K-Nearest Neighbor in Assessing Trends of Cameroonians Most Attractive Communal and Cultural Diversity Cities in Poland Based on Natural Language Processing and Artificial Intelligence

Pascal Muam MAH,
AGH University of Science and Technology, Krakow, Poland
mahpascal01@gmail.com

Gilly Njoh AMUZANG,
Silesian University of Technology Gliwice, Poland
tahgilly151@gmail.com

Micheal Blake Somaah ITOE

Silesian University of Technology Gliwice, Poland
michaelblake4u@gmail.com

Ning Frida TAH
Silesian University of Technology Gliwice, Poland
fridatah@gmail.com

Abstract

Introduction: Time and distance are the most essential factors in life. The time spent on content reveals the NLP structure and distance between key content words reveals KNN algorithm representation. One of the most widely used classification algorithms in machine learning is the K-nearest neighbor (KNN). In recent years, the trend of migration has increased tremendously as compared to a decade ago.
Aims; Aim to evaluate the impact of migration in recent years using KNN to understand the most essential time-distance factors promoting such adventures. Also, revealed how time (NLP) and distance (KNN) necessitated Cameroonians' zeal for Poland and city choice.
Problem: Time on data content (NLP) and distance on data content (KNN) is causing millions of pieces of information to go unnoticed. Also, a lot of cacophony has been in the air about Cameroonians scrambling in the Polish embassy for Polish visas, and in Polish borders that sends a mixed signal to social media, western world about the Cameroon government.
Material and method: This study uses natural language processing to capture key comments, survey distance between key content and assess how artificial intelligence automates content to the understanding of Cameroonians to identify the most attractive communal and cultural diversity cities in Poland. The study uses questionnaires with the help of some embodiment factors that explain migration trends of most Cameroonian to Poland and choice of cities. The study uses a TF-IDF and bag of words model to identify K value for K-nearest neighbor approach and a concise NLP classified key content.
Results: Based on K-nearest neighbor analysis TF-IDF and bag of words model, where results were classified into most attractive, and least attractive. Warsaw city is the most attractive communal and cultural diversity city in Poland while Gdańsk is the least. A total of 17 questionnaires reveal that Warsaw is the most attractive communal and cultural diversity city in Poland with about 49% respondents and Gdansk is the least attractive with about 2% respondents.
Conclusion: Warsaw is the K-nearest city closest to Cameroonians most attractive communal and cultural diversity city. Based on K-nearest neighbor analysis, the study concluded that the fabulous and exotic facilities, services, amenities, education system, job opportunities and affordable life are amongst the reasons for the migration trends and choice of cities.

Keywords; K-nearest neighbor, natural language processing, artificial intelligence, data classification, migration

Introduction

K-nearest neighbor is supervised learning that deals with regression problems and classification challenges [1,3]. K-nearest neighbor algorithm used machine learning technique to predict group attributes of data instances [2]. Text classification has been widely achieved using K-nearest neighbor algorithms. K-nearest neighbor has become a more popular algorithm in text classification [7]. The art of classifying data or a text content to enable and ensure understanding is a natural language processing approach that aims at fulfilling the K-nearest neighbor algorithm that holds the most essential information.

The KNN and NLP fulfill the process of monitoring and classifying public opinion based on different aspects of communication such as understanding, reporting, informing, investigating, revealing, forecasting and predicting.

Artificial intelligence helps in automating all the processes and steps that require human efforts [5]. The process of monitoring public opinion in order to reduce the time required to understand inside and outside of data content has been achieved recently with automated AI applications. Thus AI narrows down the distance between actual receivers of data content and time required. Most applications of natural language processing use artificial intelligence to automate classified information. Fertile conducting of more complex activities that require a greater vision require artificial intelligence [4]. Text classification, and clustering uses application of natural language processing to capture and redistribute public opinions in a proper network approach and in a well-organized structure. Natural language processing in the field has presented one of the best approaches and applications of public opinion with the help of keywords.

Time on data content (NLP) and distance on data content (KNN) is causing millions of pieces of information to go unnoticed. That is why this study uses the Cameroonian migration trends to explain the situation that is not fully supported by data but by human emotions. There's a need to tackle the causes to allow ample time to solve this issue than throwing accusations on individual actions. Cameron has been notorious for corrupt practices and bad governance. In recent years, since 2016 till date there's a lot going on in Cameroon but this has not been attempted to resolve by the international community. This study decided to attempt to evaluate the impact of migration in recent years using KNN to allow the public global community and the media to understand the most essential time-distance factors promoting migration trends of Cameroonians. This study can be applied to other countries.

