Milestone 4: Sentiment analysis of Wikipedia articles

Jinghong Zeng
276425917, jzen696@aucklanduni.ac.nz

Goal

The goal is to analyze the sentiment in English Wikipedia articles. Ideas are to find the sentiment in each article, how the sentiment is associated with the topic, and how the sentiment is associated with the article popularity.

Data source

Data of English Wikipedia articles comes from English Wikipedia API. In RStudio Cloud, `GET` requests from the httr package will be made to access data in JSON format from the API. Information that I want to know about the articles are nominated as the API parameters in the URL, with reference to MediaWiki API documentation.

9 main categories are selected from the main categories of English Wikipedia, except of two main categories, Human Activities and General Reference. Across the 9 categories, 22 topics are selected, with a comparable number of articles, to achieve a more representative sample of all the English Wikipedia articles and help us identify a broader scope of topics, though this sampling method cannot guarantee all the articles have the same amount of text. In total 1740 articles are extracted. Table 1 is created to summarize the topics and numbers of articles.

Topic	Category	Number of articles	Topic	Category	Number of articles
Literature	Culture	78	Spiritualism	Religion	42
Language	Culture	56	God	Religion	11
Hospitals	Health	46	Politics	Society	82
Safety	Health	144	Motivation	People	111
Diseases	Health	51	Love	People	106
Historiography	History	154	Hobbies	People	62
Measurement	Mathematics	176	Entertainment	People	52
Proofs	Mathematics	40	Sexuality	People	40
Theorems	Mathematics	18	Robotics	Technology	128
Ethics	Philosophy	144	Manufacturing	Technology	85
Buddhas	Religion	48	Software	Technology	66

Table 1: Summary of topics

Data processing

The data is sourced via API for each topic separately. It is in JSON format and then converted to `list` objects by the `fromJSON` function. The output is the converted data for the topic Literature, which is a nested list that contains two lists. Ideally there should be a third list named `batchcomplete` to indicate if the data from the current brunch have been returned completely. This list is missing possibly because some data are missing.

Then I used functions including `map`, `mutate`, `left_join` to extract variables from a sublist into a list for each topic and create the number of logged-in contributors, redirects and pageviews from original

information, then bind all the data sets. Next the release date of the first revision is sourced independently and added to the data.

The complete data has 25 columns and 1740 rows. Summary of 4 attributes is shown, including the numbers of logged-in contributors, pageviews, visiting watchers and watchers. These attributes have more than 50% missingness, 3 of which have more than 90% missingness.

```
 numcontributor     pageview      visitingwatchers    watchers
 Min.   : 15.0   Min.   :    10   Min.   : 30.00   Min.   :  30.0
 1st Qu.:108.0   1st Qu.:   455   1st Qu.: 37.00   1st Qu.:  48.0
 Median :168.0   Median :  2503   Median : 44.00   Median :  95.0
 Mean   :244.4   Mean   : 15542   Mean   : 54.23   Mean   : 174.2
 3rd Qu.:388.0   3rd Qu.: 10169   3rd Qu.: 58.00   3rd Qu.: 195.0
 Max.   :500.0   Max.   :849483   Max.   :181.00   Max.   :2339.0
 NA's   :1695    NA's   :1627     NA's   :1625     NA's   :1022
```

The data contains missing values in the number of contributors, watchers and redirects. It is implausible to impute missing values by 0 because the `batchcomplete` list has not been returned and the total number of contributors can be 0 for some pages if I impute missing values by 0. Instead, missing values are imputed by the `mice` function from the `mice` package, using the random forest method. Summary of 4 imputed attributes is shown:

```
 numcontributor     pageview      visitingwatchers    watchers
 Min.   : 15.0   Min.   :    10   Min.   : 30.00   Min.   :  30.0
 1st Qu.: 63.0   1st Qu.:   438   1st Qu.: 33.00   1st Qu.:  36.0
 Median : 99.0   Median :  2503   Median : 38.00   Median :  49.0
 Mean   :126.1   Mean   : 13815   Mean   : 41.31   Mean   : 102.7
 3rd Qu.:143.2   3rd Qu.:  7935   3rd Qu.: 47.00   3rd Qu.:  97.0
 Max.   :500.0   Max.   :849483   Max.   :181.00   Max.   :2339.0
```

