# CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

|  |  |
| --- | --- |
| Module Title: | Advanced Data Analytics  Big Data Storage and Processing |
| Assessment Title: | Game Rank Prediction by Player Performances |
| Lecturer Name: | David McQuaid  Muhammad Iqbal |
| Student Full Name: | Sule Beste Kapci |
| Student Number: | 2022327 |
| Assessment Due Date: | 26/05/2023 |
| Date of Submission: | 26/05/2023 |
| GitHub: | https://github.com/besteli/CCT\_Asigment2022-2023 |

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Introduction**

Natural language processing is a subcategory of artificial intelligence. There are two different languages ​​in the computer world. These are programming languages ​​and natural languages. By natural languages, we mean the natural languages ​​that humans speak; Languages ​​such as Turkish, English and Korean can be given as examples. The process by which machines receive and process the language spoken by humans is called natural language processing. [1]. Utilizing Twitter's API, the data for this study were generated. The purpose of this study is to examine how the attitudes of Twitter users shifted between the semifinals and the finals by contrasting the sentiments conveyed in their tweets at both stages of the competition. Simultaneously, the most essential topic to the majority of users was identified by analyzing the most popular terms and keywords. At the same time, the most popular terms and hashtags were investigated, and a survey was conducted to determine the subject matter that users discussed the most. In this way, it can be learned that people's thoughts and feelings about the Eurovision 2023 contest, which topics they focus on the most, which country and which song they talk about the most. As a result of these analyzes, people's final predictions can be researched and it can be estimated which country they support.

**Literature Review**

With the development of natural language processing and artificial intelligence systems, it accelerates the work of imitating human speech. The literature study has shown that natural language processing studies have accelerated in many areas of life and are quickly integrated into daily life. A summary of the literature studies has been presented in this study.

In their study, Locke and Basshal (2021) conducted a rewiev study that using natural language processing to predict patient outcomes through medical notes can create models that diagnose early-stage chronic diseases. [2]

Shankar and Parsana (2022) tried to understand which marketing models are suitable for natural language processing (NLP) models used in the study of conversations and marketing interactions through digital technologies and media. [3]

Ruder and Peters (2019) examined the transfer learning method. They talk about the development of transfer learning methods and architectures in the field of natural language processing, modern transfer learning methods, the training stages and results of the models. [4]

Dodoo and Wen (2021) analyzed people's twitter posts in the field of advertising using natural language processing in their study. Sensitivity analysis was performed using the Python programming language. As a result of the study, it was seen that people approached this issue in eight different ways. The most positive issue was "Artificial intelligence supported marketing tools" and the most negative issue was "Artificial intelligence participation in social media campaigns". [5]

**Method**

In this study, the Hadoop ecosystem will be utilized and this study will be carried out on the Linux operating system.

**- Hadoop**

Hadoop is an open source library used to process big data on ordinary servers. [6] This library, which has very high processing power, provides the ability to manage multiple tasks simultaneously. The data was loaded with sparkSession.read.load() and visualized as a schema. By getting information about the data, the most common words in the texts and their frequencies were found. Hashtag analysis was made and the most used ones by users were found.

- **MySQL**

MySQL provides access to data via SQL while structuring data in tables. Hadoop is used to store and process large amounts of data while MySQL is used for structuring, managing and querying data. In this study, query operations were carried out by transferring our data to the table.

- **MongoDB**

MongoDB is an open source NoSQL database application. It has a scalable structure. It supports range queries, search by field, and regular definition searches. It can create multiple copies of the original data, thus preventing data loss. [7]

In this study, MongoDB and MySQL databases were selected in order to compare the two databases. Data loading times of the two databases, respectively; It is 3.98 seconds in MySQL and 0.97 seconds with MongoDB.

**- Sentiment Analysis Transactions**

**1. Twitter API**

The Twitter API account used in this study is a basic account. There is a monthly data limit of 10000 in the Basic account and 1000 data can be withdrawn daily. There is no day limit when capturing the data. For this reason, as a result of each data extraction process, 1000 data were retrieved daily from the time of processing.

During the data extraction process, the filtering process was applied to the data retrieved through the parameters in the query command. “#Eurovision” has been determined for the hashtag. Retweets and tweets with links are not included in the dataset. Only tweets with English as language selection are included in the dataset.