Natural language processing has established a lot of opportunities to automate human activities through software with a number of automated process steps using machine learning and artificial intelligence. Digital transformation is the major area that has established the opportunity for different applications to automate human activities through software [6].The K-nearest neighbor algorithm is one of the most used machine learning techniques that automate human activities through spoken, written and documentary

Study Validation Technique and methodological approach

The process flow of the study follows application of model experimental validation. The model validation and verification stages in this study are summarized in figure 1 below.

In this section, the experimental validation uses deep learning models and natural language processing. Deep learning focuses on the K-nearest neighbor algorithm and natural language processing focuses on frequency inverse document frequency (TF-IDF) and bag of words (BOW) Model.

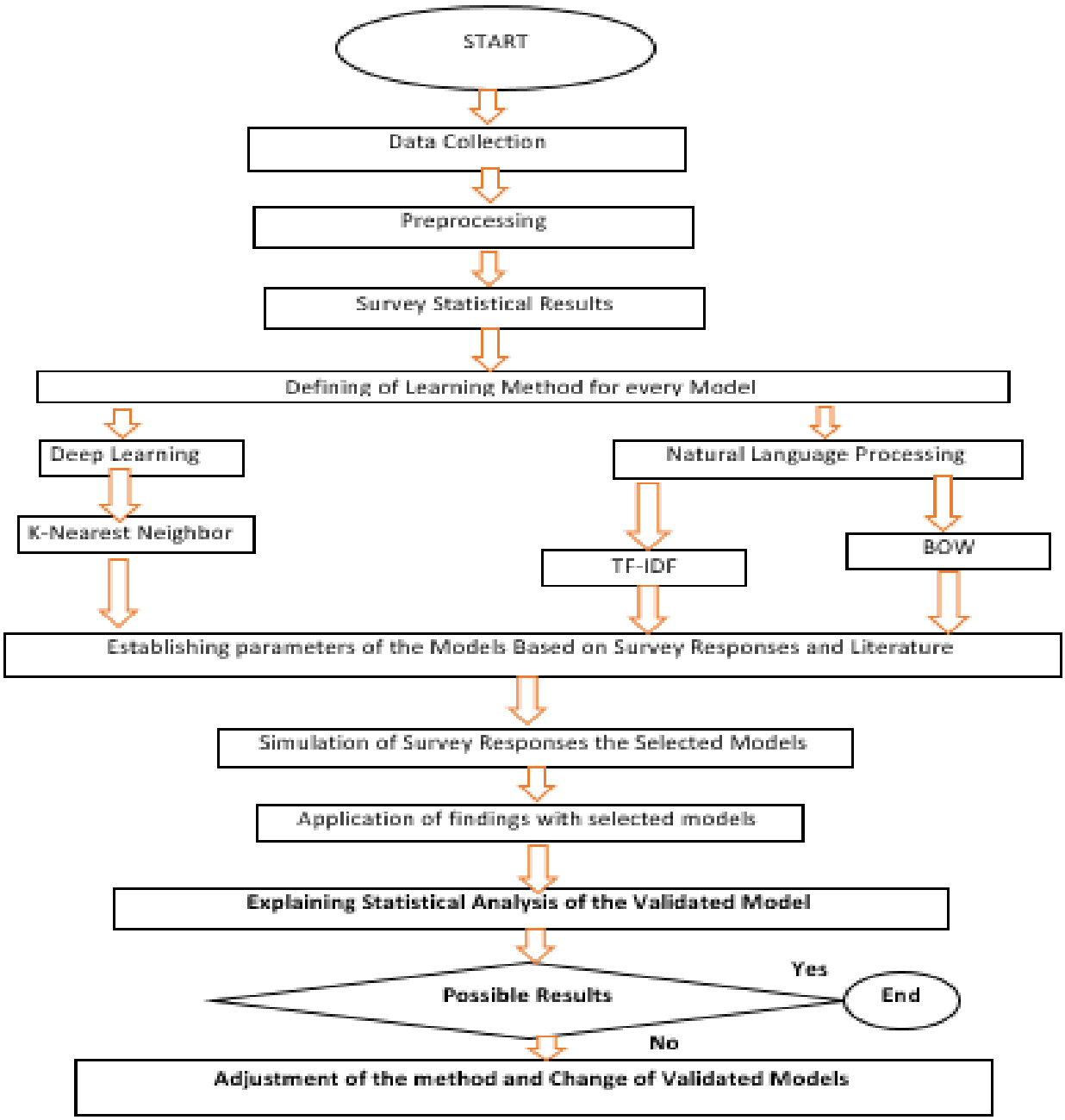

Figure 1: validation technique and material flow

Figure 1 represent preprocessed and tested by statistical validation methods, and it is divided into sample technique and testing sets validation using the K-nearest neighbor algorithm and bag of words model of natural language processing. The sample survey responses were used to learn to achieve the bag of words model algorithm.

