

Text Analytics of Course Reviews on Coursera Platform

Chan Huan Yang
School of Computer Science
University Science of Malaysia
Penang, Malaysia
chyng2000@gmail.com

Ramindhran Raja Mohan
School of Computer Science
University Science of Malaysia
Penang, Malaysia
rampg08@hotmail.com

Abstract— 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. In this research work, 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. The proposed text analytic pipeline was created in Python Jupyter Notebook and three different Python-related courses were chosen for the study and demonstration. The proposed text analytics pipeline solution was proved able to achieve our research objective. It can help course seekers to gain a quick insight including the positive and negative reviews into the courses as well as enables them to make a quick comparison between multiple courses. The n-gram analysis and word cloud generated were sufficient enough to provide an accurate and informative glance into the course. However, it fell short on sentiment analysis especially in detecting the negative reviews.

Keywords—text mining, n-gram, word cloud, reviews, sentiment analysis, Textblob, Vader

I. INTRODUCTION

The concept of traditional education has changed drastically within the last couple of years. With the rise of the internet and new technologies, being physically present in a classroom isn't the only learning option anymore. Nowadays, we can access a quality education whenever and wherever we want, as long as we have access to a computer. We are now entering a new era - the revolution of online learning.

Some of the best online learning platforms include Coursera, Skillshare, Udemy, Codecademy, Edx, Pluralsight, Future Learn, and Moodle. These online courses deliver a series of lessons to a web browser or mobile device which are conveniently accessible by the internet users at any time, anyplace. From working professionals to recent high school graduates, especially after the advent of Covid-19, many of them have found the reasons to take all or some of their courses online.

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. In this research work, two different text analytics techniques such as term frequency and sentiment analysis are used to develop a simple model for online course seekers to gain quick insight into their shortlisted courses. The data source used in this work is the course reviews and ratings given by students on the Coursera platform. Three different Python-related courses from Coursera were chosen for the study and demonstration of the various text analytics techniques. The outputs of this work are word cloud that represents the positive and negative reviews

of each course, which can provide quick insights about each course as well as offering simple comparison or similarity between courses.

II. PROBLEM STATEMENT

There are many advantages to online education, but there are plenty of factors to consider before becoming an online student. It is always better to explore all the preferable courses before jumping into any decision. Opting for a course that is not appropriate for us can knock down our confidence and harm our potential and abilities. Although the plethora of choices available online can confuse us and making a decision can turn out to be very difficult.

In the world of online shopping, often buyers will purchase an item only after they see that other people also like it. And the easiest way to find out consumer sentiment is based on reviews. According to Fan & Fuel [1], 94% of online customers read reviews before making any purchasing decisions. For product-specific information, Spiegel Research Centre [2] shows that 95% of shoppers read reviews before making a purchase. Small Business Trends [3] shows 83% of job seekers use reviews to support their decisions on which companies to apply to and 84% of patients use online reviews to evaluate physicians before checking in. Authentic reviews are a valuable tool when we try to compare among different courses. They can help us make important decisions by learning about someone else's experience

However, it could be time-consuming to review all the comments to identify the positives and negatives of a course. Furthermore, a course searching process will usually involve multiple courses. Therefore, an effective system is needed to assist in the review of all the student's comments.

III. RESEARCH QUESTION

In this research work, our objective is to create a text analytics-based solution for online course seekers to improve their course seeking process efficiency. The research questions for this work were:

- How text analytics techniques such as n-gram analysis, Wordcloud, 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, Wordcloud, and sentiment analysis?

IV. PURPOSE OF THE STUDY

In this research work, 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.

V. RESEARCH METHOD

We have created a text analytic pipeline in Python Jupyter Notebook that can help course seekers to gain a quick insight into their shortlisted courses. The overall flow of text analytics is as below:

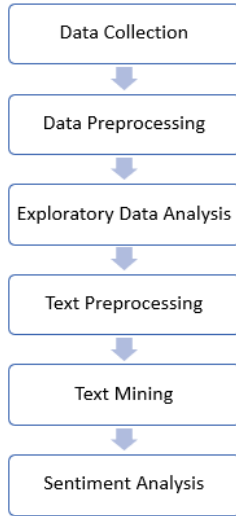


Figure 1. The flow of text analytics used in Coursera’s course review analysis

A. Data Collection

The data used in this work is freely available at Kaggle <https://www.kaggle.com/imuhammad/course-reviews-on-coursera>. This dataset has a total of 1.4M student reviews on 604 different Coursera courses. The table below describes the column features of the dataset.