Title and content are combined together as text. To preprocess text, I first use the `gsub` function to clean hidden information, file uploads, references, URL and other non-useful structures out of text. My cleaning function may lead to some mistakes, for example some non-useful characters that I do not know about may not be cleaned out, but all the sentimental words should have been retained. An example text extracted from the article "Proverb" is:

```
[1] "Proverb {{short description|Short traditional saying that expresses a perceived"
[2] "truth}} '''proverb''' (from {{lang-la|proverbium}}) is a simple and insightful,"
[3] "traditional [[saying]] that expresses a perceived truth based on common sense"
[4] "or experience."
```

After cleaning, this text becomes:

```
[1] "Proverb '''proverb''' ( ) is a simple and insightful, traditional <<saying>>"
[2] "that expresses a perceived truth based on common sense or experience."
```

Then I use Spark NLP annotators to further clean the text. The process is to first split the text into tokens (words), normalize the tokens (remove dirty characters) and lemmatize the tokens (find the root words) based on a Spark lemma dictionary. The cleaned text now becomes:

```
 [1] "proverb"     "proverb"     "be"          "a"           "simple"
 [6] "and"         "insightful"  "traditional" "say"         "that"
[11] "express"     "a"           "perceive"    "truth"       "base"
[16] "on"          "common"      "sense"       "or"          "experience"
```

Analytical plan

Analyze the sentiment of the articles

To find the sentiment of each article, I plan to use the Sentiment Detector from `SparkNLP`. My approach is to score the sentiment of the articles based on a user-defined sentiment lexicon via the Sentiment Detector and then classify the articles into 3 sentiment classes based on the calculated scores. The sentiment is positive if the score is positive, neutral if the score is 0, otherwise negative. This approach is suitable when the actual sentiment label for each article is unknown.

Analyze the association between the sentiment and the topic

I plan to use multinomial logistic models and decision trees from `SparkML`, with the topic as the explanatory categorical variable and the sentiment as the response variable. These models can classify the response variable into positive, neutral and negative sentiment classes. It is easy to obtain interpretable results such as the probabilities of three sentiment classes from both models. The most likely sentiment and the sentiment probabilities associated with each topic will then be computed from the models and compared to understand the association between the sentiment and the topic.

Analyze the association between the sentiment and the article popularity

I plan to use multinomial logistic models and decision trees from `SparkML`, with the sentiment as the response variable. The explanatory variables are all the popularity-related attributes, including the numbers of contributors, watchers and redirects. The multinomial logistic models can produce interpretable results such as coefficients to show the association, and they can also include the interaction terms easily when the interaction between some attributes is considered. The decision trees are extra interesting applications to see if they can provide additional information from the tree structure.

Results

Analyze the sentiment of the articles

The base lexicon is NRC Word-Emotion Association Lexicon, with more than 10 thousand common English words and their associations with eight basic emotions and two sentiments (negative and positive). The sentiments of each word is extracted from the NRC lexicon, and I add some words to increase, decrease or revert the sentiment. For example, the word "very" can increase the positivity of the word "good". Some comparative words are also added. In addition, root words of NRC words are added and words related to both positive and negative sentiments are excluded.

To prove applicability of the modified NRC lexicon, I check the most common words in all the articles. These words contain few sentiment words, however, from which the relevant sentiment words ("not", "more", "no", "most", "love") have the correct sentiment in the modified NRC lexicon.

```
                1      2         3         4      5
word         love    not      more      most     no
sentiment positive revert increment increment revert
```

Then I add the Sentiment Detector into the same pipeline trained for text preprocessing in `SparkNLP` before the final step and obtain the sentiment scores for all the articles. Finally I classify the articles into 3 sentiment classes based on the calculated scores. The model output for 2 articles is shown below.

```
  score sentiment                                                text
1    50  positive Stylistics       '''Stylistics''', a branch of  style>
2    -1  negative Imagery   '''Imagery''' is  Phenomenological>>'', perta
```

502 articles are classified as negative, 540 articles as neutral, and 698 articles as positive. The three proportions are quite comparable, where positive articles tend to be more. This may indicate that the Wikipedia articles do not have a tendency towards a single sentiment.