**2. Data Set**

In the dataset, there are tweets of Twitter users regarding the Eurovision 2023 contest. The data collection process is as follows. Data were started to be collected 2 days after the 1st Semi-finals, which took place on 9 May. A total of 7632 data were collected from 11 May to 16 May after the Finals. The data set consisting of 8 parts was combined using the pandas library and turned into two different data sets. These are prepared before and after the Eurovision Finals. In the data sets, there are columns for Id, twitter text data and date of creation.

**3. Data Understanding**

**3.1. Before the Eurovision Finals**

The data set created before the finals contains 4000 data. There are 3 columns named “id, text, created\_at”.

**3.1.1. Cleaning the Data**

The data posted by Twitter users includes many different characters, emojis and symbols. When performing a natural language processing exercise, the data needs to be cleared of such special characters, emojis, symbols and numbers. In this study, the module named “re” was used to clean the text data in the data set. This module allows to quickly check whether a given string matches or contains a particular pattern. While deleting expressions such as numbers, special characters, links, and hashtags in the data set, we also convert all uppercase letters to lowercase and make the characters a single standard. We add the result of the text cleaning process to our dataset as a new column.

Stopwords are the identification of words that are used extensively in texts and have no meaning (“and”,”the”,”to). These words are not important in feature extraction.

**3.1.2. Analysis of Tweets**

**a. Most Used Words**

In this section, the most used words and hashtags by the users were examined by examining the tweets. The top 10 words most frequently used in tweets before the finals are as follows: “Eurovision, tonight, song, im, final, year, day, like, vote, time”. The 10 most popular words used after the finals are as follows: “Cha, song, year, Eurovision, Like, Last, People, im, Sweeden, Loreen”. At the end of this analysis, it was seen that before the finals, users talked about topics such as competition, song, voting. After the competition, it is seen that the contestants mostly talked about the Finnish song "Cha Cha Cha". This song came in 2nd place as a result of the competition. At the same time, Loreen, who is the winner of the competition, and her country are in the list of the most used words, but they are in the 9th and 10th places, respectively. This shows that although the Finnish song Cha Cha cha is in the 2nd place, it is spoken more than the 1st Loreen. If this list is expanded to the first 20 words, it is seen that the word Finland is also included in the list. From here, we can say that users tweet about the first two songs the most. The results obtained are visualized through different graphics, making them easier to examine.

**b. Hashtag Analysis**

Hashtag analysis is performed to find which hashtags users use most in their tweets. Here, the words starting with the "#" symbol are selected and the number of times they are used is calculated. Among the top 20 hashtags in tweets before the finals, 6 hashtags are #Eurovision. #Liverpool, #Ukraine, #Austria, #Australia are the most mentioned country hashtags. When the hashtags thrown after the finals are examined, 5 hashtags in the first 20 hashtags are #Eurovision and do not contain hashtags belonging to any country.

**c. Sentiment Analysis**

Sentiment analysis is the name given to the process of calculating and classifying the opinions/statements indicated by a piece of text through various algorithms in order to evaluate the positive, negative and netural attitude of a user, author or article towards a certain subject.

In this study, it will be examined which emotion is intense from the tweets of twitter users before and after the finals and how the emotions of the users change with the competition. By determining the polarization scores on the cleaned data using "SentimentIntensityAnalyzer()", it is said that it has positive emotion for values ​​greater than 0, neutral for values ​​equal to 0, and negative emotion for values ​​less than 0.

**1 a. Before the Competition**

When sentiment analysis was applied to the data collected before the competition, 2105 positive, 641 negative and 1254 neutral-minded user data were obtained from 4000 text data. The ratios of the results obtained in the data set are given below.

A picture containing text, screenshot, design

Description automatically generated

Şekil 1: Sentimental Distribution

When the most used positive words are examined, "eurovision, song, final, tonight, day" are in the first 5 rows. When the unique positive words are examined, "beautiful, ok, proud, celebrate, wins" is in the first 5 rows.

When the most used neutral words are examined, "tonight, vote, time, eurovision, song" are in the first 5 rows. When the neutral words that are unique are examined, "rediscova, usvintage, sustainable, ding, gooooo" are in the first 5 places.