Literature Review

This section contain definition of key terms, and different between K-nearest neighbor and natural language processing

Definition of key terms

K-nearest neighbor: This is a supervised simple machine learning algorithm used to solve classification and regression problems [8]. K-nearest neighbor is one of the common classification methods, which function on the basis of the use of distance measures [9].

Natural language processing: This is a branch of artificial intelligence that deals with the design and implementation of algorithm systems able to interact through human language [10]. Natural language processing creates unique opportunities that systematically document users' details from digitized free-text [11]. Natural language processing is the task of converting unstructured human language [12].

Artificial intelligence: This is the science and engineering of making intelligent [13]. Artificial intelligence is the study of how to build or program computers to enable them to do what minds can do [14].

The K-nearest neighbors algorithm is a non-parametric supervised learning method first introduced by Evelyn Fix and Joseph Hodges in 1951 [15, 16]. The components that affect the importance of words in a document called Term Frequency factor and Inverse Document Frequency (IDF) factors [17, 18]. Term frequency of words in a document (TF) is weight that depends on the distribution of said document words. The importance of words in s document are understood based on their frequency. When talking about inverse document frequency, each word in the document and its weight depend on the distribution frequency [19, 20]. There are different ways to compute independence words especially particular words class. The chi-square statistic is a score that can be used to determine the features of document with the highest values.

Natural language processing provides approaches that are used to achieve data normalization and data dimensionality reduction [21, 22]. The birth of modern digital services improved the classification effect thanks to natural language processing. This study uses K-nearest neighbors to analyze the closest factors leading migration trends of Cameroonians to Poland and their most attractive communal and cultural diversity cities. Amongst eight factors that determine the nearest desire for travel to Poland and the most attractive communal and cultural diversity cities, there exist the most attractive factor. The study uses the "K" in K-nearest neighbor (KNN) to be the push factor trends or parameter that determines the closest neighbor to include in the alternative process for the most attractive communal and cultural diversity cities and migration motivation. The study uses the following cities (Warsaw, Krakow, Silesian, Wroclaw, Poznan, Gdansk, Lodz, and Lublin) as data points with the help of eighteen alternative push factors of parameters that forces Cameroonian to focus on the closest factor "K" kNN. Amongst seventeen factors there exist only one or few closest desires "K".

Differences between K-nearest neighbor and natural language processing

The main distinction between natural language processing when performing content classification and clustering with the K-nearest neighbor algorithm is time and distance.

Natural language processing engages in various processes to reduce the time we require to understand a long line of content or document while the K-nearest neighbor algorithm represents the distance between datasets that belong to the same layers of document content.

Natural language processing required a Preprocessing approach to enable and ensure feature extraction while the K-nearest neighbor algorithm represents distance between datasets with features to ensure a proper choice.

Natural language processing required preprocessing and introduction of datasets that K-nearest neighbors require adjusting the parameters to achieve a classifier. K-nearest neighbor (KNN) provides a more helpful method of data classification that helps us understand the likelihood that something is going to happen or has happened or is close to happening.

On the other hand, natural language processing breaks down the steps, stages and present data of what has happened or will happen or data of what is close to happen.

Applied Method

This section contain methodological analysis on data preprocessing, construction of word vector structure with TF-IDF, analysis of Bag of Words with survey responses, Weight Strategy, and TF-IDF Application Calculation Process of data Classification in the following paragraphs provide details on how the study was carried out.

