

Data Analytics Project Report:

Tweets with language labels

Xiating Cai 521285

Rui Lyu 509141

Table of Contents

1. Problem Statement and Background	4
1.1 Background	4
1.2 Dataset characteristics	4
1.3 Descriptive problem	5
1.4 Predictive problem	5
2. Data Cleaning	6
2.1 Missing Value	6
2.2 Noisy Value	7
2.3 Inconsistent Value	7
2.4 Cleaning Methodology.....	8
3. Data Integration, Transformation, Discretisation and Reduction	9
3.1 Data Integration	9
3.1.1 Schema integration.....	9
3.1.2 Redundancy Handling.....	9
3.2 Data Transformation	9
3.2.1 Coordinates_geo Attribute Transformation	9
3.2.2 Language Attribute Transformation	10
3.2.3 Text Attribute Transformation	10
3.2.4 Aggregation, Normalisation and Generalisation	10
3.3 Data Discretisation	11
3.4 Data Reduction	11
4. Feature Design and Selection	12
4.1 Create Feature Bilingual	12
4.2 Create Feature Area	12
4.3 Feature Selection	13
5. Exploratory Data analysis	14
5.1 Data Structure.....	14
5.2 Attribute/Observation Characters.....	14
5.3 Key Finding	15
6. Descriptive Data Mining.....	18
6.1 K-means algorithm	18
6.2 DBSCAN algorithm.....	18
6.3 Dataset Grouping Result	19
7. Predictive Data Mining	21
7.1 Naïve Bayes Algorithm	21
7.2 Decision Tree Algorithm	21

7.3 Rule-Based Algorithm	21
7.4 K-NN Algorithm	22
8. Evaluation and analysis of results.....	23
8.1 Parameter Optimising	23
8.2 Performance Measure	24
8.2.1 Performance Measure	24
Naïve Bayes Model Performance Vector:	26
8.2.1 Attempt to improve accuracy	26
8.3 Evaluation Methodology.....	26
8.4 Analysis of Results.....	26
9. Others	28
9.1 Lessons Learned	28
9.2 Tools Utilised	28
9.3 Team Contribution	29
Reference	30

1. Problem Statement and Background

1.1 Background

In big data era, the volume of data is getting larger and larger, the scale of data with all kinds of format is extremely huge and even countless, one of the examples is that countless tweets are spread all over the network, as of October 2019, on average, around 500 million tweets sent each day or, 200 billion tweets per year according to the newest statistics from Sayce, D. (2019).

Usually, a tweet message contains lots of information besides the text content itself, such as the location and date of tweets sent, senders' account information associated with the tweet including user id or profile.

1.2 Dataset characteristics

The TweetData.xlsx tweets dataset given with manual annotation of language attributes including tweet_id, coordinates_type, coordinates_geo, place.country, place.name, user_id, tweet_language, and tweet_text.

Almost of dataset attributes are non-numerical type, which means the country, content, language and other attributes are based on the text information, the format of attributes' values is mainly string type, nominal/ordinal data process is one of the major tasks.

Name	Type	Missing	Statistics	
✓ tweet_id	Real	1	Min 439375187159814140	Max 450782336251351040
✓ coordinates_type	Polynominal	6236	Least Point (28749)	Most Point (28749)
✓ coordinates_geo	Polynominal	6236	Least 43.74889 [...] 31518 (1)	Most 43.02472 [...] 059 (209)
✓ place_country	Polynominal	15	Least Spanien (2)	Most España (27134)
✓ place_name	Polynominal	15	Least Zamora (1)	Most Lugo (5518)
✓ user_id	Polynominal	1	Least zzarrillo (1)	Most IRUKLugo (319)
✓ tweet_language	Polynominal	1	Least es/eu/gl/pt (1)	Most es (21417)
✓ tweet_text	Polynominal	1	Least ðŸ™ˆ~ðŸ™‰% [...] s :) (1)	Most Buenos dÃ—as. (18)

Figure 1.1: Dataset attributes type

1.3 Descriptive problem

Problem: what are the popular languages that commonly used in a specific area or region?

Different areas (community, city, country) usually have a preferred language or some commonly used language type, sometimes this language distribution cannot be determined simply by the country or area as the use of language is affected by several other factors such as culture, religion as well as the ethnic group, so determining the popular language in a specific place could be a hard task sometimes.

1.4 Predictive problem

Problem: how to determine the language type of tweets according to tweet's other information without human labor labeling?

Sometimes it is necessary to know the tweet content language used, for example, the tweet company may wish to categorize them according to language labels, but unlike some other attributes that can be attained automatically with tweet itself, getting language type is a tough task.

In the given dataset, the tweet is labeled manually, which could be a troublesome and time-consuming task. With the huge amount of tweet messages in recent years, this human manual process is harder or even impossible, whereas the intelligent artificial technique which is based on data analysis and the automatic process could save the manual labor from the labeling task and make this process much easier.

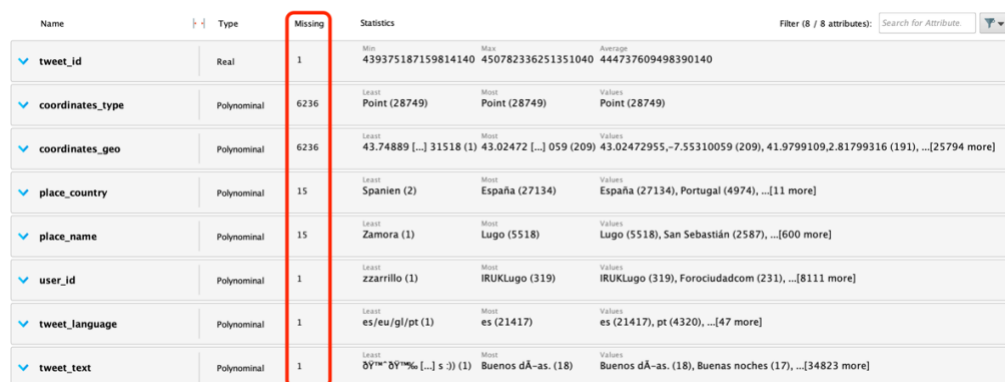
2. Data Cleaning

2.1 Missing Value

After importing and append two dataset tables given, we found there are three types of the missing data value in 34,985 observations:

- First, the major one is some examples' values of attributes of `coordinates_geo` are missing, the number is 6236. See *figure2.1*.
- Second, there are 15 examples' `place_country` and `place_name` values are missing. See *figure2.1*.
- Third, one record's `coordinates_type`, `coordinates_geo`, `place_country`, and `place_name` values are all missing, and another record does not have `user_id`, `language`, and `text` value.

For the third type missing value, consider just two special cases compared with 34,958 dataset records, and their value is hard or impossible to impute because of too many missing attributes and missing attributes are hardly relevant to others, we just removed these two records.