When the most used negative words are examined, "eurovision, dont, year, tonight, points" are in the first 5 rows. When the unique negative words are examined, "receives, po, janelle, monae, horrible" are in the first 5 rows.

**1.b. After the Competition**

When sentiment analysis was applied on the data collected after the competition, 2047 positive, 774 negative and 811 neutral-minded user data were obtained from 4000 text data. The ratios of the results obtained in the data set are given below.

A picture containing text, screenshot, font

Description automatically generated

Şekil 1: Sentimental Distribution

When the most used positive words are examined, "year, song, like, cha, Eurovision" are in the first 5 rows. When the unique positive words are examined, "vinyl, grand, fab, vintage, jazz" are in the first 5 places.

When the most used neutral words are examined, "year, Eurovision, next, points, follow" are in the first 5 rows. When the unique neutral words are examined, "rediscova, usvintage, cryptobox, filtered, innovations" are in the first 5 places.

When the most used negative words are examined, "cha, im, year, people, didn't" are in the first 5 rows. When the unique negative words are examined, "poor, die, body, journalists, dead" are in the first 5 lines.

As a result of the transactions, it has been seen that while the positive and neutral emotions of the users have decreased, negative thoughts have now occurred. From here, it can be commented that users are not satisfied with the results of the competition.

**4. Modeling Data**

In this study, a model was created using Bernoulli Naive Bayes Classifier, SVM (Support Vector Machine), Logistic Regression and the results obtained at the end of the predict process were compared. Before creating the model, the data was pre-processed and made ready.

**4.1. Data Preparation**

Before starting the modeling processes, the data before and after the Eurovisin final competition were combined. First of all, the “Sentiment” line, which contains our sentiment analysis results, has been converted into numerical data. Here it is represented by Positive: 1, Neutral: 0, Negative: -1. Emoji, punctuation, hashtag, etc. for preprocessing. The "clean\_tweet" line is used, which has been cleared of stuff. Text clearing processes are described below.

**- Cleaning Stop Words**

These words are commonly used words like “the”, “a”, “an”, “in” that a search engine is programmed to ignore when both indexing and retrieving entries for search. These words do not want to take up space in the database and prolong the processing time. For this, a list of such words is saved and these words are extracted from the dataset.

**- Cleaning and Removing Repeating Characters**

This operation is used to clear duplicate characters. It is used to avoid repeating characters accidentally entered by users. These repetitive characters can lead to errors during data analysis processes. Therefore, repeating characters are removed from the dataset.

**- Cleaning and Removing Numeric Numbers**

This process is used to clean the numeric data in the text data.

**- Getting tokenization of tweet text**

This process is used to break paragraphs and sentences into smaller parts that can be handled more easily.

**- Applying stemming**

This process is used to separate English words into stems. It is also a type of text normalization that does not allow to reduce text dimensionality and standardize on certain expressions, also called roots.

**- Applying lemmatizer**

This is the process of grouping inflected forms of a word to analyze it as a single word in linguistics. Lemmatization is more useful for seeing the context of a word in a document than root extraction. Lemmatization uses some of the speech tags and the meaning of the word in the sentence to see the main context of the document. Therefore, NLTK Lemmatization is an important step towards understanding a text and using it in Natural Language Processing and Natural Language Understanding applications.

**- Train Test Split**

X\_train, X\_test, y\_train, y\_test were obtained by dividing the data set prepared for training by train\_test\_split operation.

- Fit the TF-IDF Vectorizer

This process is used to convert text data into numeric vectors. It is used to tokenize documents, learn vocabulary and reverse document frequency weights and allow encoding new documents.

**4.2. Data Modeling**

**- Model 1: Bernoulli Naive Bayes Classifier**

Bernoulli Naive Bayes is part of the Naive Bayes family. It works faster than other classification models. In case of small amount of data or small documents (eg in text classification), it gives more accurate and precise results than other models. The Accuracy value obtained as a result of Bernoulli Naïve Bayes applied in this study was observed to be 0.55.

**- Model 2: SVM (Support Vector Machine)**

SVM is used to split data into two or more operations. SVM, which can be used in the field of natural language processing, works effectively on high-dimensional data and the performance obtained as a result of classification operations is better than other machine learning algorithms. As a result of the SVM applied in this study, it was observed that the accuracy value was 0.75.