Data Preprocessing

Data were collected from 40 respondents in seventeen (17) categorical survey questions, for a choice of eight (8) attractive cities. The cities were Warsaw, Krakow, Silesian, Wroclaw, Poznan, Gdansk, łódź and Lublin. The respondent indicated a random identity (respond sheet) answer sheet that includes cities as class of dataset like Warsaw, Krakow, Silesian, Wroclaw, Poznan, Gdansk, łódź, and Lublin. 40 respondents were identified randomly distributed as per survey as training data, including Warsaw 418, Krakow 181, Silesian 65, Poznan 55, Wroclaw 33, Gdansk 18, łódź 38, and Lublin 44,

Respondents are registered in the study survey as a list, and total response data are considered as a list. Based on the survey responses, a segmentation method was applied based on bag of words model. This study segments the city names as word counts that appear in each question for use in training and test sets.

Construction of word vector structure with TF-IDF

TF-IDF means term frequency-inverse document frequency. TF-IDF is a statistical method that measures and evaluates how relevant a word is to a document in a collection of documents. This section deals with structured representation of survey responses. Here we constructed a Word Vector Structured representation of text categorization that mainly counts the frequency of words in the survey response.

The study uses eighteen questions classified into eight cities to examine the most and least attractive communal and cultural diversity city.

This was done by multiplying the metrics on how many times a word (city name) appears for each question in the survey responses, and the inverse response frequency of all the total words (city names) across the set survey datasets. TF-IDF has many uses. The most significant is its ability in automating text analysis which is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

Analysis of Bag of Words with survey responses

Bag of Words (BoW) is a process that simply counts the frequency of words in a document. The vector for a document, unlike the study survey, has the frequency of each word as indicated by studies in the corpus for documents. The key difference between bag of words and TF-IDF is that the bag of words does not incorporate any sort of inverse document frequency (IDF) but it is only a frequency count (TF).

Weight Strategy

TF-IDF indicates that the word frequency reverses the frequency of the document. This study uses TF-IDF with the intended stress that Word frequency should not reduce the frequency of documents but instead make the documents with highest frequency words are more visible. The sole aim is to stress the importance of the survey responses. TF-IDF assumed that a word or phrase that appears in a high frequency in a document rarely appears in other documents and when this happens, it is considered suitable for Classification.

TF-IDF Application

Determining the importance of a word is to a document is useful in many ways.

Information retrieval. TF-IDF invention was for document search and is also used to deliver results of survey, search queries and keyword automation. TF-IDF gives frequently appear words a higher score. Most search engines use TF-IDF scores in its algorithm. The study uses TF-IDF to obtain results for the study based on the strict survey responses.
Keyword Extraction. TF-IDF is very useful for keywords extraction from text. The study sees this very close to the questionnaires used in this study. The questionnaires were tailored to the basic and concise needs of Cameroonians in Poland to understand their most favorite attractive communal and cultural diversity city. The site uses search engine to identify the most desired attention for foreigners and a survey questionnaire was built. The highest scoring words in the survey responses indicated the most relevant needs of Cameroonians and therefore was considered keywords for that survey.

The Classifier

K-nearest neighbor algorithm calculates that most of the k nearest neighbors in a dataset or document belong to a certain category, and the sample also belongs to this category. The algorithm in this study involves a simple main objective of distance measurement. The k-value selection is based on a bag of words (BOW) model. The experiment of K value selection is performed using a bag of words model and the optimal k value is selected from the eight (8) categories of datasets set in this study for the most city and respond from survey determine the k value.