Name	Type	Missing	Statistics
tweet_id	Real	1	Min: 439375187159814140, Max: 450782336251351040, Average: 444737609498390140
coordinates_type	Polynomial	6236	Least: Point (28749), Most: Point (28749), Values: Point (28749)
coordinates_geo	Polynomial	6236	Least: 43.74889 [...], 31518 (1), 43.02472 [...], 059 (209), 43.02472955, -7.55310059 (209), 41.9799109, 2.81799316 (191), ...[25794 more]
place_country	Polynomial	15	Least: Spanien (2), Most: España (27134), Values: España (27134), Portugal (4974), ...[11 more]
place_name	Polynomial	15	Least: Zamora (1), Most: Lugo (5518), Values: Lugo (5518), San Sebastián (2587), ...[600 more]
user_id	Polynomial	1	Least: zzarrillo (1), Most: IRUKLugo (319), Values: IRUKLugo (319), Forocidadcom (231), ...[8111 more]
tweet_language	Polynomial	1	Least: es/eu/gl/pt (1), Most: es (21417), Values: es (21417), pt (4320), ...[47 more]
tweet_text	Polynomial	1	Least: ðŸ™ˆ ðŸ™ˆ% [...], s :) (1), Most: Buenos dÃ¡-as. (18), Values: Buenos dÃ¡-as. (18), Buenas noches (17), ...[34823 more]

Figure 2.1: Missing values

For the first and second type, though there are only 15 of the second type records, their missing attribute value has a strong relationship with country and place attributes, so we kept them. Because the number of 6236 cannot be ignored compared with the whole dataset, and their missed attributes are nominal type, we cannot simply remove or using mean/medium value to fill them(also consider the geographic location has relationship with country/place), so we applied the k-NN algorithm on both the first and second type record to impute the missing value, using the most similar case's (nearest neighbor) value to fill them.

Through trial and error, we found the geographic attribute has a strong relationship with county and place attributes, other attributes disturbed the distribution of imputed values due to the weak relationship, so just use these two fields as the feature to fill the missing value.

2.2 Noisy Value

Our team find there are only three countries after processed the inconsistent values: Spain, Portugal, and France, the main dataset is comprised of Spain and Portugal, there are only 135 records, compared with the whole 34,985 records, see figure 1.3. By comparing the amount of the data, almost all data point is distributed in Spain and Portugal later in the exploratory data analysis process. We think this 135 of data examples are noisy values, hence we removed them from our training data.

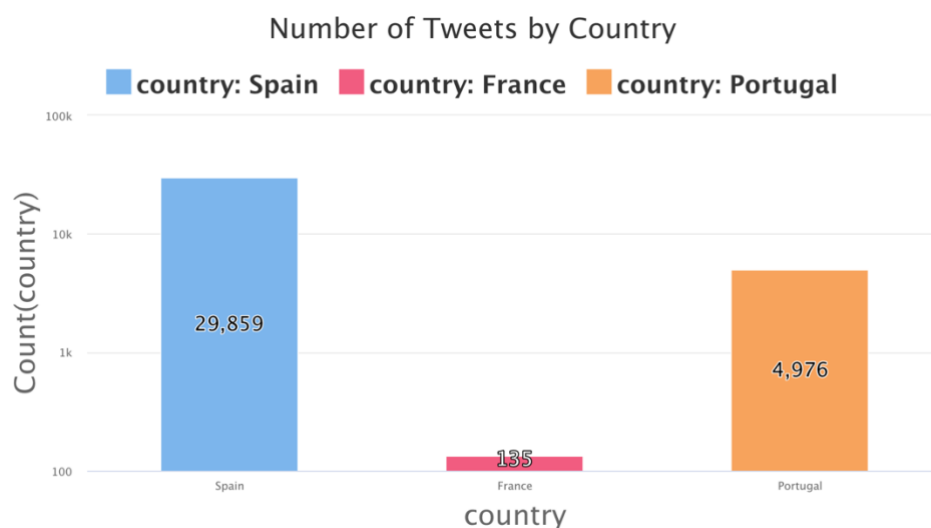


Figure 2.2: Number of tweets by country

2.3 Inconsistent Value

After data analyze, we found the country attribute's value consist of country names in many diffident languages, they simply just are three countries' names, so we deal with this name convenient inconsistency by translating all of them into English country name.

```
1 if(place_country=="España" || place_country=="Espainia" || place_country=="Espagne" || place_country=="Espanha" ||
2 place_country=="Espanya" || place_country=="Spagna" || place_country=="Spanje" || place_country=="Spanien", "Spain",
3 if(place_country=="Portogallo", "Portugal", if(place_country=="Francia", "France", place_country)))
```

Figure 2.3: Deal with inconsistent country names

2.4 Cleaning Methodology

We utilized the following process for cleaning the missing value. In the “Impute Missing Value” process, we decided to use k-NN to fill in the missing value of geolocation (main), and country, place at the same time.

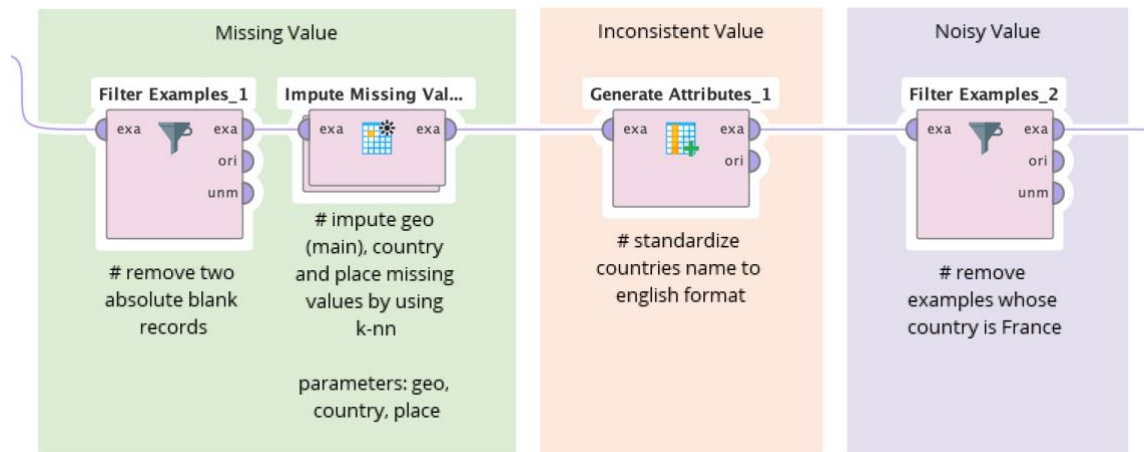


Figure 2.4: Data Cleaning

3. Data Integration, Transformation, Discretisation and Reduction

3.1 Data Integration

3.1.1 Schema integration

After the imported dataset, we noticed there are two data tables in the TweetData.xlsx, which have different attributes name convention, so before appended them into one table we unify the attributes name by integrating metadata from different sources into a formatted table. See *figure 1.4 Schema integration*.

