Text Analytics of Course Reviews on Coursera Platform

Chan Huan Yang, Ramindhran Raja Mohan

School of Computer Science, University Science of Malaysia, Penang, Malaysia

Introduction

We are now entering a new era - the revolution of online learning. From working professionals to recent high school graduates, many of them have found the reasons to take all or some of their courses online in platform such as Coursera, Udemy, and Edx

Problem Statement

Ratings and reviews are always the major consideration factor by online course seekers before they joining the course. However, it can be time-consuming to read all the information especially the course reviews.

Research Question

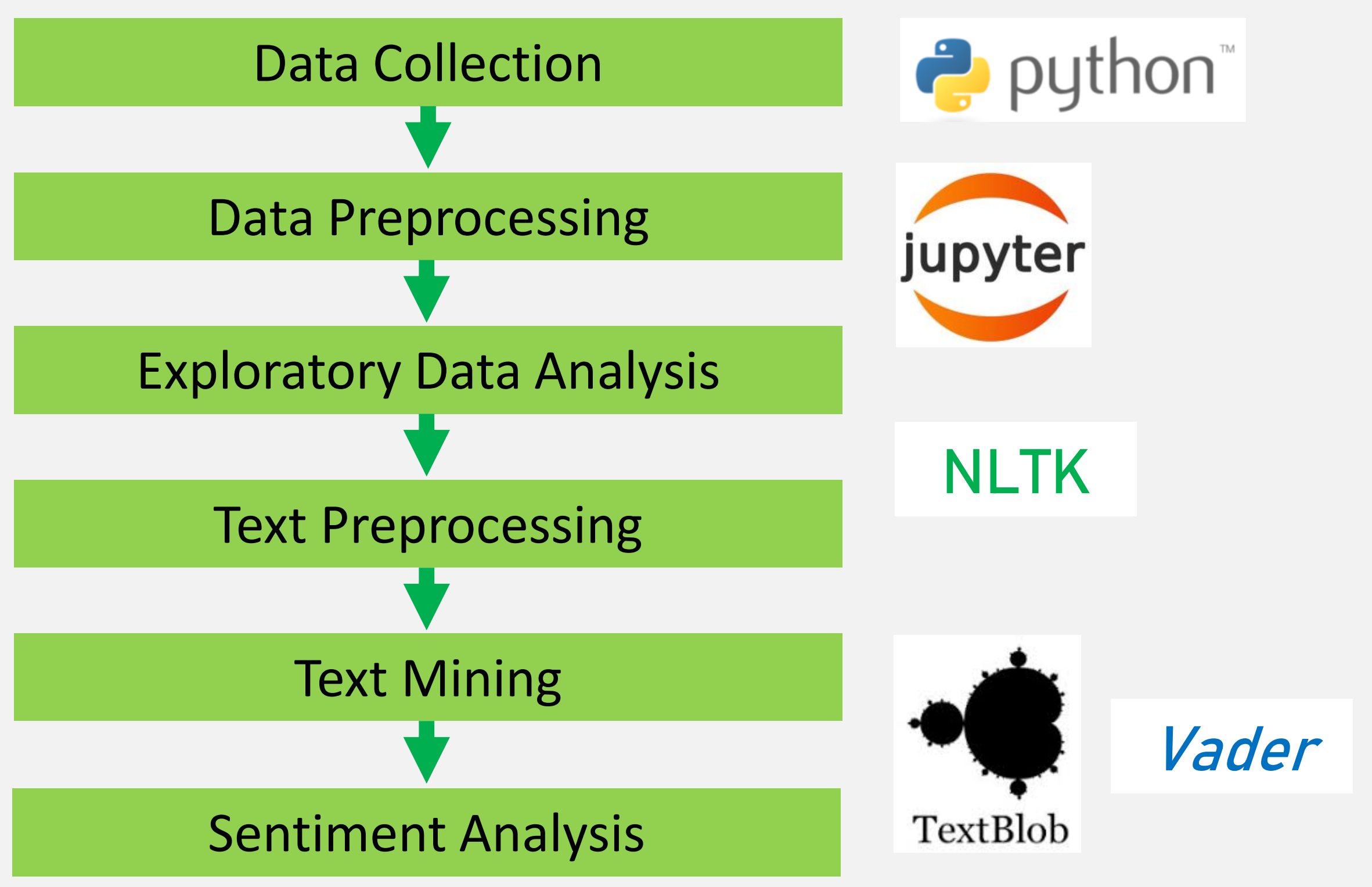
The research questions for this work were:

- How text analytics techniques such as n-gram analysis, word cloud, and sentiment analysis can be applied to improve the online course searching process?
- What insights can be obtained by using text analytics techniques such as n-gram analysis, word cloud, and sentiment analysis?

Purpose of The Study

Our objective is to propose a text analytics pipeline that includes text cleaning, text lemmatization, sentiment analysis, text mining, and visualization that can help course seekers to gain a quick insight into the courses as well as enables them to make a quick comparison between multiple courses

Research Method



Data Collection → Data Preprocessing → Exploratory Data Analysis → Text Preprocessing → Text Mining → Sentiment Analysis

Data Collection

The data used in this work is from Kaggle.

Course Name	URL	No of Review
Programming for Everybody (Getting Started with Python)	https://www.coursera.org/learn/python	45218
Python Data Structures	https://www.coursera.org/learn/python-data	33543
Introduction to Data Science in Python	https://www.coursera.org/learn/python-data-analysis	14289

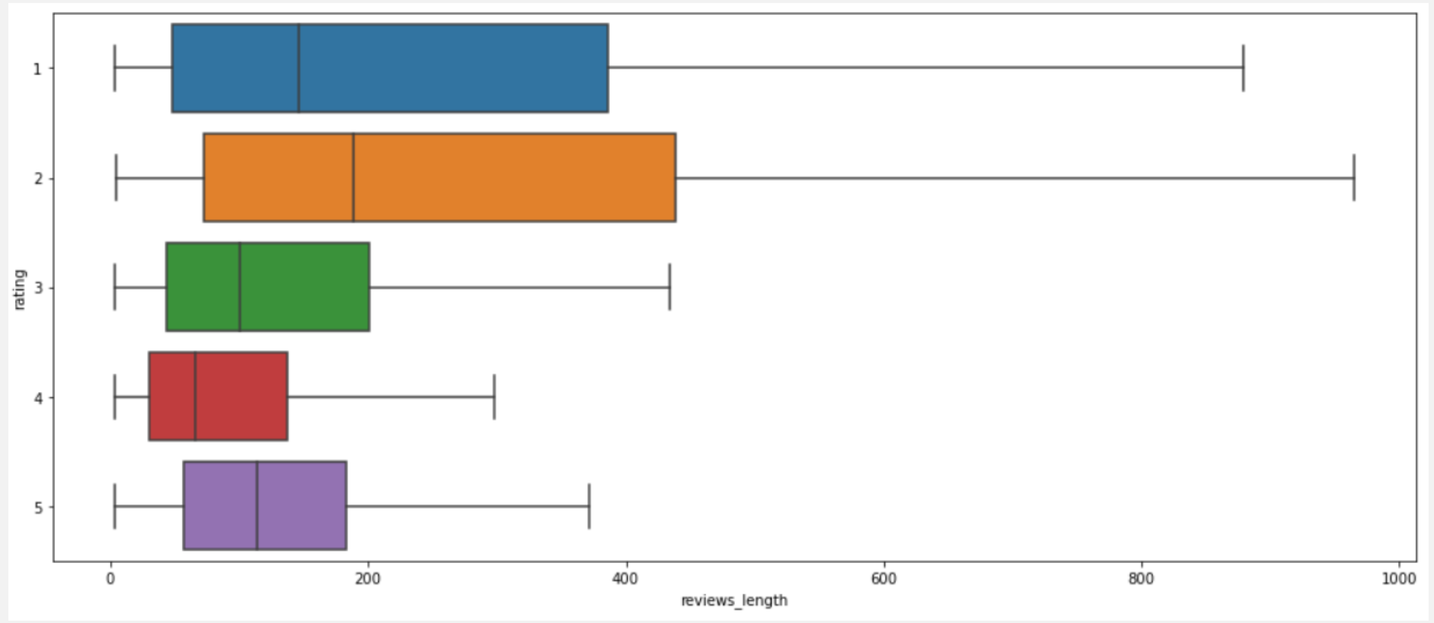
Data Preprocessing

Data preprocessing steps include:

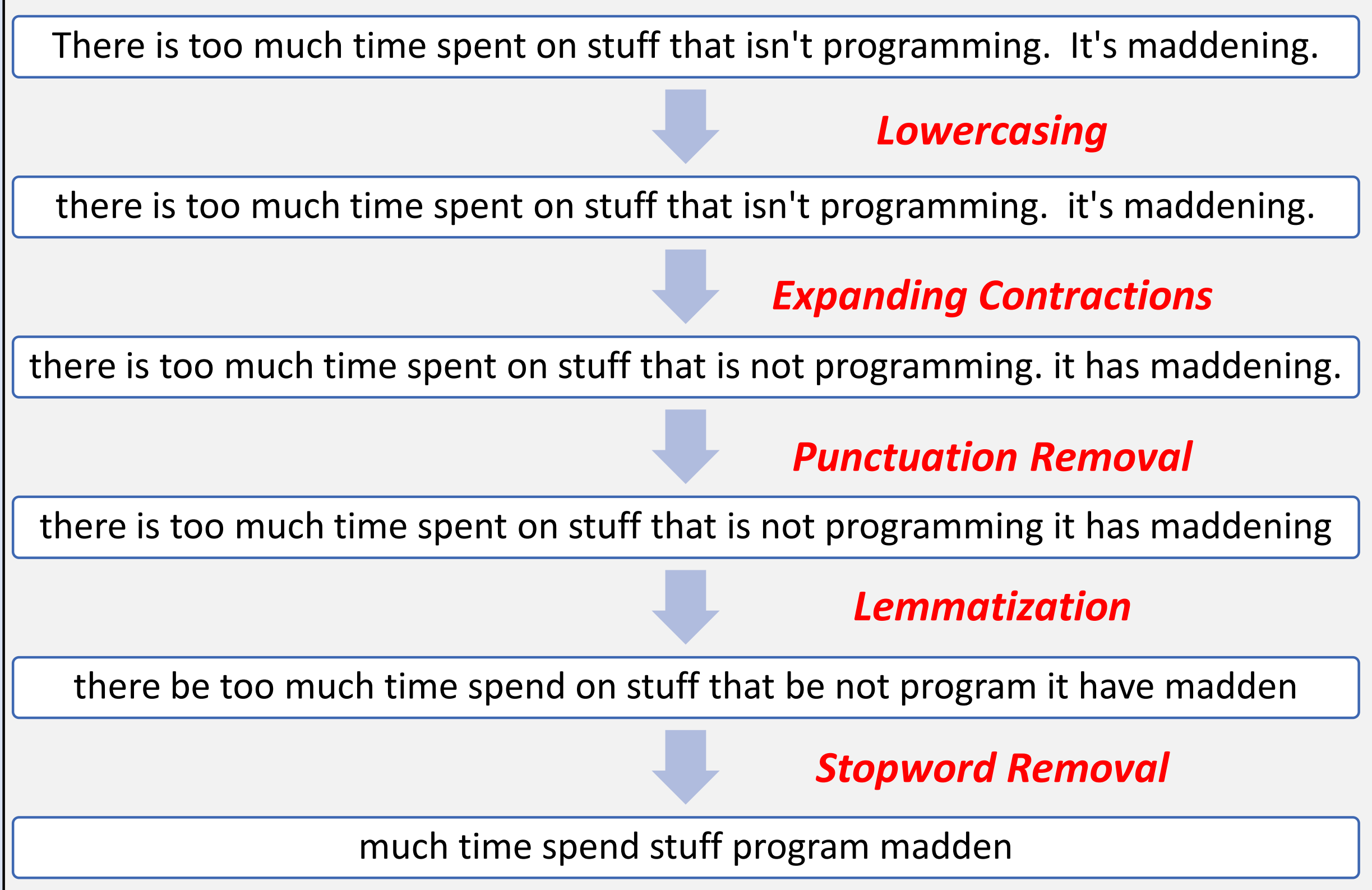
- removed the duplicate reviews
- removed the reviews with string's length less than three
- selected the English labeled reviews only using package "langid"