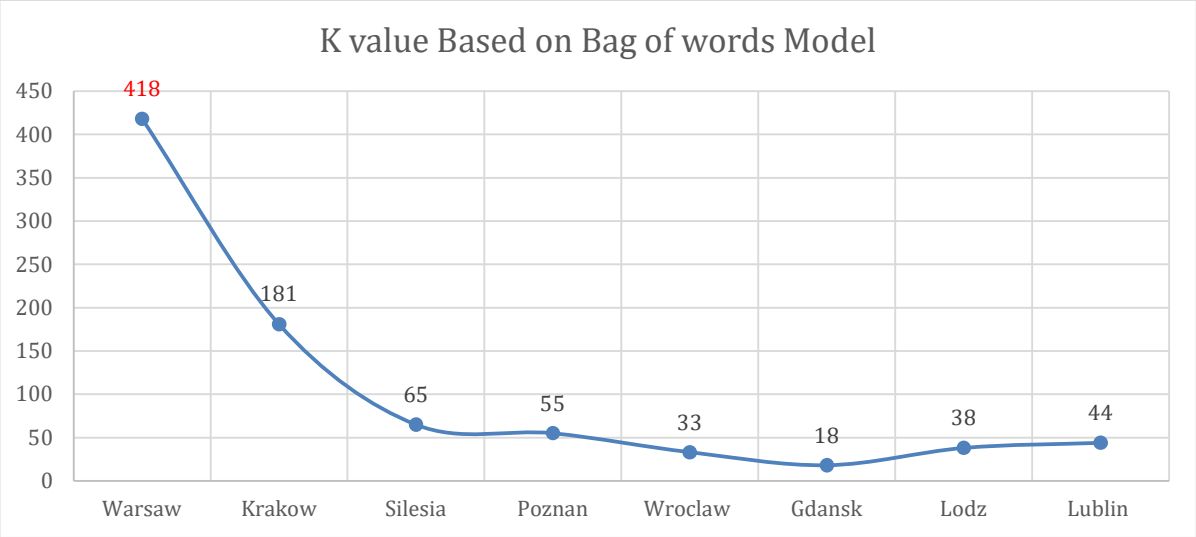

Figure 2 K value based on bags of words

The seventeen (17) survey questionnaires generate data by a simple cross-validation method of eight (8) categorized selected cities. The test set is validated by using the datasets obtained as the training set interfere with by a sample selection of frequently appeared words using bag of words model .

The results show that the effect is the best appearance or frequency based on bag of words model is 418. K=418. Warsaw was the most frequency amongst all the eight categories set in this study for seventeen (17) sample questionnaires.

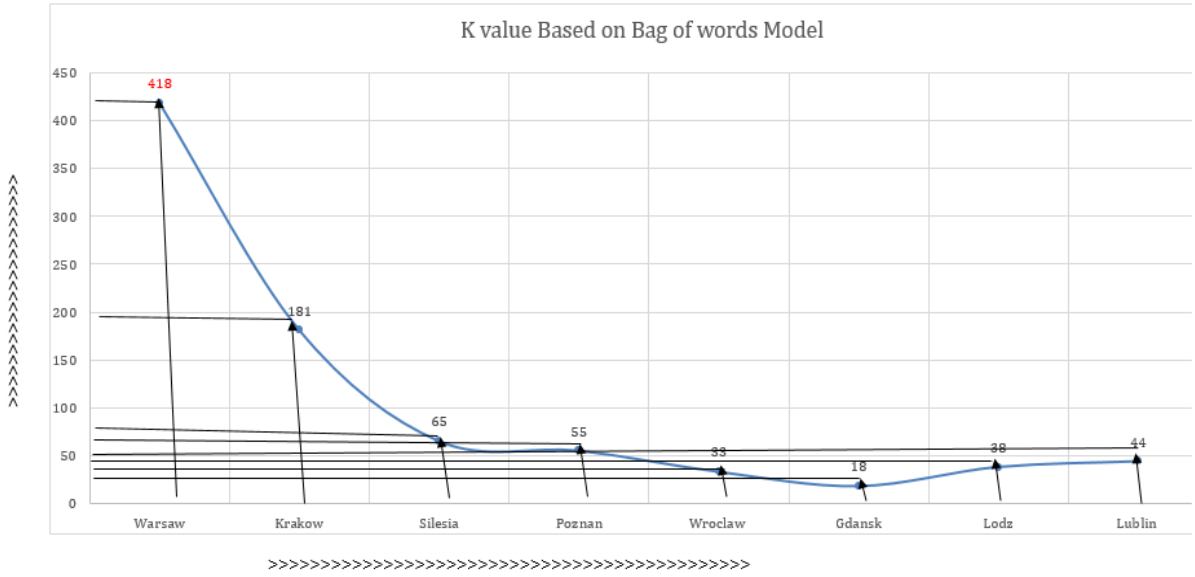

Figure 3: frequency and dimensionality law of BAW model

Figure 3 represents BAW model assumption that the higher the frequency of words the lower the dimensionality. Based on the study, Warsaw contain the highest frequency and represents the K value in the study.

Accuracy validation technique $= \frac{TP+TN}{TP+TN+FP+FN}$

Warsaw=TP, Gdansk=TN, Krakow=FP, and Wroclaw=FN

True Positive=TP, True Negative=TN, False Positive=FP, and False Negative=FN

$$\frac{450+18}{450+18+181+33} = 0.68\%$$

$$= 68\%$$

Results

This study uses sample questions to determine the most attractive communal and cultural diversity city in Poland. The sample questionnaires were shared only amongst Cameroonians. The sample questions were share in the WhatsApp group for Cameroonians in Poland. The following statistics were collected. The study score words with the chi square function and sorted the words frequency using the bag-of-words model is clearly better than the filtered mode [23]. Survey questions and respondents are possible with the application of bag of words model and chi square [24]. Empirical analyses are based on chi square, especially non-filtering findings like the survey applied in this study.