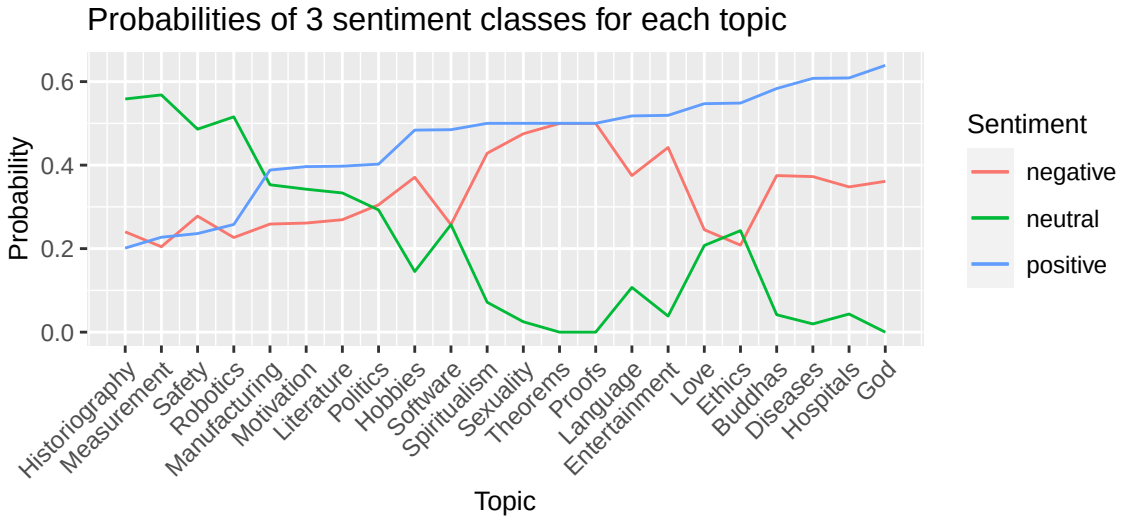
Analyze the association between the sentiment and the topic

An explanatory multinomial logistic model is fitted. Overall, all the topics are as either neutral or positive from the model. Table 2 shows the topics and their most likely sentiment class. Most topics are associated with the positive sentiment, which indicates that the overall sentiment in the Wikipedia articles tends to be positive.

Sentiment	Topic
Neutral	Historiography, Measurement, Safety, Robotics
Positive	Manufacturing, Motivation, Literature, Politics, Hobbies, Software, Spiritualism, Sexuality, Theorems, Proofs, Language, Entertainment, Love, Ethics, Buddhas, Diseases, Hospitals, God

Table 2: Summary of the most likely sentiment associated with each topic

To understand how each sentiment class is associated with the topic, a graph is plotted below to show the association between the probability of each sentiment class and each topic.



The topics are ordered in the ascending positive sentiment probability. The neutral sentiment probability has a downward trend and the negative sentiment probability has an upward trend. Except the first 4 topics, from "Historiography" to "Robotics", the positive sentiment probability dominates. Table 3 summarizes the results from the graph. It is plausible that some topics are associated with both positive and negative sentiments. For example, "Hobbies" can include both good and bad hobbies, associated with opposite sentiments.

It is surprising that "Theorems" and "Proofs" have 0 probability to be neutral but an equal probability to be positive and negative. From the sample, the two topics only have an equal number of positive and negative articles. Hence, "Theorems" and "Proofs" can involve sentiments. On the other hand, the two topics can be viewed as no sentiment overall, because the positive and negative sentiment cancel out each other. This interpretation may suit the common sense more.

Accuracy is defined as the percentage of the correct predictions in the sample. The multinomial logistic model has about 50% accuracy. The accuracy is not high, possibly because some attributes that can explain more variability in the sentiment are not included in the model, but the topic is still a useful explanatory variable.