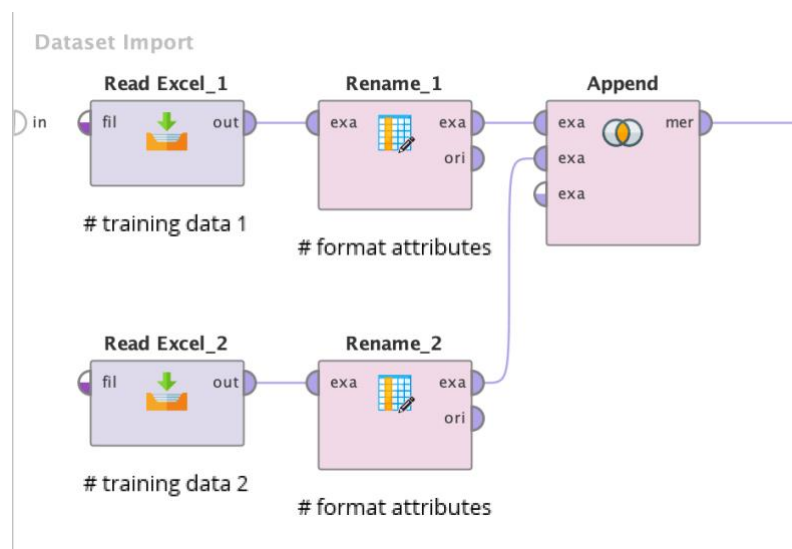


Figure 3.1: Schema integration

3.1.2 Redundancy Handling

There are two pair of records have the same tweet id with completely different information that we found after appended, so we used a new id identifier instead.

3.2 Data Transformation

3.2.1 Coordinates_geo Attribute Transformation

The `coordinates_geo` is a ploy-nominal type attribute, combining the latitude and longitude information in a string format, we split it into latitude and longitude field and converted them into numerical type for better geographical visualization.

3.2.2 Language Attribute Transformation

There are three kinds of value in language attribute for example:

en: means the tweet content is a single language.

en/es: means the labeler could not determine which language type should be but in a specific range.

en+es: means there is more than one language used in the tweet text.

Because there is an uncertainty in the en/es language type, we thought it would be better using the most likely case's language tag to replace instead of using an ambiguous value range, so in this process, we remove the value contains "/" and imputed using the relevant feature fields (for example country, use id, latitude, longitude).

3.2.3 Text Attribute Transformation

The tweet content contains many unimportant or useless information such as @/mention, #/topic, and https:/link, for a better text mining result it is better to remove these unnecessary first, we got a clean text by dropping these noisy contents.

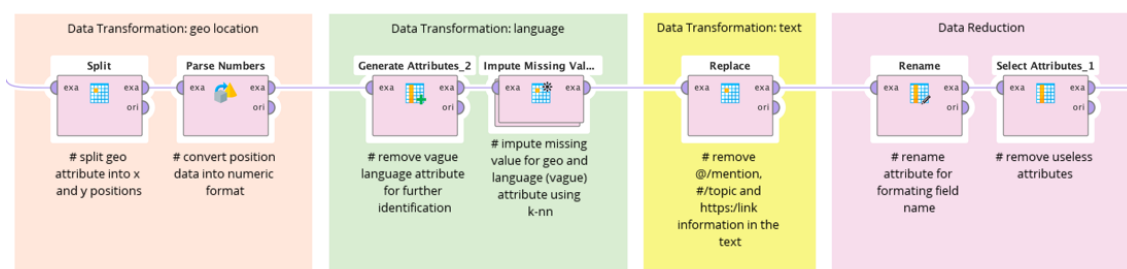


Figure 3.2: Data Transformation

3.2.4 Aggregation, Normalisation and Generalisation

For data aggregation, normalization and generalization, there are only two numerical type attributes after the above processes: id, latitude, and longitude. For id obviously, there is no need to do any further process, for latitude and longitude we think the original value can reflect the distribution of data points geographically, normalization may be not a good choice for visualization and accuracy.

Most of the attributes in the table are non-numerical type, this characterizes of dataset determines that it is not suitable for data aggregation, normalization or generalization.

3.3 Data Discretisation

As except the id attribute is numerical type, all other fields are categorical data types, so there is no need to do the data discretization process.

3.4 Data Reduction

In this process, we made data dimensionality reduction by dropping some unrelated attributes such as tweet id for more accurate feature value.

4. Feature Design and Selection

4.1 Create Feature Bilingual

After visualized the processed data, we found the tweets containing more than one language have a distribution trend associated with geographic location, obviously, there are more bilinguals than other areas in two of Spain areas, we think there is a strong relationship between the bilingual and specific area (there are four cluster area together), which means the tweet sent in one specific place is more likely to not be a single language content if a tweet is written in single language or not can imply some geographic distribution vice versa, so we used this created feature with other related features to better predict the target language label. See *Figure 1.6*.

4.2 Create Feature Area

Considering most of the fields are nominal type and data is collected in the form of different countries, using a distance-based cluster algorithm is not a good choice so we chose the DBSCAN algorithm which is based on density to cluster the data points. See *Figure 1.6*.

After visualizing the above-clustered data, the distribution of data points clearly in four areas (three areas belong to Spain), again, we found several specific languages have a very strong relationship with specific areas, for example, es is mainly used in Spain and pt is popular in Portugal, which is obvious, also "ca" and "gl" are used commonly in two different Spain areas, which reveals certain areas tend to have one or more specific language commonly used, it can be a very important feature to predict which language tag that a tweet should be labeled.

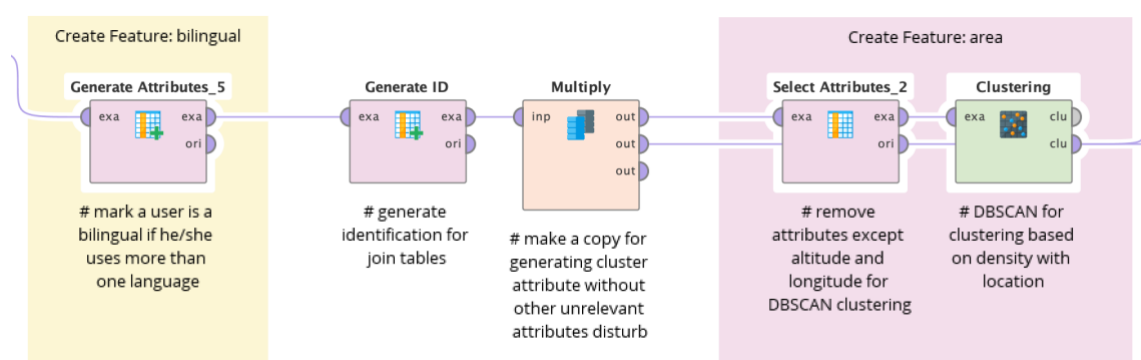


Figure 4.1: Feature Design and Selection

This area feature created is a little bit similar with the latitude, longitude, and country, but it offers another dimension to classify language according to block belonged to instead of the nearest data point location, also within the same country (Spain) there are still three different

blocks which have different language tendency, so area could be a necessary feature and it contributes a more accurate result later we found.