Formula of Chi Square

$$X^2 = \sum . \frac{(Oi - Ei)^2}{Ei}$$

$$x^2 = Chi\ Square$$

$$Oi = Observed\ value$$

$$Ei \cdot = Expected\ value$$

The statistics were evaluated based on chi square observation from sample questionnaires to statistical presentation.

.

Table 1: Survey Based on Bags of word (BOW)

No	Surey Questions	Warsaw	Krakow	Silesia	Poznan	Wroclaw	Gdansk	Lodz	Lublin	Total Responses per question
1	The Most Learning Institutions City	28	10	3	3	2	2	2	2	52
2	The Most Beautiful City	26	12	4	2	2	1	3	2	52
3	The Most Learning Facilities City	25	11	3	3	0	0	1	2	45
4	The Most Foreign Learning Courses City	26	10	3	4	1	0	3	4	51
5	The Most Polish Language Learning Centers	27	13	0	4	0	1	4	2	51
6	The Most Attractive Jobs Opportunities	29	11	5	4	3	1	2	2	57
7	The Most City with Foreign Recipes	28	13	2	5	4	1	4	1	58
8	The Most City with Foreign Access to Housing	22	13	7	4	4	0	4	2	56
9	The Most Foreign Accessories City	26	15	5	4	2	1	0	4	57
10	The Most City with Foreign Organizational Units	30	11	4	4	2	1	0	3	55
11	The Most Foreign Shops City	26	9	9	4	4	1	2	3	58
13	The Most Secure City	25	6	5	5	1	1	3	4	50
14	The Most Inhabitant City with Cameroonians	23	10	6	3	5	2	6	3	58
15	The Most Fastest City with Legalization of stay	22	16	4	4	1	2	1	5	55
16	The Most Expensive City	28	10	4	2	0	2	2	2	50
17	The Most Welcoming City	27	11	1	0	2	2	1	3	47
	Grand Total	418	181	65	55	33	18	38	44	852

Table 1 was made up of seventeen (17) questions. The target of the questionnaires was to determine the most attractive communal and cultural diversity city in Poland for Cameroonians. In the survey, eight (8) cities were targeted.

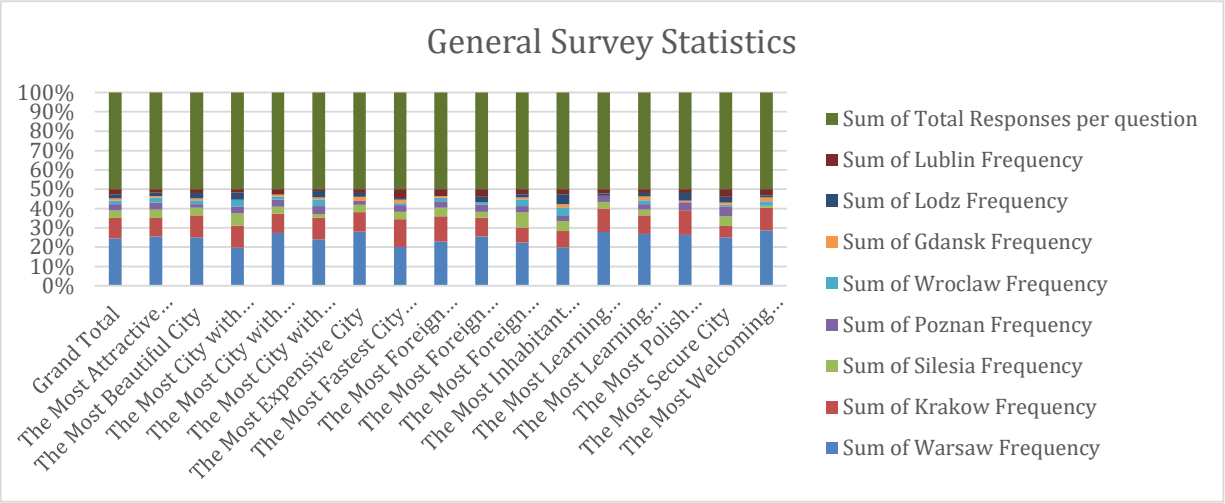

Figure 4 general survey statistics.