Next, an explanatory decision tree with depth of 19 is fitted through `SparkML`. 22 topics are converted into 21 dummy variables, with the topic "God" as the reference level. Depth of 22 is chosen because I want all the topics to be considered by the decision tree and also want to know which topics are most

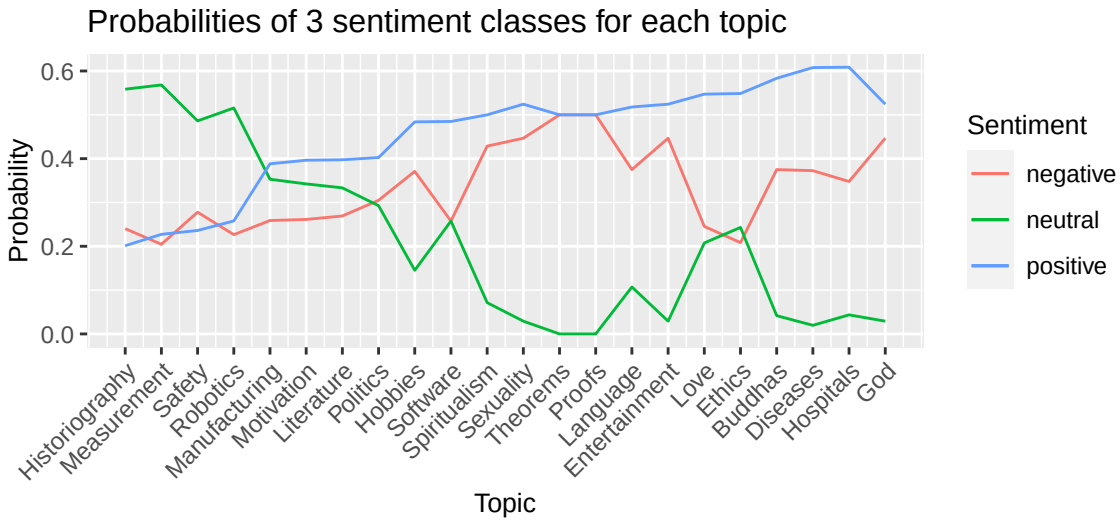
Topic	Sentiment probability	Association
Historiography, Measurement, Safety, Robotics	Neutral sentiment probability is highest, the other two probabilities are similarly low	Mainly associated with neutral sentiment
Manufacturing, Motivation, Literature	Higher probability for neutral and positive sentiments	More associated with positive and neutral sentiments
Politics	Similar probability for negative and neutral sentiments that is comparable to positive sentiment probability	Associated with all the sentiments
Software, Love, Ethics	Similarly low probability for neutral and negative sentiments	Mainly associated with positive sentiment
Hobbies, Spiritualism, Sexuality, Theorems, Proofs, Language, Entertainment, Buddhas, Diseases, Hospitals, God	Very low probability for neutral sentiment, high probability for positive and negative sentiments	Associated with both positive and negative sentiments, tend to be more sentimental

Table 3: Summary of association between the sentiment and the topic

important. The 4 most important features are the dummy variables for the topics "Historiography", "Measurement", "Safety", "Robotics". They are the top 4 nodes in the tree structure. The other dummy variables contribute little to the decision tree. The dummy variables for the topics "Entertainment" and "Sexuality" are even not used.

Except that the topics "Theorems" and "Proofs" are classified to be negative, the most likely sentiment associated with the other topics from the decision tree are the same as from the multinomial logistic model. All the 3 sentiment probabilities for "Theorems" and "Proofs" are identical in both the decision tree and the multinomial logistic model, so the difference in the classified sentiments may just be the different model default.

A graph is plotted to show the association between the topic and the probability for each sentiment.



The probabilities produced by the decision tree are not exactly identical to the probabilities produced by the multinomial logistic model, but from the graph of the decision tree, the patterns of the probabilities for all the topics are nearly identical to the patterns in the graph of the multinomial logistic model. One significant difference is that the topic "God" from the decision tree has a similarly high probability for positive and negative sentiments, but this does not change the conclusion that "God" is associated with both positive and negative sentiments. The conclusions about the association between the sentiment and the topic are identical from the graphs of the decision tree and the multinomial logistic model.