4.3 Feature Selection

After data analyze, we found the language has a relationship with the following feature attributes:

- user: a user more likely use the same language to write a tweet (not always but almost)
- bilingual: certain areas have much more bilinguals than others
- area: different location areas have different language using trend
- latitude and longitude: location affects the distribution of language
- country: two different countries (Spain and Portugal) have different commonly used language range
- place: In regards to k-NN, if the two attributes are closely related to each other, it will affect the performance of the model. As the "place" attribute is highly related to the "geolocation" attribute. We decided to remove the "place" attribute to improve the performance of k-NN.

5. Exploratory Data analysis

5.1 Data Structure

The TweetData.xlsx tweets dataset given with manual annotation of language, attributes including tweet_id, coordinates_type, coordinates_geo, place.country, place.name, user_id, tweet_language, and tweet_text.

Almost of dataset attributes are non-numerical type, which means the country, content, language and other attributes are based on the text information, the format of attributes' values is mainly string type, nominal/ordinal data process is one of the major tasks.

Distributions: the distribution of data is neither normal nor skewed, it is collected based on the location where the tweet sent, in other words, the data distribution represents a geographic visualization based on latitude and longitude.

Data quality problems: the missing values in coordinates_geo cannot be ignored (6235 records are missing), some values that are a range in language field are ambiguous

Outliers: 135 records that are collected from France.

Correlations and inter-relationship: language attributes have relationships with other fields more or less, also place, there also is a strong relevance among country and geographic location.

Subsets of interest: All attributes except for the tweet_id and coordinates_type.

5.2 Attribute/Observation Characters

tweet_id: numerical, the identifier of the tweet records, having 4 examples of tweet_id conflicts after appended two tables.

coordinates_type: nominal, 6236 records miss value, all other value in coordinates_type attribute is "point".

coordinates_geo: nominal, 6236 records miss value, combining latitude and longitude value into a string format.

place.country: nominal, 15 records miss value, there are inconsistent values in this field, that means the same country name is written in different languages.

place.name: nominal, 15 records miss value, have a potential relationship with language.

user_id: nominal, there are several records belong to the same user.

tweet_language: nominal, there are three types of value generally in language attribute, one single language (es), multiply languages (es+en) and an unknown specific range (es/en).

tweet_text: nominal, contain some useless or noisy information such as @/mention, #/topic and https:/link content.

5.3 Key Finding

1. The data collected have an obvious geographic distribution tendency. After transformed coordinates_geo attribute if we use latitude and longitude as x and y position, there is a clear visualization of 2 countries' (Spain and Portugal) tweet data including 4 distinctive areas (3 of 4 belong to Spain).

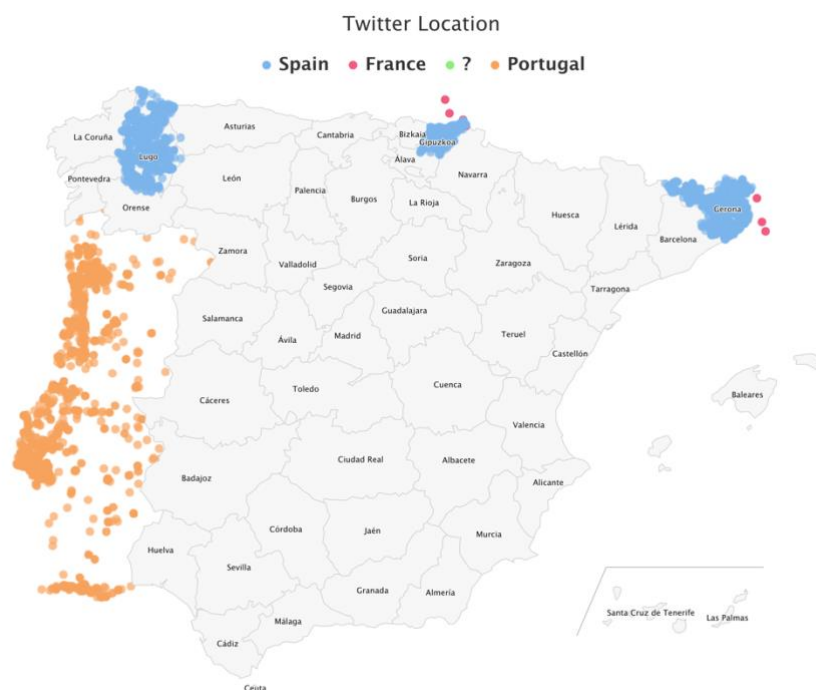


Figure 5.1: The map of tweets location

2. The geographic distribution tendency in a zoom-in level shows the popular location that the user sends a tweet, it almost like a traffic map, but instead of the traffic of vehicles, it shows the traffic of the tweets that sent in a certain location by the users.

