal of stopwords, lemmatization and converting words into lowercase to ease the analysis process
iii)Feature extraction- The feature to be extracted is word frequency where we will use the bag of words algorithm to determine the frequency of occurrence of specific services which we will define in a dictionary of words.

Datasets

OCHA SERVICES DATA: This dataset is a detailed dataset containing data regarding the many different services and amenities available in Kenya, and the date when they were recorded. We will use this data in training our model to identify the specific service or amenity a resident wants improved. Below is the link to the dataset:
https://data.humdata.org/dataset/world-bank-infrastructure-indicators-for-kenya

PPRA Awarded contracts - Kenya - This dataset shows the tenders awarded to various organizations and the services they offered in 2019. We will use this data to train our model in identifying the services a user feels should be improved
https://staging.openafrica.net/dataset/250600e3-8f6d-47d4-832f-fb18ccb0f0ab/resource/6ea2833c-7762-46fa-bcd3-355c7995e3ae/download/ke_contracts.csv

3.4 Evaluation

1. **User Interface (UI) Testing:** UI testing evaluates the website's user interface to ensure it is visually appealing, intuitive, and responsive. It focuses on verifying the proper display of elements, proper navigation, and consistent design across different devices and browsers. Additionally, UI testing can also cover accessibility testing to ensure the website is usable for people with disabilities.
2. **Unit Testing**: This type of testing focuses on verifying the individual components or units of the website. It involves testing the functionalities of isolated modules, such as the data collection forms, database operations, and algorithms used for sentiment analysis or opinion classification.

3. **Integration Testing**: Integration testing is crucial for a website that relies on various components working together seamlessly. It ensures that different modules or subsystems of the website integrate correctly, exchange data accurately, and function as intended. It validates the communication between the front-end and back-end systems, database connectivity, and API interactions.
4. **Acceptance Testing**: This assesses the overall user experience of the website. It involves gathering feedback from real users to identify any usability issues, difficulties in navigating the site, or areas where improvements can be made. This type of testing helps ensure that the website is user-friendly and meets the needs and expectations of residents using it to express their opinions.

Analytical validation

3.5 Ethical considerations

1. **Bias and Fair Representation -** To ensure a fair representation, it's crucial to recognize and address potential biases in data collection and analysis. Gathering opinions from the diverse population of the county ensures representation. Striving for inclusiveness demands seeking feedback actively from underrepresented groups while addressing any data imbalances.
2. **Transparency and Accountability -** The website's purpose and functionality should be fully disclosed to promote openness and honesty for all users. It is essential to communicate how user feedback will be utilized in decision-making processes that contribute to the improvement of services. Include a contact channel allowing users to easily express their concerns, clarify data usage or seek support from the operator.
3. **Data Protection and Security**: Take appropriate measures to securely store and handle user data. Implement encryption, access controls, and regular security audits to protect against unauthorized access, data breaches, or misuse of information. Avoid sharing personally identifiable information with third parties without explicit user consent.
4. **Responsible Use of Data**: Handle user data responsibly and refrain from using it for purposes beyond the scope of the website's stated objectives. Avoid selling or sharing user data with advertisers or other entities without explicit consent. Emphasize the importance of data anonymization and aggregate reporting to protect individual privacy.

3.6 Expected outcomes

4. References:

1. SHAHOW A.A (February 24th 2023)-Devolution Has Not Delivered for the People of North-Eastern Kenya.https://www.theelephant.info/features/2023/02/24/devolution-has-not-delivered-for-the-people-of-north-eastern-kenya/

2.WAO.L. (2020)Community-Media Engagement on Water,Sanitation, and Hygiene by KEWASNET in Kilifi County https://kewasnet.co.ke/community-media-engagement-on-water-sanitation-and-hygiene-by-kewasnet-in-kilifi-county/

4. Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June 1). *An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges*. IEEE Xplore. https://doi.org/10.1109/IEC47844.2019.8950616

5. Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019, November 1). *Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset*. IEEE Xplore. https://doi.org/10.1109/SMART46866.2019.9117512

6. Retweek, C., & Sandeep, A. (2020). REACT.JS AND FRONT END DEVELOPMENT.