Accuracy of the decision tree is about 51%. It is comparable to the accuracy of the multinomial logistic model. In conclusion, two models may both be suitable, and their results would be reasonably reliable. On the other hand, the decision tree can provide additional information that the 4 most important

topics are strongly associated with the neutral sentiment, and this association may be stronger than the association between the sentiment and the other topics.

Analyze the association between the sentiment and the article popularity

An explanatory multinomial logistic model is fitted initially on all the popularity-related attributes. These attributes are the numbers of logged-in contributors, anonymous contributors, watchers, visiting watchers, pageviews, redirects and years. Before imputation, the numbers of logged-in contributors, visiting watchers and pageviews have more than 90% missingness, and the number of watchers has about 59% missingness. Then a multinomial logistic model without attributes that have more than 90% missingness before imputation and another multinomial logistic model without attributes that have more than 50% missingness before imputation are also fitted, to compare with the full model. My idea is to see if the imputation can improve the model. Accuracy and F1-score are used as model comparison metrics and defined as in usual context. The comparison results are in table 4. These models are not predictive. However, if a model has good explanatory ability, it should fit the data relatively well.

Model	Accuracy (%)	F1-score (%)
Full	48.2	43.0
No attributes that have more than 90% missingness	48.4	43.6
No attributes that have more than 50% missingness	48.0	43.4

Table 4: Comparison of the models with or without attributes that have many missing values

The model without attributes that have more than 90% missingness is best. Excluding the number of watchers from the best model seems to worsen the model. In addition, the differences in two metrics are quite small among the models, which indicates the excluded attributes do not explain much variability of the response variable. Since the imputation on the attributes with more than 90% missingness do not improve the model, these attributes are excluded.

Then two multinomial logistic models with interaction terms are fitted. One model includes the interaction between the number of redirects and the numbers of anonymous contributors and watchers. The other model includes the interaction between years and the numbers of anonymous contributors, watchers and redirects. The comparison results are in table 5.

Interaction added	Accuracy (%)	F1-score (%)
With the number of redirects	48.3	44.6
With the years	48.3	43.8

Table 5: Comparison of the multinomial logistic models with interaction terms

Combining the two kinds of interaction in a single model helps slightly, but the increased model complexity makes the interpretation of the association harder. Hence, the two kinds of interaction are not used together. The model with interaction from the number of redirects gains more F1-score increase over tiny accuracy loss. Hence, it is chosen as the best model. Its coefficients are summarized in table 6. Contributors mean anonymous contributors. "A:B" means the interaction term between A and B.

Probability	Contributors	Watchers	Redirects	Contributors:Redirects	Watchers:Redirects	Years
Negative	0.000780	0.001561	0.027139	-0.000005	-0.000036	0.056484
Neutral	-0.001995	-0.001547	-0.048713	0.000040	0.000019	-0.076875
Positive	0.001216	-0.000013	0.021574	-0.000035	0.000017	0.020391

Table 6: Coefficients for each sentiment probability from the best model

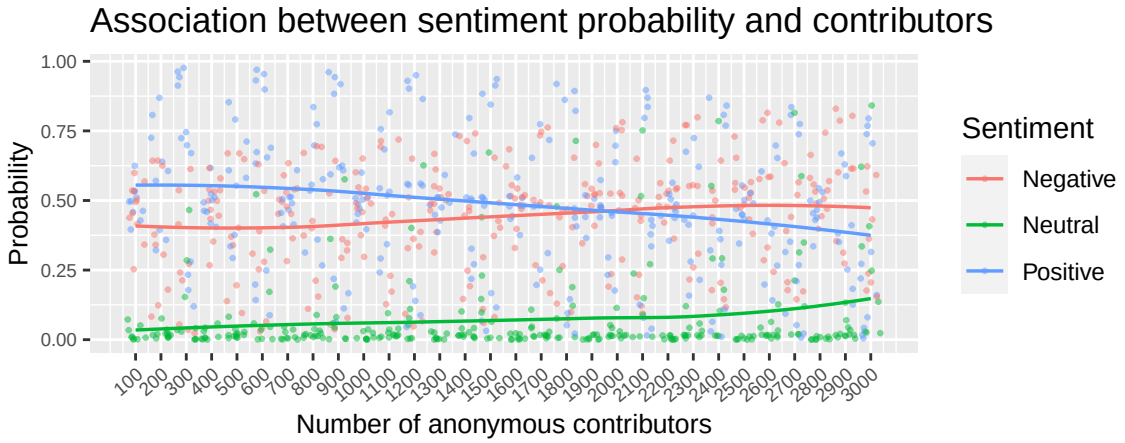
The direct association is the coefficient for a single attribute. The interaction association is the coefficient for an interaction term. The results for the direct associations are summarized in table 7. It is indicated that the article popularity is associated with non-neutral sentiments. Interaction between