Exploratory Data Analysis

Reviews length of each rating



Text Preprocessing



There is too much time spent on stuff that isn't programming. It's maddening.

Lowercasing

there is too much time spent on stuff that isn't programming. it's maddening.

Expanding Contractions

there is too much time spent on stuff that is not programming. it has maddening.

Punctuation Removal

there is too much time spent on stuff that is not programming it has maddening

Lemmatization

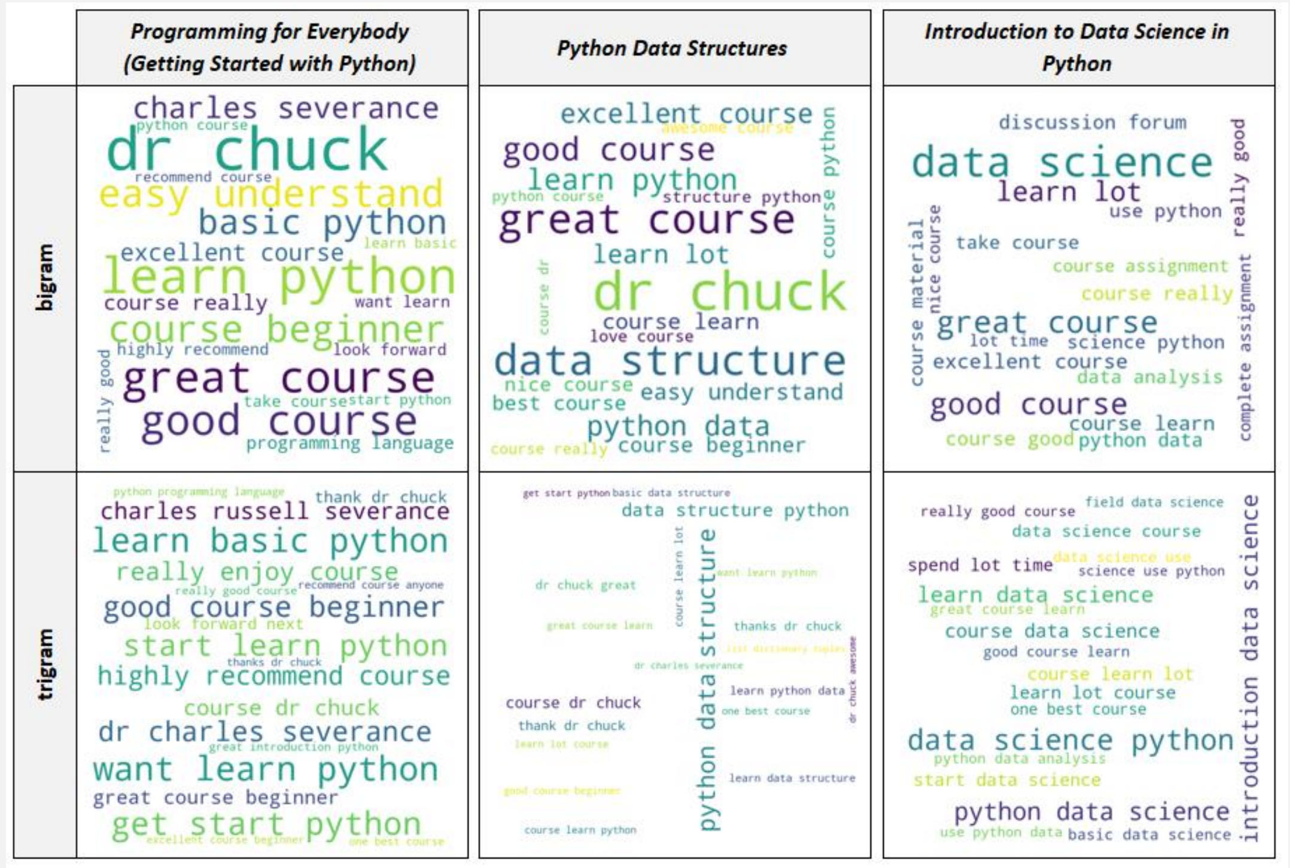
there be too much time spend on stuff that be not program it have madden