**- Model 3: Logictic Regression**

This algorithm is used in classification problems, and this model can be preferred in classifying the texts in the data set as positive, negative, neutral in sentiment analysis processes. As a result of the logistic regression applied in this study, it was observed that the accuracy value was 0.71. In line with the result obtained from here, it is observed that Bernoulli Naive Bayes Classifier has the lowest accuracy rate. It shows that the effect of this model on the data set is less than other algorithms. The fact that the SVM algorithm has the highest accuracy rate shows that it is an effective option in sentiment analysis.

**- Time Series**

Time series analysis (Time Series Analysis) is a system that allows the analysis of events and transactions that occur within a certain time period, and then allows it to be converted into an insight. In this study, firstly, the data were made ready for analysis by applying preprocessing.

In this study, the most suitable model was selected by comparing SARIMAX and ARIMAX models. Work continued with the ARIMAX model. Model section shows the (p,d,q) values ​​of the ARIMA model. Dept. Variable" section shows the name of the dependent variable applied to the model. Statistical values ​​such as "Log Likelihood" and "AIC" indicate how well the model fits, and lower AIC values ​​indicate better fit of the model. "coef" columns indicate the model's fit. "std err" columns represent the standard errors of the predicted coefficients “P>|z|" columns indicate the p values ​​indicating the statistical significance of the coefficients. Generally, p values ​​less than 0.05 indicate that the coefficients are statistically significant. The "sigma2" column shows the value of the variance predicted by the model.

For this study, the performance of the ARIMA(1,1,1) model was analyzed and the results were evaluated. Below are the results of this analysis.

Log Likelihood and AIC; Log Likelihood value and AIC show how well the model fits. A higher Log Likelihood value and a lower AIC value indicate better fit of the model. In this case, the Log Likelihood value is 3.361 and the AIC value is -0.723.

Coefficients and p-values; The "coef" columns represent the coefficients predicted by the model, while the "P>|z|" columns show the p values ​​indicating the statistical significance of the coefficients. In this case, it can be considered that the ar.L1 and ma.L1 coefficients are not statistically significant since the p values ​​are quite high and greater than 0.05. However, you may need to use a larger dataset and other evaluation methods for a more reliable result.

Other statistics: Statistics such as Ljung-Box (Q), Jarque-Bera (JB), Heteroskedasticity (H), and Kurtosis are also used to evaluate the fit and performance of the model. Interpreting these values ​​and deciding whether the model is a good one often depends on the data set and the specific conditions of the analysis.

As a result, it cannot be said with certainty whether the model is a good model based on this output alone. Using a larger dataset, trying different model parameters, and making a more comprehensive evaluation can provide a more reliable result.

**REFERENCES**

[1] Yildirim, E., Fatih Çetin, Gülşen Eryiğit and Tanel Temel (2014). The Impact of NLP on Turkish Sentiment Analysis. 7(1), pp.43–51.

[2] Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A. and Kitchen, G.B. (2021). Natural language processing in medicine: A review. *Trends in Anaesthesia and Critical Care*, [online] 38, pp.4–9. doi:https://doi.org/10.1016/j.tacc.2021.02.007.

[3] Shankar, V. and Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *Journal of the Academy of Marketing Science*. doi:https://doi.org/10.1007/s11747-022-00840-3.

[4] Ruder, S., Peters, M.E., Swayamdipta, S. and Wolf, T. (2019). *Transfer Learning in Natural Language Processing*. [online] ACLWeb. doi:https://doi.org/10.18653/v1/N19-5004.

[5] Wu, L., Dodoo, N.A., Wen, T.J. and Ke, L. (2021). Understanding Twitter conversations about artificial intelligence in advertising based on natural language processing. *International Journal of Advertising*, pp.1–18. doi:https://doi.org/10.1080/02650487.2021.1920218.

[6] Katal, A., Wazid, M. and Goudar, R.H. (2013). Big data: Issues, challenges, tools and Good practices. 2013 Sixth International Conference on Contemporary Computing (IC3). doi:https://doi.org/10.1109/ic3.2013.6612229.

[7] Celesti, A., Fazio, M. and Villari, M. (2019). A Study on Join Operations in MongoDB Preserving Collections Data Models for Future Internet Applications. *Future Internet*, 11(4), p.83. doi:https://doi.org/10.3390/fi11040083.