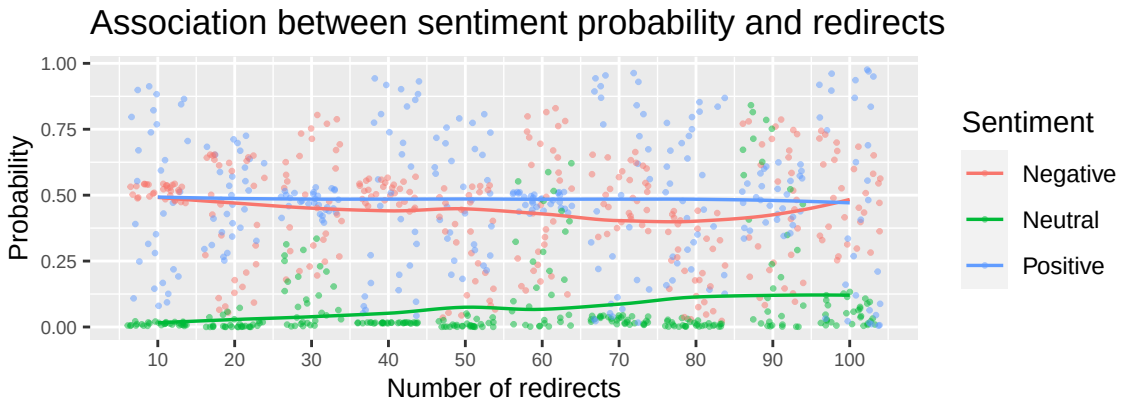
contributors and redirects is positively associated with neutral sentiment, while interaction between watchers and redirects is positively associated with non-negative sentiments. They are different from direct associations. Graphs are used to show the total association of the numbers of contributors, watchers and redirects with the sentiment.

Attribute	Association
Contributors number	Positively associated with non-neutral sentiments, more with positive sentiment
Watchers number	Positively associated with negative sentiment, negatively associated with positive and neutral sentiments
Redirects number	Similarly positively associated with non-neutral sentiments
Years	Positively associated with non-neutral sentiments, more with negative sentiment

Table 7: Summary of direct association between the sentiment and single attributes

Two graphs below show the association of the numbers of anonymous contributors and redirects with the sentiment. Predictions are made under the ranges of the two attributes in the data as shown in the graphs, while holding the other attributes fixed at a reasonable level. At each value of the number of anonymous contributors, 10 data points are plotted for each sentiment probabilities, corresponding to 10 values of the number of redirects. Then the overall trend in each sentiment probability is drawn, showing the average positions of 3 sentiment probabilities. This is similar to fixing the number of redirect at a level. But this approach can avoid the problem that the vertical position of each sentiment probability changes as the number of redirects changes.