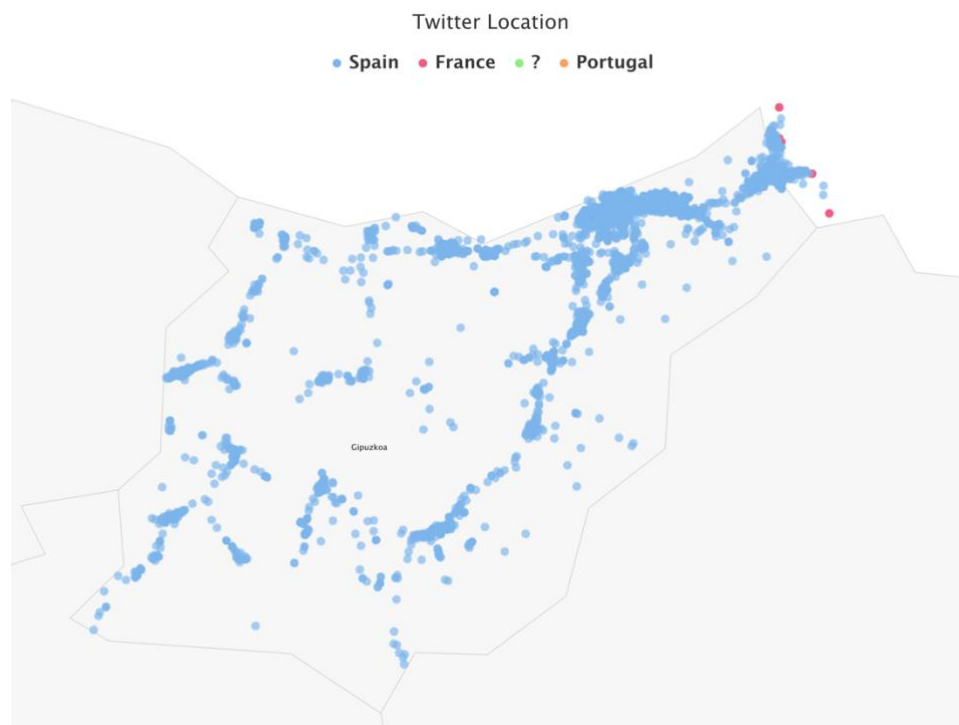


Figure 5.3: Zoom-in tweets location in city Gipuzkoa

3. The data are given shows Spanish language dominance among all kinds of languages. After applying the aggregate function with language attribute as x value and count as function, the distribution shows a unimodal shape, which has a single prominent peak in es. Overall the language used in the tweet language in the dataset is shown in the following figure:

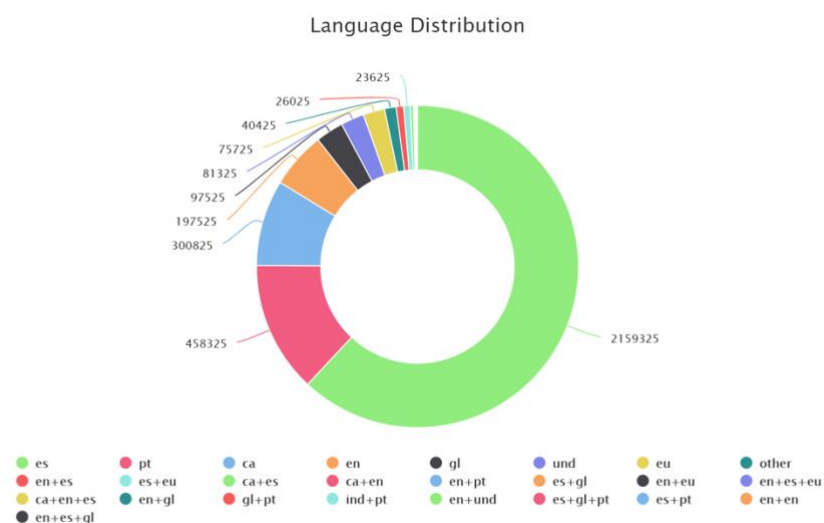


Figure 5.3: Tweets language distribution

4. The tweet's language distribution is strongly related to location information (geographic position, country, and place). We not only found different countries have their dominant language type, but also even different blocks in the same country (Spain) could represent the different languages using preference.

6. Descriptive Data Mining

Our descriptive problem is what are the popular languages that commonly used in a specific area or block, that means if want to find the language distribution the first thing that we need to do is find the appropriate clusters of the data points, then visualizing the dominant language type according to specific cluster that is based on some criteria.

6.1 K-means algorithm

Firstly, we tried the k-means algorithm for clustering the data given, after visualizing the result of we found the shape of data points have a geographic distribution trend, the data points fell into four main distinctive blocks which represent a density-based distribution.

As the k-means is a clustering algorithm that is based on the distance between data points, even though alter we modified to 4 for the k value (the number of clusters) the visualization result is still not very satisfied, as the k-means algorithm is not suitable for the arbitrary shape of data distribution.

Also considering most of the fields are nominal type and data is collected in the form of different countries, using a distance-based cluster algorithm is not a good choice so we chose the DBSCAN algorithm which is based on density to cluster the data points.

6.2 DBSCAN algorithm

Finally, we replaced the k-means with the DBSCAN algorithm, which is a density-based cluster algorithm that can be used to cluster the specific shape dataset, obviously, it is suitable for our dataset with a specific shape related to geographic location.

At first, we used the parameters 0.3 as epsilon and 5 as min points and then got a poor result which displayed many clusters more than four, then we calculated the distance of any members between or within different areas and found 0.3 is too small so more than expected clusters are created, the minimal value of the epsilon should be larger than the maximum distance between any pairs of data points within the same block, as the same time, the maximum epsilon value should be smaller than the minimal distance between any members between different clusters, finally we chose the 0.5 for epsilon value, 2 min points and got a much better visualization result according to the data shape distribution.

6.3 Dataset Grouping Result

With our descriptive data mining representation, we used the language label as x value and a count function for the value in the coordinate system, also using the color for different areas.

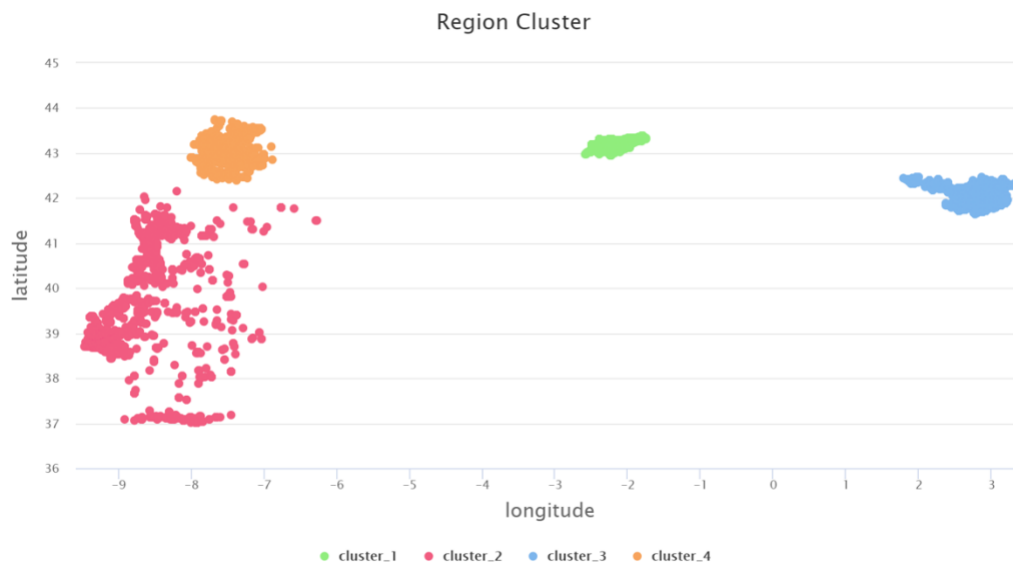


Figure 6.1: Four Region Cluster

By comparing to the figure, the four cluster regions are Cluster 1: Gipuzkoa (Green), Cluster 2: Portugal (Pink), Cluster 3: Gerona (Blue) and Cluster 4: Lugo (Orange).