Figure 4 represent the type of seventeen (17) questions, number of response for each question. Each question was to evaluate and examine the most attractive communal and cultural diversity cities.

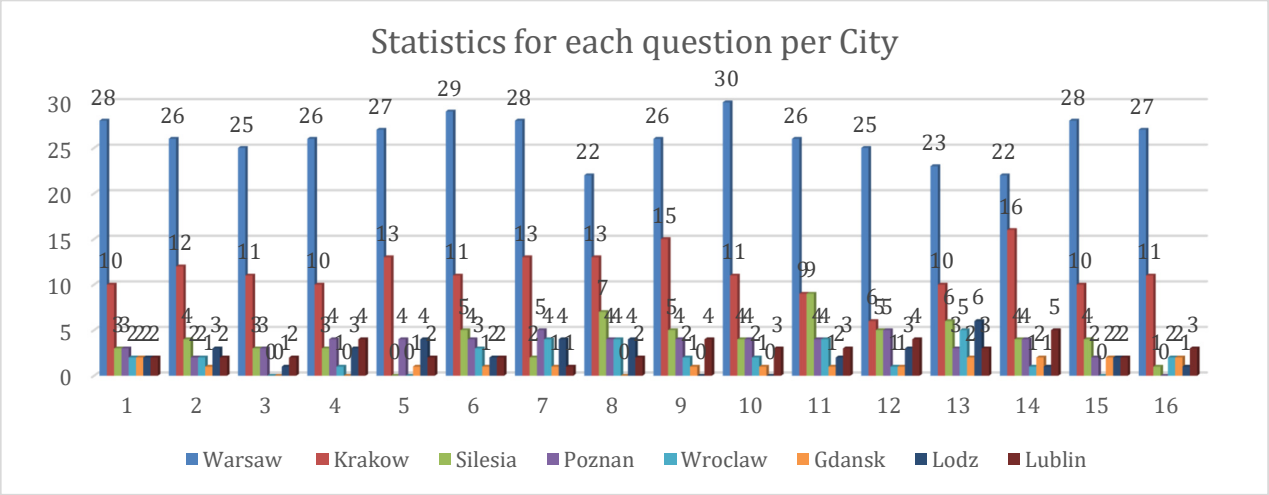

Figure 5 statistics per city

Figure 5 represents the eight cities and the amount of respondents to each. From figure 5, Warsaw is the highest city with a positive response. The least city was Gdańsk with about 18 respondents

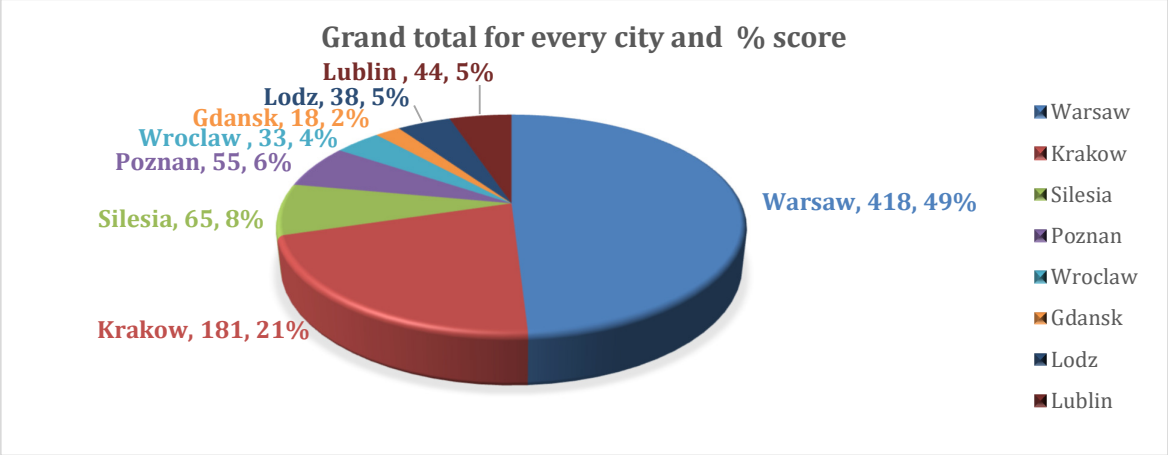

Figure 6: percentage score per respondents

Figure 6 represents eight cities with Warsaw with 49%, Krakow with 21%, Silesia with 8%, Poznan with 6%, Wroclaw with 4%, Gdansk with 2%, łódź with 5% and Lublin with 5%.

