1.State a business reason for selecting your tools (problem you would like to solve).

You will state your reasons 😁

2. Document how you used the tool. Many tools are super rich in features and you probably won't be exploring them all, explain the parts you did use.

I used a python script to do a sentiment analysis of marvel movie mentions in tweets within days of the movie release. Using Social Network Scrape library to scrape tweets, NLTK library to process natural language, scikit learn library to do classification and get models to train the sentiment analyzer developed.

3. When you choose a data mining algorithm(s) for you mining model, tell us why you chose that

one (or that category of algorithms)

For the algorithms, I tried three different data mining algorithms. They include:

Naive Bayes: The reason being they are fast and effective in processing large amounts of text data

SVM: the reason being they are a powerful and versatile classification algorithm that can handle complex and high-dimensional data

Logistic regression: the reason being it is simple, fast, and provides interpretable results.

Comparing the models can show the various different results and how the various algorithms perform.

4. Document how/where you got your data (if it is publicly available, or internal for a work project).

I fetched the data by doing a twitter scrape for tweets containing the movie titles. First, I had to get a list of Marvel movies released from 2005 and their release dates. From there I added an up until date as a search filter for searching and scraping tweets that contained mention of the movie name. I fetched 1000 tweets per movie bringing the number of tweets to 30,045 tweets as it seems some tweets were omitted. After that, the data was processed and cleaned for use to test the models. Training data was fetched from the various supported and imported libraries.

5. Given an explanation/analysis of the output (What did you learn or uncover).

The table below contains the results

Classification Algorithm	Positive Tweets	Negative Tweets	Neutral Tweets	Accuracy
Naive Bayes	14461	153351	233	95%

SVM (predict)	16257	13788	0	91%
Logistic Regression	17200	12845	0	89%
SVM (cross validation)	6561	23484	0	83%

The tweets were scrapped, numbering 30,045. The criteria for selecting tweets was to collect tweets between the day the movie was released and two days after. The total movies searched were 31, and on average looking for 1000 tweets a movie. This data was unsupervised. A supervised dataset was fetched in order to train our models and to test its accuracy. Using an inbuilt Multinomial Naive Bayes analysis to test for the testing data accuracy which was at 0.95 or 95% accuracy. After training the Naive Bayes sentiment analyzer we came up with, testing was done with the Marvel tweets. The results show that there is negative sentiment towards Marvel with 15351 being negative and 14461 being positive and 233 being neutral. The SVM using the predict function had a result of 13788 negative tweets and 16257 positive tweets with an accuracy of 0.91 or 91%. It is to be noted that there was no neutral sentiment. The second SVM using the cross validate function had a result of 12845 negative tweets and 17200 positive tweets with an accuracy of 0.83 or 83%. The logistic regression model had 23484 negative tweets and 6561 positive tweets with an accuracy of 0.89 or 89%.

Conclusion

The general sentiments towards Marvel movies is negative, but the intensity isn't that skewered toward extreme negativity as can be seen in the various graph plots apart from the logistical regression model. The reasons vary but may be associated with inconsistencies within the selected data as There could be many duplicates such as retweets, the size dataset may not be fully representative of the actual sentiments of the actual movie and since there are a number of different movies, each having their own different receptions and may affect the results of each other during the analysis

6. Conclude with the 3 W's (What Went Well, What Did NOT go Well, What Would you do Differently Next Time).

What went well

The data collection process was challenging but interesting. Seeing how you can fine tune a

search and see the results, compare it to some of the personal hypotheses that I had and thinking about how insightful the information was made the process worth the sweat.

What went wrong

At first my tool of choice was rapid miner. Unfortunately, due to the large dataset and hardware limitations of my computer, I was forced to switch to using another tool. I settled for coming up with a python script as a challenge to myself, with some assistance from the internet and friends who are well conversant with the inner workings of the language and libraries to develop the program.

What would I do differently

I would try to associate the tweets with the ratings of the movie on critique sites such as Imdb to see if there is some similarities to the movie reception from a social platform and a rating site.I would further investigate each individual intellectual property and see how the reception has been over a function of time i.e. if for example, Iron Man has seen a more positive reception growth from the first movie to the last Iron Man or any movie containing the character.