When we color the map by the language used in the text of the tweet. We can see the language distribution in each region in the figure below.

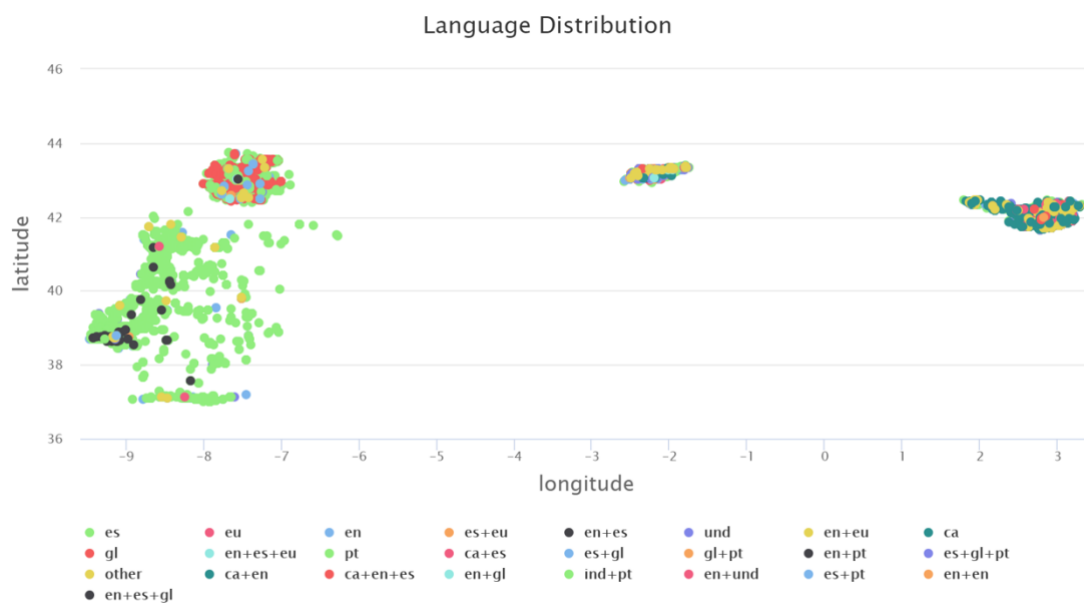


Figure 6.2: Language Distribution Map

After the data processing, it is much easier to notice that different areas have one or more dominant languages as the following pictures showing:

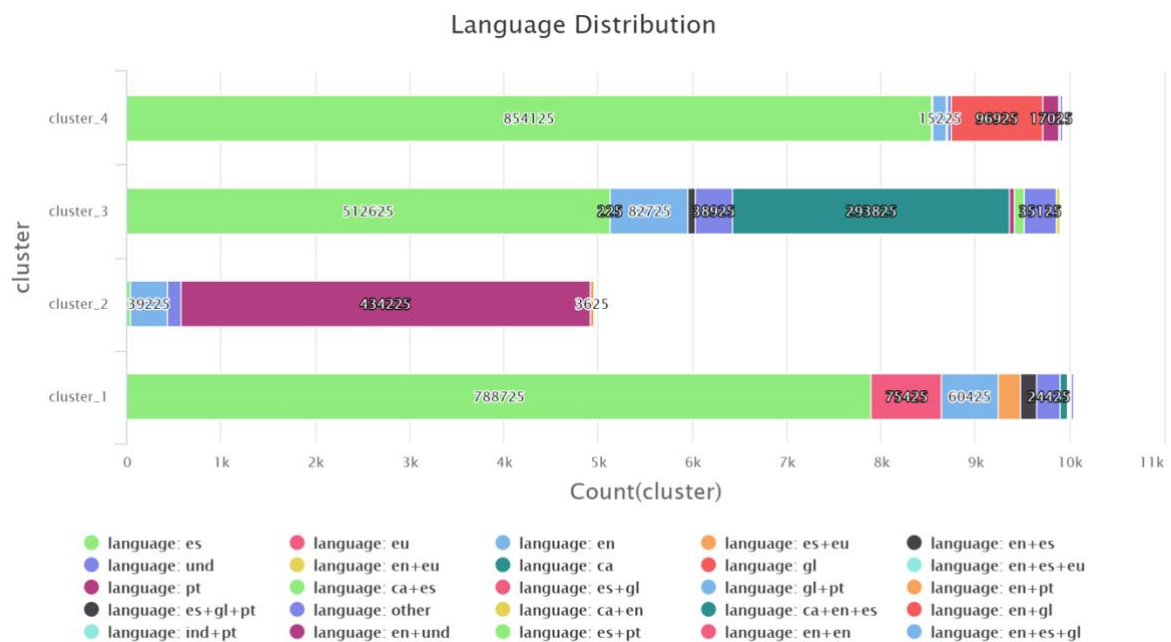


Figure 6.3: Language Distribution in Each Region

For Cluster 1: Gipuzkoa and Cluster 4: Lugo, the domain language is "es" (Spanish) and has been reached to 80% and 86% accordingly.

For Cluster 3: Gerona, the domain language is still "es" (Spanish) which has 52% and an interesting finding from the data is that "ca" (Catalan) is the second domain language in Gerona region, which has taken 30% among all the other language.

For Cluster 2: Portugal, the domain language is "pt" (Portuguese) which makes sense as it is in the country of Portugal.

7. Predictive Data Mining

7.1 Naïve Bayes Algorithm

In the predictive data mining analysis process, before any implementations of the algorithm to classify we analyzed several characters of the algorithm, in general, to guess whether an algorithm is suitable for our dataset or not.

Firstly, we thought the Naïve Bayes algorithms, as there are relationships between different attributes in the dataset given, and this algorithm has an assumption that all features are independent, which is not true in our case, so obviously Naïve Bayes is not a good choice for our classification task, we assumed it will perform very poor with a low accuracy and we can not use it, but for comparison, we still tried this algorithm in our data mining process to compare with other algorithms.

7.2 Decision Tree Algorithm

During the data analysis process, it can be easily noticed that some main factors are controlling which language type a tweet should belong to if we want to predict the language tag using other attributes, it is easy for people to think tweet sent in the specific country has a specific language label. Actually, the country is a very important factor when deciding language type, then in the same country there may still have some language distribution based on the place or city and so on, so we first thought the decision tree is naturally similar to the people decision-making process with this issue.

Reason for choice:

- Naturally closer to the human thought process compared with other types of algorithms, very easy to come up with.
- Basic logic pretty straightforward, which is easy for either explaining or interpreting
- Have a clear and vivid presentation for displaying the classifying process

7.3 Rule-Based Algorithm

After tried decision tree algorithm, the accuracy of the model is not that high and we found that predicting a tweet's language is not only based on a single factor at every prediction, usually multiply factors contribute to the final result, and sometimes there is not a hierarchical relation between different test nodes in decision tree. For example, sometimes there are no obvious classified clusters under a country node regarding the latitude or longitude, which means the language type is decided mutually by location and country

information, so we thought beyond decision tree the rule-based algorithm would be a better choice.

