Hello everyone. This is Natallia Casey. I welcome you all to my project presentation. Let me begin with the agenda. I will start with the introduction to the project and its objectives. I will also focus on the EDA, visualizations, actionable insights and best performing models. Finally, I will present my results and recommendations.

**Objectives**

* There are a few main objectives:
* Identifying the overall sentiment of the application TuneIn Radio
* Identifying the sentiment trends
* Determining the sentiment of TuneIn Radio main aspects and features
* Creating a model capable of effectively predicting a sentiment
* Pinpointing the possible problems with the product and concerns of the users
* Creating recommendations for customer support for user churn prevention

**Steps**

To approach any data science project we need to understand a basic data science life cycle. This is where the notion Cross-industry standard process for data mining, known as CRISP-DMcomes into play. The stages of CRISP-DM that I followed throughout the project are: Business Problem Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Recommendations for deployment

**Problem statement**

I will provide a short description of the background of the problem and explain why sentiment analysis with topic modelling.

* Online reviews influence customers’ purchasing behavior, and sentiment analysis is an effective business method of gaining insights about customer’s opinion of the product.
* Analyzing reviews manually is time-consuming but opinion mining creates an opportunity for a fast analysis of large volumes of data
* The sentiment of the product may vary based on its different aspects. Topic modeling used in sentiment analysis is essential for identifying possible problems, customer’s likes and dislikes sorted by specific features and aspects
* TuneIn Radio could benefit by monitoring the sentiment trends which could alarm the customer support if customer churn is a potential threat.
* Sentiment of the product can be extracted from a manifold of sources. The resulting approach (if works well) could be used to predict the sentiment from different origins (other internet app stores reviews and social media comments and etc.)

**Data**

A few details about the data used in this analysis. The data was scraped from [Google Play reviews](https://play.google.com/store/apps/details?hl=en&id=tunein.player) of TuneIn Radio application on March 16th, 2021. The dataset contains 8 variables and 194,229 records. It consists of 154,103 reviews with positive scores of 4 and 5, 10,036 observations with a neutral score of 3, and 29,936 with a negative rating score of 1 and 2. Personal information such as name and photo were removed for privacy protection.

The variable “content” is a categorical variable used in the task of topic modeling with Laten Dirichlet Allocation. In the classification modeling, the variable “score” is a target variable and the variable “content” is a predictor variable.

**Neutral Class Complexity.**

The sentiment analysis in this project is a multiclass classification problem. Besides the positive and negative classes, the neutral class is also used.

Neutral class is often excluded from sentiment analysis classification problems creating an assumption that the text could only be either positive or negative. This is not always the case because there is a lot of text data that doesn’t convey any sentiment. For example, the word “OK” could mean neither good nor bad. Just neutral. Indicating indifference.

There is text in the neutral class that conveys a mixed sentiment. This is especially hard for algorithms to comprehend, For example, “The app is good but I don’t like the absence of my favorite station”. 75% of the neutral class reviews in this data set contain mixed sentiment.

The neutral class often contains questions where the users express uncertainty about some aspect of a product, or simply need additional information. For example, “When is the next release?”.

As you can see, there are several semantic representations of neutral class which makes it especially complex and hard to predict.

**Results of EDA analysis.**

Identification of the number of missing values for the variable “replyContent” allowed to see which negative reviews haven’t been addressed yet. Especially important are the reviews whose relevancy is high (those are the reviews where the variable thumbsUP has a higher value). As a result, for the first quarter of 2021, a list of review IDs with high relevancy and lack of response by customer support professionals was created and suggested to be considered as a customer churn prevention measure. Also, the trend over time for the customer support responses to the reviews was displayed using the same information derived from missing values. If we look at the line chart, we can see that in comparison to 2018, there has been progress in addressing the concerns of users, and just a slight decrease in customer support responses for 2020 and 2021.

Another list with the review IDs for the observations containing the questions that haven’t been addressed by the customer support was also created based on detection of missing values.

Those are part of the strategy to establish contact and build a healthy relationship with the customer.

**Data Visualization. Distribution of overall Sentiment**

The main takeaways here are the fact that the overall sentiment of TuneIn Radio is positive and that there has been a slight decline in the positivity of the sentiment since 2016 (the year of the most positive opinion) as seen in the grouped bar chart.

**Data Visualization. Sentiment of TuneIn Radio Aspects.**

The identification of the aspects of the product and the visualization of the data related to them led to the following conclusions:

* Content is the aspect that users are most happy with.
* Subscription is the aspect that has the most negative sentiment.

The multiple line chart on the right shows that all of the aspects of the product noticed a significant increase in the number of negative reviews compared to 2016. If we look at the year of 2021, the negative opinion has seen an increase for all of the aspects.