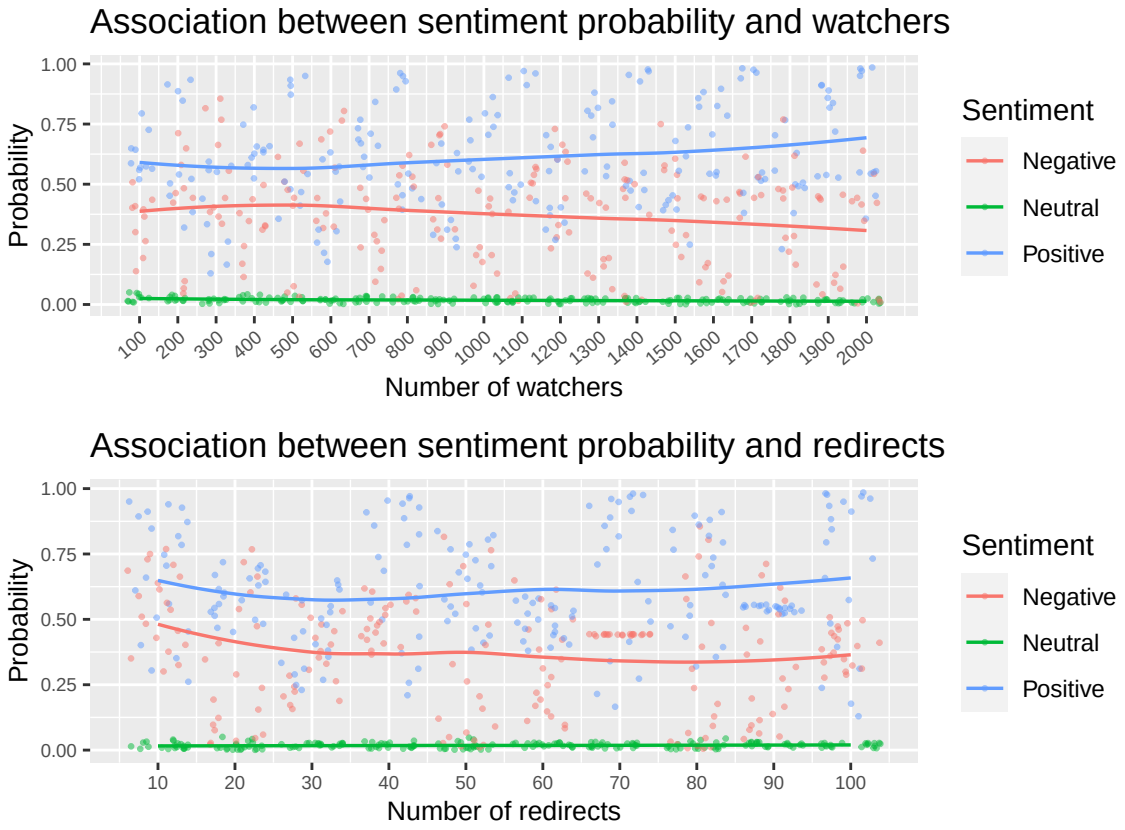


When the number of anonymous contributors increases, the positive sentiment probability decreases from the highest while the other two probabilities increases. Hence, the article popularity in anonymous contributors tends to be positively associated with non-positive sentiments. When the number of redirects increases, the neutral sentiment probability increases more significantly. Hence, the article popularity in redirects tends to be positively associated with neutral sentiment.

Another two graphs below show the association of the numbers of watchers and redirects with the

sentiment. Predictions are made similarly.

When the number of watchers increases, the positive sentiment probability increases and dominates while the negative sentiment probability decreases. Hence, the article popularity in watchers tends to be positively associated with positive sentiment.



The association between the number of redirects and the sentiment is slightly different from the association under the interaction with contributors. The negative sentiment probability decreases, while the positive sentiment probability first decreases and then increases up to the initial value. It suggests the article popularity in redirects tends to be negatively associated with negative sentiment.

Next, explanatory decision trees with depth of 2 and 3 are fitted through `SparkML` using the 4 attributes without more than 90% missingness before imputation. A deeper tree depth makes it harder to interpret the association between the popularity-related attributes and the sentiment, especially when the same attributes are used in different depths in a single tree. Hence, only 2 small tree depths are chosen for modelling. The comparison results of two trees are in table 8.