As we have a relatively large dataset (34,985 records), processing and classifying new examples could be time-consuming and computing expensive for the machine learning algorithm, for example, artificial neural network, whereas with rule-based method a quicker classification process could be easily got.

7.4 K-NN Algorithm

After the above two types of algorithm trying, we found through the rule-based approach slightly improves the performance of the model, the accuracy is still not that satisfied, so we were thinking maybe we need to choose a more complex algorithm accomplished with more accurate result.

Although using K-NN algorithm takes more time than the others, we got a noticeable improvement in the pattern performance, this method uses different features to calculate the distance between data points and classify the tweet's language label according to its nearest neighbors, taking accounting into many factors in determining the language tag.

8. Evaluation and analysis of results

8.1 Parameter Optimising

For the predictive problem, in this classification issue which belong to supervised learning we tried the following multiple parameters to find the best performance model for each of three machine learning algorithms:

Naive Bayes: laplace_correction: true/false

Decision Tree Algorithm

Criterion: gain_ratio, information_gain, gini_index, accuracy

Confidence: Min 0.1, Max 0.5, Steps 5

Maximal_depth: Min 1, Max 10, Steps 10

Minimal_leaf_size: Min 1, Max 10, Steps 10

Minimal_size_for_split: Min 1, Max 10, Steps 10

Rule Based Algorithm

Criterion: gain_ratio, information_gain, gini_index, accuracy

Confidence: Min 0.1, Max 0.5, Steps 5

Maximal_depth: Min 1, Max 10, Steps 10

K-NN Algorithm

K value: Min 1, Max 20, Steps 20

We used the Optimize parameter Operator to run the above parameter ranges for each algorithm and find these optimized parameters with the highest accuracy:

Naive Bayes: laplace_correction: true

Decision Tree Algorithm: Criterion: information_gain, Confidence: 5.0, Maximal_depth: 10
Minimal_leaf_size: 2, Minimal_size_for_split: 5

Rule Based Algorithm: Criterion: information_gain, Confidence: 4.00000002, Maximal_depth: 10

K-NN Algorithm: K value: 12

8.2 Performance Measure

8.2.1 Performance Measure

In supervised learning, there are more specific performance or accuracy measures than unsupervised learning as follows:

Predictive accuracy: correctly predict the class label of new or previously unseen data

Speed: computation costs involved in generating and using the model.

Robustness: make correct predictions given noisy data or data with missing values

Scalability: construct the model efficiently given a large amount of data.

Interpretability: level of understanding and insight that is provided by the model

Simplicity: decision tree size, rule compactness.

In RapidMiner, we used the following measures to evaluate whether our model is performing well for the predicting/classification task:

Accuracy: firstly we used the accuracy as the performance measure, which reached approximately 80%-90%, but later we found the distribution of the class in the dataset is not equal, most of (around 70%) language is es that dominates the value distribution, so accuracy is not a good measure in this case.

Weighted mean root precision: for the unequal distribution of class value, the weighted mean root precision is a more reliable measure for evaluation.

Weighted mean root recall: for the unequal distribution of class value, the weighted mean root recall is a more reliable measure for evaluation.

Absolute error: we also used absolute error to measure the model's performance, to weigh how much the difference between the actual and predicted value.

Relative error: compared with the absolute error, the relative error is a better way to compare the pattern performance among different algorithms.

The classification performance of our final trained model with different measures as follows:

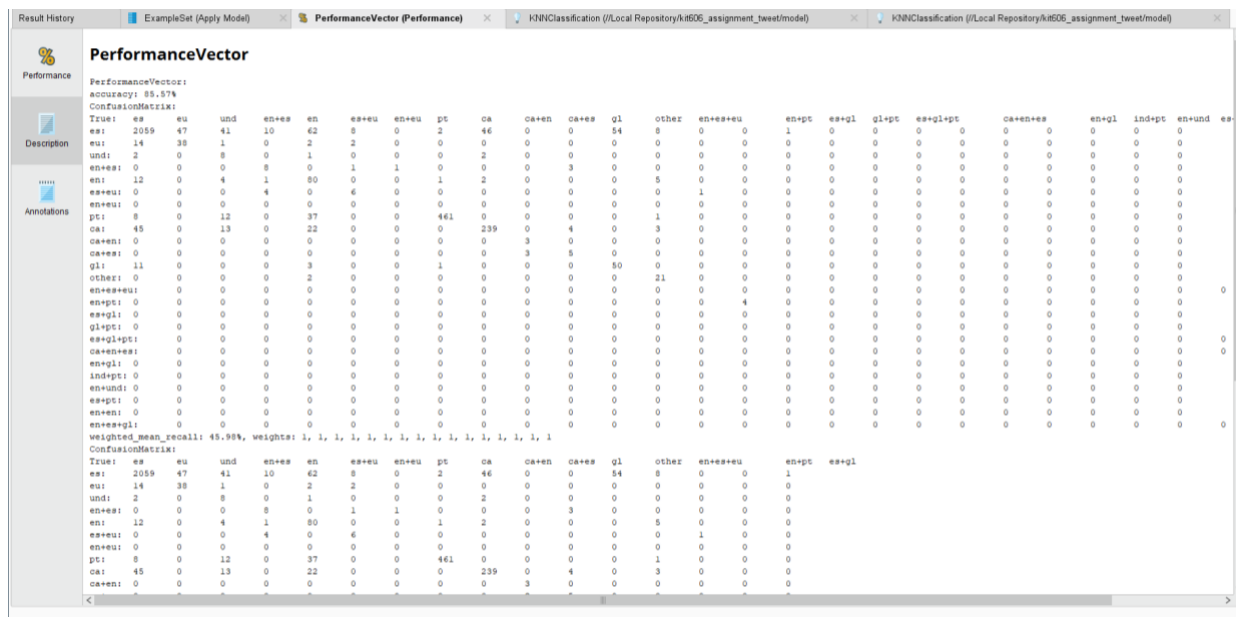
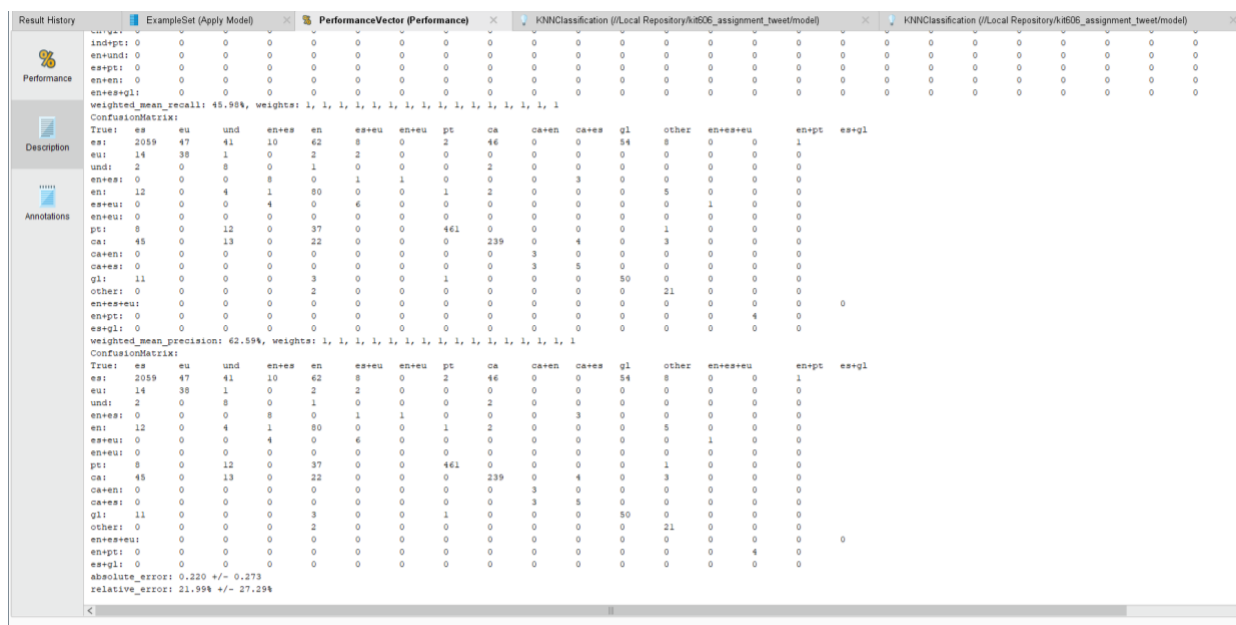