Stopword Removal

much time spend stuff program madden

Text Mining

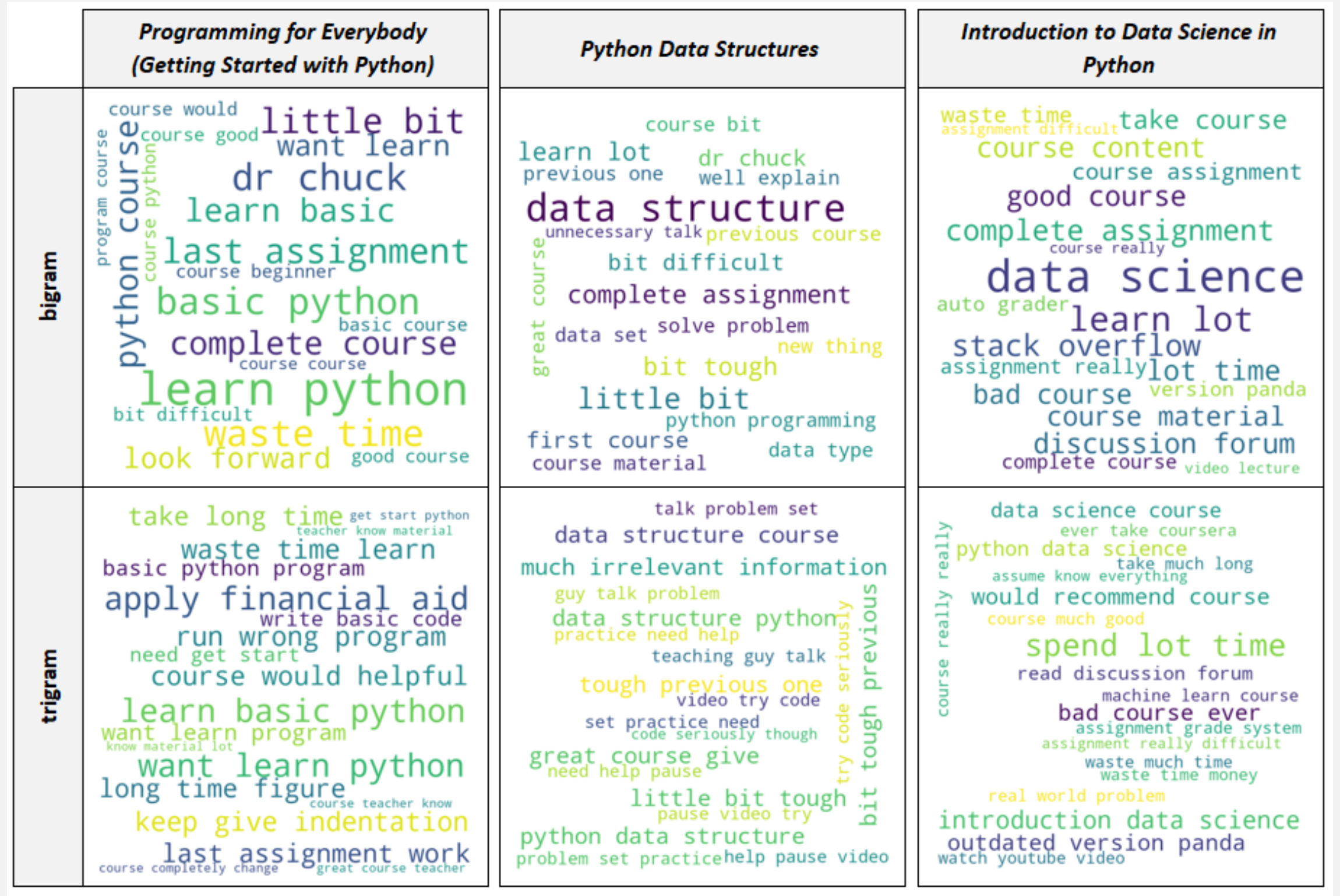
Word cloud of overall reviews



Sentiment Analysis

	Average of Polarity		
name	**Textblob**	**Vader**	**Overall**
Programming for Everybody (Getting Started with Python)	0.40	0.68	0.54
Python Data Structures	0.49	0.62	0.55
Introduction to Data Science in Python	0.34	0.45	0.40

Word cloud of negative reviews



Discussion and Finding

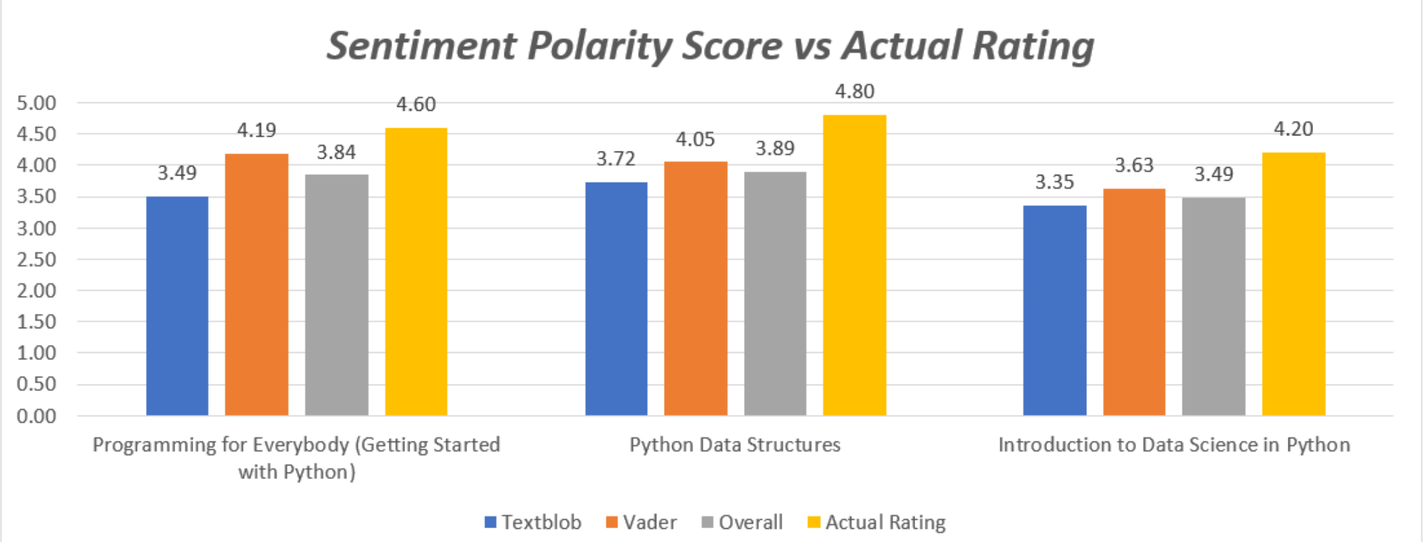
Evaluation of accuracy and usability of n-gram

Programming for Everybody (Getting Started with Python)		
n-gram	**accuracy**	**usability**
dr chuck	1	1
learn python	1	1
great course	1	1
good course	1	1
easy understand	1	1
course beginner	1	1
basic python	1	1
charles severance	1	1
excellent course	1	1
course really	0	0
want learn python	1	1
learn basic python	1	1
get start python	1	1
start learn python	1	1
good course beginner	1	1
dr charles severance	1	1
highly recommend course	1	1
really enjoy course	1	1
charles russell severance	1	1
course dr chuck	1	1