**Data Visualization. Q1 2021.**

Based on the results displayed in the word-cloud

* UK is the most frequent word in the negative reviews for Q1, 2021 which indicates potential problems with the product offered for the UK-based users.
* MSNBC and BBC are also frequently mentioned in the negative reviews which could indicate that checking the streaming quality for these particular channels might be needed.
* The frequent use of “30 seconds” in negative reviews could potentially indicate a problem 30 seconds into the start of streaming.
* The frequently used phrase “7 days” might suggest that users are unhappy with the 7-day trial.

The distribution of the replies to the negative reviews in 2021 with the help of the bar chart displays the highest performance of customer support in the month of March.

**Tackling imbalanced dataset.**

The original dataset used for this analysis is highly imbalanced with most of the reviews being positive. Several techniques were tested before the final balanced sample was created.

Random oversampling, oversampling with SMOTE and undersampling the data were examined using the cross-validated logistic regression model. Even though the accuracies ended up being higher for the oversampled data samples and the original imbalanced data, their macro averaged F1 scores are smaller than the F1 of the undersampled data (which is at 57%). It is mostly due to the fact that they struggle with the neutral class identification which as previously mentioned is an important class. The highest F1 score of the oversampled data is more than two times less than that of the undersampled data. The decision was made to proceed using the undersampled data sample.

**Best performing models.**

The project consisted of two different types of modeling. The topic modeling with the Latent Dirichlet Allocation, and a classification task tested with several models.

The number of components for the best performing LDA model is 4, and the number of iterations is 60. The results of the topic modeling task included 4 topics mapped to the following product’s aspects:

Streaming

Subscription

Content

Technical features

Classification modeling

The evaluation of the models showed that the Support Vector Machines with the default parameters was the best performing model. The metrics of F1, accuracy, recall, and precision are all the highest for this model (at about 70%). All of the models tended to struggle with the neutral class identification while the Support Vector Machines identified the neutral and negative reviews the most correctly (especially as compared to neural networks) which contributed to its overall better performance. This is important because the negative and neutral classes of reviews are prioritized for the replies by customer support specialists.

**Additional evaluation of the tested models**

Additional testing and evaluation of the created models, the popular sentiment analysis tool Vader, and human’s sentiment classification (my own in this case) was performed on a small sample of 15 balanced observations. It looks like our best model (support vector machines) classified the sentiment of these observations as well as I did. It is important to remember that sentiment analysis is a complex problem even for humans. It could be very subjective. Humans tend to agree on the sentiment overall only about 80% of the time (Barba, P.).

**Results of the project**

The best performing model that I built, the Support Vector Machine, with the metrics of accuracy, F1, recall and precision at about 70% implies that 70% of the time the expected predictions of the sentiment would be correct. This is a satisfactory result considering the complex nature of sentiment analysis.

Topic modeling with Latent Dirichlet Allocation resulted in creating 4 clear aspects of TuneIn Radio application: Content, Subscription, Technical features, and streaming which helped expand the analysis by deriving valuable insights about each of these topics.

The project allowed to create an overall picture of the sentiment of TuneIn Radio:

It is positive overall.

Most positive in 2016 with a slow decrease in positivity afterwards.

for the Q1 2021, the sentiment is most positive in the month of March

Highest positive sentiment is expressed for the Content aspect

Most negative - for the Subscription aspect

Most relevant negative opinions have been addressed mostly successfully by customer support

The overall responses to the negative reviews have increased compared to the lower number in 2018

The project allowed to create a list of reviewIDs that still need to be addressed and identify potential problems of the service.

These findings are useful and could be the base of a new business or customer support strategy for product improvement and customer churn prevention.

And finally, **Challenges, limitations and recommendations.**

One of the main challenges of this project is the imbalanced dataset which resulted into a smaller size sample for the final classification modeling stage. Another challenge was getting an acceptable classification results for the neutral class due to the complexity of it. The results of the Support Vector Machines are acceptable and could be improved by better examining the multipolarity present in this class.

The main limitations of the project included the complexity of the language such as interpretation and handling of the following linguistic phenomena: sarcasm, irony, hyperbole, and ambiguity that create challenges and decrease in the accuracy and performance for most of the types of text data analysis.

And to conclude the presentation of the project, I would like to identify the main recommendations:

For the support team:

to address all of the named problems and concerns of the users, and answer the recent questions of the customers.

And a recommendation for the project improvement is to address the multipolarity issue of the neutral class and separate the polar sentiments found in the observations. I believe that this could be a way to improve the accuracy of predictions.

Thank you for listening to my presentation. I hope you enjoyed it and learned a lot.