Figure 8.1: K-NN Performance Vector_1



Rule Based Model Performance Vector:

accuracy: 75.70%
weighted_mean_recall: 24.01%
weighted_mean_precision: 20.95%
absolute_error: 0.408 +/- 0.001
relative_error: 40.80% +/- 0.11%

Decision Tree Model Performance Vector:

accuracy: 75.79%
weighted_mean_recall: 24.43%
weighted_mean_precision: 26.96%
absolute_error: 0.372 +/- 0.013
relative_error: 37.18% +/- 1.34%

Naïve Bayes Model Performance Vector:

accuracy: 64.63%
weighted_mean_recall: 37.47%
weighted_mean_precision: 30.23%
absolute_error: 0.369 +/- 0.006
relative_error: 36.94% +/- 0.65%

8.2.1 Attempt to improve accuracy

In the accuracy improving stage, we found using the place as the feature will degrade the performance of the model, as the k-NN algorithm is easily misled by unrelative attributes, after we dropped place fields there is a slight improvement in the accuracy.

8.3 Evaluation Methodology

In the model evaluation process, we chose the 10-fold cross-validation, which is a better way to avoid the overlapping test data, before we finally decided to use 10-fold cross-validation we also tried some other evaluation methods for example hold-out approach with partition split, whereas these methods did not achieve a good result like 10-fold cross-validation which makes the best use of data to evaluate.

8.4 Analysis of Results

In our tweet data analysis assignment, we came up two questions based on the dataset given, one of them is a descriptive type: what are the popular languages that commonly used in a

specific area or region, and for predictive problem, we wanted to know how to determine the language type of tweets according to tweet's other information without human labor labeling?

For descriptive problem, we processed the data mining (clustering) to visualize different areas' language distribution feature, we got a series of pictures of characters of language distribution based on the distinctive blocks with different dimensions, basing on these plots we could much easily to find which language type is more popular or commonly used in a specific area.

For predictive problem, we did the data preparation and data mining (classification) process to train the model based on the training data given, using the several performance measures and evaluation methods to determine if the detected pattern is good or not. After many tries, though the accuracy of trained model reached to around 85%, the weighted mean root recall and precious is not that high as accuracy, only around 45.9%, and 62.59%, we assume the cause is characters of dataset given which does not have equal class value distribution, but we tried to control the relative error to a minimal range which is 21.99% +/- 27.29%.

9. Others

9.1 Lessons Learned

We learned a wide range of things from the exploratory data analysis to some machine learning algorithms for classification/prediction and clustering, also the data preparation process and practical exercise with RapidMiner software.

The data preparation is a very significant stage during the whole data analysis process, the quality of processed data directly determines the quality of data mine result. There are several aspects to perform this task such as data cleaning, transformation, discretization and reduction, especially data cleaning plays a crucial role during this process.

We also learned several ways to visualize the dataset for exploratory data analysis, different types of graphs such as scatter plot, bar plot, and histogram, there are lots of characters of data structures associated with different data distributions.

For the prediction task, we mainly focused on the some commonly used clustering and classification algorithms, for example, the Naïve Bayes or K-nearest neighbor algorithm for supervised learning (classification task), and k-means or DBSCAN for unsupervised learning (clustering task), we studied the basic logic and examples for each of them.

We also got some real experience with the exercise from tutorial and assignment except for the theory learned in the lecture, which helped us to fully understand some theory-based concepts in these practical activities.

9.2 Tools Utilised

RapidMiner: data preparation; machine learning; text mining; predictive analytics.

Python: using numpy and pandas module to make some text processes.

9.3 Team Contribution

	Contribution	Main Contribution and activities in the project
Name: Rui Lyu ID: 509141	50%	<ul style="list-style-type: none"> • Discussed the data structure in the group meeting. • Brainstormed on the data questions and products. • Project proposal: Created the project proposal document ahead of the presentation document. • Analyzed the dataset given and discussed the data structure and process. Analyzed the missing value, inconsistent and noisy data. • Wrote the draft of the report with framework and structure. • Data Product: Designed the RapidMiner data product. • Model Training and Selecting: Trained the final model and made some improvements to the accuracy of the model.
Name: Xiating Cai ID: 521285	50%	<ul style="list-style-type: none"> • Discussed the data structure; Brainstormed on the data questions and products. • Project Presentation: Organized the presentation powerpoint, represented the team in the project early stage presentation. • Project Proposal editing, improving and proofreading. Project report finalization. • Data Cleaning: Worked on the inconsistent and noisy value in RapidMiner. • Data Transformation: Worked on data transformation such as geolocation data split, cleaning tweet text @mention, #topic, https:// link, etc in the tweet text. • Data Visualization: Selecting and fitting the best visualization tools to represent the data analysis result.

Reference

Sayce, D 2019, *The Number of tweets per day in 2019* | David Sayce, viewed 9 December 2019, <<https://www.dsayce.com/social-media/tweets-day/>>.