Conclusion

In this study, we investigate the most attractive communal and cultural diversity cities in Poland for Cameroonians with the used of sample questionnaires. Results indicated Warsaw as the most attractive communal and cultural diversity city in Poland. The study findings were validated using deep learning with K-nearest neighbor algorithm structure analysis and natural language processing validated model of bag of words and frequency-inverse document frequency (TF-IDF). The study confirm the following:

I. The accuracy of bag-of-words can be achieve by simple procedure and lay down rules.
II. The performance of classifier of a set data for any task can be achieve depending on the technicality of the dataset and intended purposed.
III. The application of higher thresholds for chi-square model can enhances feature dimension without necessarily optimization

Declaration Conflict of Interest

We certify that we have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. We have no financial or proprietary interests in any material discussed in this article.

Declaration of material used

All data underlying the results are available as part of the article and no additional source data are required or reserved somewhere.

References

- Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modelling and graphics (pp. 99-111). Springer, Singapore.
- Soofi, A. A., & Awan, A. (2017). Classification techniques in machine learning: applications and issues. Journal of Basic & Applied Sciences, 13, 459-465.
- Hackeling, G. (2017). Mastering Machine Learning with scikit-learn. Packt Publishing Ltd.
- Jarrahi, M. H. (2019). In the age of the smart artificial intelligence: AI's dual capacities for automating and informating work. Business Information Review, 36(4), 178-187.
- Romao, M., Costa, J., & Costa, C. J. (2019, June). Robotic process automation: A case study in the banking industry. In 2019 14th Iberian Conference on information systems and technologies (CISTI) (pp. 1-6). IEEE.
- Schmitz, M., Dietze, C., & Czarnecki, C. (2019). Enabling digital transformation through robotic process automation at Deutsche Telekom. In Digitalization cases (pp. 15-33). Springer, Cham.
- Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. Procedia Engineering, 69, 1356-1364.
- Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.
- Modaresi, F., & Araghinejad, S. (2014). A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water resources management, 28(12), 4095-4111.
- Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing, 470, 443-456.
- Rouillard, C. J., Nasser, M. A., Hu, H., & Roblin, D. W. (2022). Evaluation of a natural language processing approach to identify social determinants of health in electronic health records in a diverse community cohort. Medical Care, 60(3), 248-255.
- Voytovich, L., & Greenberg, C. (2022). Natural language processing: Practical applications in medicine and investigation of contextual autocomplete. In Machine Learning in Clinical Neuroscience (pp. 207-214). Springer, Cham.
- Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. Metabolism, 69, S36-S40.
- Boden, M. A. (Ed.). (1996). Artificial intelligence. Elsevier.

- He, C., Ding, C. H., Chen, S., & Luo, B. (2021, November). Intelligent Machine Learning System for Predicting Customer Churn. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 522-527). IEEE.
- Hu, X., Wang, J., Wang, L., & Yu, K. (2022). K-Nearest Neighbor Estimation of Functional Nonparametric Regression Model under NA Samples. Axioms, 11(3), 102.
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of documentation.
- Polettini, N. (2004). The vector space model in information retrieval-term weighting problem. Entropy, 34, 1-9.
- Rathi, R. N., & Mustafi, A. (2022). The importance of Term Weighting in semantic understanding of text: A review of techniques. Multimedia Tools and Applications, 1-23.
- Ozyegen, O., Kabe, D., & Cevik, M. (2022). Word-level text highlighting of medical texts for telehealth services. Artificial Intelligence in Medicine, 127, 102284.
- Nistor, A., & Zadobrischi, E. (2022). The influence of fake news on social media: analysis and verification of web content during the COVID-19 pandemic by advanced machine learning methods and natural language processing. Sustainability, 14(17), 10466.
- Kaczmarek, I., Iwaniak, A., Świetlicka, A., Piwowarczyk, M., & Nadolny, A. (2022). A machine learning approach for integration of spatial development plans based on natural language processing. Sustainable Cities and Society, 76, 103479.
- Fouzia Sayeedunnissa, S., Hussain, A. R., & Hameed, M. A. (2013). Supervised opinion mining of social network data using a bag-of-words approach on the cloud. In Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012) (pp. 299-309). Springer, India.
- Obasa, A. I., Salim, N., & Khan, A. (2016). Hybridization of bag-of-words and forum metadata for web forum question post detection. Indian Journal of Science and Technology, 8(32), 1-12.

respond from survey determine the k value