Python Data Structures		
n-gram	**accuracy**	**usability**
dr chuck	1	1
data structure	1	1
great course	1	1
good course	1	1
learn python	0	1
python data	1	1
excellent course	1	1
learn lot	1	1
easy understand	1	1
course learn	0	0
python data structure	1	1
data structure python	1	1
course dr chuck	1	1
thank dr chuck	1	1
dr chuck great	1	1
thanks dr chuck	1	1
learn python data	1	1
learn data structure	1	1
course learn lot	1	1
good course beginner	1	1

Introduction to Data Science in Python		
n-gram	**accuracy**	**usability**
data science	1	1
great course	1	1
good course	1	1
learn lot	1	1
course learn	0	0
excellent course	1	1
discussion forum	0	0
course really	0	0
course good	1	1
python data	0	1
data science python	1	1
python data science	1	1
introduction data science	1	1
learn data science	1	1
course data science	1	1
learn lot course	1	1
course learn lot	1	1
start data science	1	1
spend lot time	0	1
data science course	1	1

Evaluation of sentiment polarity score

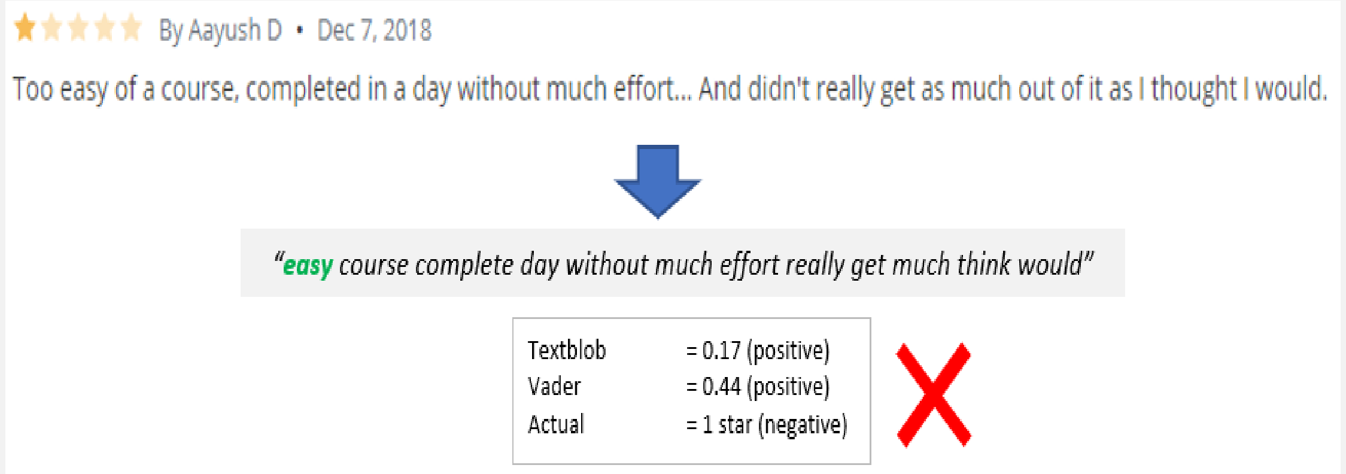


Confusion matrix and classification report

	Predicted negative	*Predicted positive*
Observed negative	326	385
Observed positive	772	28324

	precision	*recall*	*f1-score*	*support*
negative	30%	46%	36%	711
positive	99%	97%	98%	29096
accuracy		96%		29807
macro average	64%	72%	67%	29807
weightage avg	97%	96%	97%	29807

Example review from the dataset



★☆☆☆☆ By Aayush D • Dec 7, 2018

Too easy of a course, completed in a day without much effort... And didn't really get as much out of it as I thought I would.

"easy course complete day without much effort really get much think would"

Textblob	= 0.17 (positive)
Vader	= 0.44 (positive)
Actual	= 1 star (negative)

Conclusion

The n-gram analysis and word cloud are sufficient enough to provide an accurate and informative glance into the course. However, it falls short on sentiment analysis especially in detecting the negative reviews.