Table 1. Description of dataset features

variable	class	description
reviews	character	Course review
reviewers	character	The name of the reviewer who wrote the review
date_reviews	date	Date when the review was posted
rating	integer	The rating score which is given by the reviewer to the course
course_id	character	Course ID

We only selected three different Python-related courses for the study and demonstration of the text analytics pipeline. The courses selected from the dataset are the following:

Table 2. Three selected courses

name	course_id	no_of_review
Programming for Everybody (Getting Started with Python)	python	45218
Python Data Structures	python-data	33543
Introduction to Data Science in Python	python-data-analysis	14289

All three selected courses are highly related to Python. For example, the course “Programming for Everybody (Getting Started with Python)” aims to teach everyone the basics of programming computers using Python. It covers the basics of how one constructs a program from a series of simple instructions in Python. The second course “Python Data Structures” aims to introduce the core data structures of the Python programming language. It will move past the basics of procedural programming and explore how a student can use the Python built-in data structures to perform increasingly complex data analysis. The third shortlisted course “Introduction to Data Science in Python” aims to introduce the learner to the basics of the python programming environment, including fundamental python programming techniques. It will also introduce data manipulation and cleaning techniques using the popular python pandas data science library.

B. Data Preprocessing

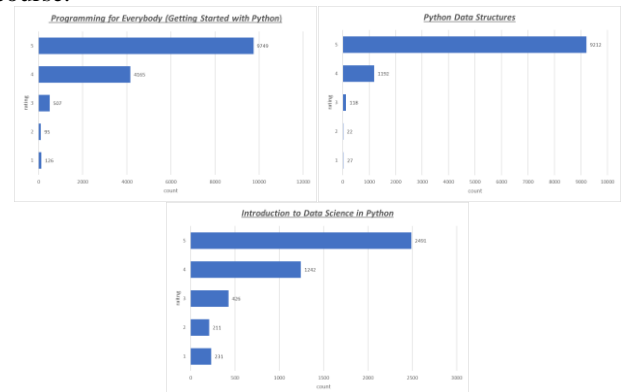
In this stage, data preprocessing and cleaning were performed to ensure the validity of the data. We have removed the duplicate reviews from the dataset. Besides, we also removed those reviews with string’s length less than three from our dataset. Finally, we used the package “langid” to label the language used in each review. “langid” is pre-trained from 5 different sources and able to recognize 97 different languages. We only selected those English labeled reviews for our study. Below is the number of reviews of each shortlisted courses after data preprocessing and cleaning:

Table 3. Before and after data preprocessing

name	no of review before data preprocessing	no of review after data preprocessing
Programming for Everybody (Getting Started with Python)	45218	14642
Python Data Structures	33543	10571
Introduction to Data Science in Python	14289	4601

C. Exploratory Data Analysis

In the exploratory data analysis stage, we analyzed data sets and summarized their main characteristics. We first took a look at the ratings given by students of each shortlisted course.



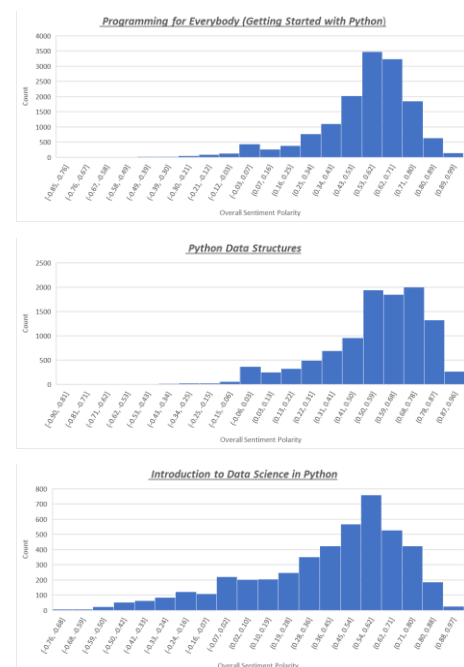
	<p>Programming for Everybody (Getting Started with Python)</p> <p>This word cloud features 'chuck' and 'severance' as the most prominent words. Other visible terms include 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>Python Data Structures</p> <p>This word cloud is dominated by 'chuck' and 'severance'. Other words include 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>Introduction to Data Science in Python</p> <p>This word cloud features 'data', 'science', 'python', and 'beginner' as prominent words. Other visible terms include 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>
bigram	<p>This bigram word cloud shows pairs of words. Prominent pairs include 'chuck', 'severance', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>This bigram word cloud shows pairs of words. Prominent pairs include 'chuck', 'severance', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>This bigram word cloud shows pairs of words. Prominent pairs include 'data', 'science', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>
trigram	<p>This trigram word cloud shows groups of three words. Prominent groups include 'chuck', 'severance', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>This trigram word cloud shows groups of three words. Prominent groups include 'chuck', 'severance', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>	<p>This trigram word cloud shows groups of three words. Prominent groups include 'data', 'science', 'python', 'beginner', 'great', 'course', 'learn', 'understand', 'easy', 'basic', 'excellent', 'really', 'good', 'recommend', 'start', 'program', 'language', 'tutorial', 'book', 'video', 'series', 'first', 'step', 'way', 'to', 'go', 'about', 'python', 'programming', 'for', 'everyone', 'who', 'wants', 'to', 'learn', 'python', 'programming', 'from', 'scratch', 'this', 'book', 'is', 'the', 'perfect', 'resource', 'for', 'anyone', 'looking', 'to', 'get', 'started', 'with', 'python', 'programming'.</p>

By looking at the bigram and trigram analysis, we can easily get some quick insights from them. For example, course 1 is conducted by someone called Dr. Chuck or Charles Severance. It is somehow a Python programming course and generally has positive reviews. For course 2, it is also taught by someone called Dr. Chuck or Charles Severance. It is somehow a Python data structure course and generally has very positive reviews too. For course 3, it is somehow related to data science and Python.