Depth	Features used	Accuracy (%)	F1-score (%)
2	Years, number of watchers	54.4	53.0
3	Years, number of watchers, number of redirects	50.7	46.0

Table 8: Comparison of two decision trees

The tree with depth of 2 fits the data better. Additionally, in the tree with depth of 3, the number of redirects is used only in one node at the final depth followed by two leaf nodes, so this attribute does not contribute much to the interpretation from the tree which slightly indicates that more redirects tend to be associated with more positive sentiment. Hence, the tree with depth of 2 is used to show the associations. The tree structure is shown in figure 1.

The classifications from the tree are the most likely sentiment at each leaf node. When the years are smaller than 15.5, increasing the years to be above 10.5 can make the most likely sentiment change from neutral to positive. This indicates more years are associated with more positive sentiment probability. It does not tell how the negative and neutral sentiment probabilities change. Given that the years are over

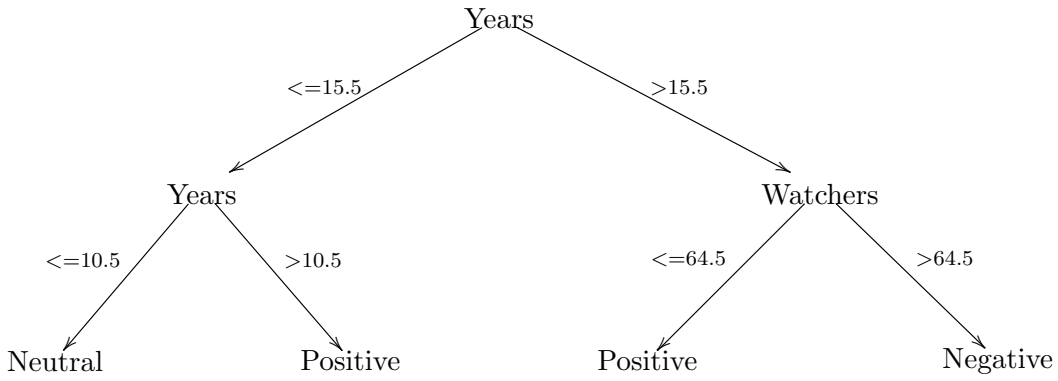


Figure 1: Structure of the decision tree with depth of 2

15.5, when the number of watchers increase from below 64 to above 64, the most likely sentiment changes from positive to negative. This indicates more watchers are associated with more negative sentiment probability. These results align well with the coefficients from the multinomial logistic model in table 6.

The decision tree with depth of 2 is more accurate and precise than the multinomial model with 2 interaction terms. The accuracy and F1-score are not high from both models, but the attributes used are still useful and the models are good to tell the associations. The multinomial logistic model includes the results from the decision tree and also produces interaction associations. Hence, it is more useful in this case.

In conclusion, the article popularity in anonymous contributors, watchers, redirects or years is associated with the sentiment differently. When interaction is considered:

1. The article popularity in anonymous contributors tends to be positively associated with non-positive sentiments.

2. The article popularity in watchers tends to be positively associated with positive sentiment.

3. The article popularity in redirects tends to be positively associated with neutral sentiment and negatively associated with negative sentiment.

4. The article popularity in years tends to be positively associated with non-neutral sentiments.

Hence, the article popularity is not associated with a single sentiment. Different aspects of popularity are associated with different sentiments.

Discussion

The data comes from a single source, English Wikipedia API, so it is relatively accurate and does not have issues related to multi-source data. In addition, the data covers a wide range of topics on the Wikipedia articles, which can enable us to understand the association of more topics with the sentiment. Some text structures from Wikipedia are known and easily identified, which facilitates the text cleaning process. The text information is relatively complete, so the computation of the sentiment is relatively reliable.

Missing values are a big problem in analysis. The data does not have many usable attributes related to the article popularity. Approximately 4 attributes, such as the numbers of logged-in contributors and pageviews, are mostly missing. Their missing values are imputed, but the imputation may not be informative, restricted by a small number of observations and the information from the other attributes that are already used in the analysis. The imputation may even affect the models. So it should be necessary to adjust the attributes in the models. Some of these attributes are not used, so I only have 4 usable attributes. These usable attributes may not be able to fully represent the article popularity. Hence, the association between the sentiment and the article popularity may not be comprehensively discovered.