After performed basic text mining techniques, we proceed to apply sentiment analysis to the course reviews. Sentiment analysis (or opinion mining) is a natural language processing technique used to determine the polarity (positive, negative, neutral) of the data. It was performed on the course reviews to help course seekers to have a quick insight into the possible positive and negative reviews of their shortlisted courses. It is extremely important for course seekers because it helps them to quickly understand the overall opinions of the previous students. By automatically sorting the sentiment behind reviews, we can make a faster and more accurate decision

- Defines two lists of polarized words.
- Counts the number of positive and negative words that appear in a given text.
- If the number of positive word appearances is greater than the number of negative word appearances, the system returns a positive sentiment

We performed the sentiment polarity on the reviews of shortlisted courses using both Textblob and Vader and then get the average polarity score of them. All three shortlisted courses show negatively skewed distribution which meant that they have a relatively positive sentiment. However, we observed that the course “Introduction to Data Science in Python” clearly has some negative sentiment in which the polarity value is lower than 0. Below is the distribution of sentiment polarity of the three shortlisted courses.



	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

After obtained the polarity of each review, we labeled those polarities that score greater than 0 as positive reviews, lower than 0 as negative reviews. Normally a course seeker will always check for negative and positive reviews before they decide whether to purchase or not. Hence, we also created n-gram analysis and word cloud to help course seeker to has a

glance of insight into the negative and positive reviews of each shortlisted course. Below is the n-gram analysis of all three shortlisted courses' negative and positive reviews.

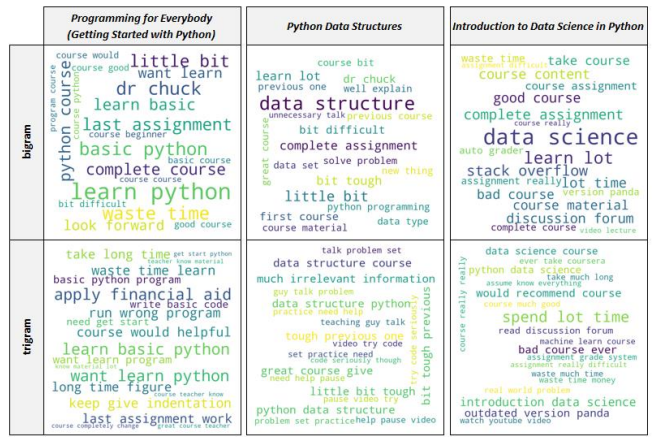


Figure 8. Word cloud of each shortlisted course's negative reviews

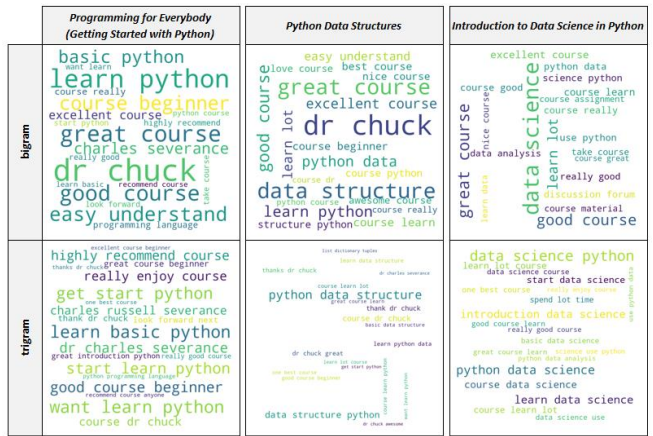


Figure 9. Word cloud of each shortlisted course's positive reviews

Our review polarity data is negatively skewed. Hence, from Figure 9, we can notice that very little new insight can be extracted from the positive review word clouds as the bigrams and trigrams generated is similar to the word clouds in Figure 6. However, some interesting insights can be found in the negative review word clouds (Figure 8). For example, for the course “Introduction to Data Science in Python” some reviews highlight that it using an outdated version of Pandas and the assignment is really difficult.

VI. DISCUSSIONS AND FINDINGS

In this work, a qualitative evaluation method was used to evaluate the usability of the n-gram/word cloud platform to represent the company's insight. In the evaluation part, the information was collected from the Internet and Coursera itself to verify whether the n-gram/word cloud in Figures 5 and 6 are accurate and useful in representing the overall insight about the shortlisted courses. The summary of the evaluation is shown in Figure 10.

Programming for Everybody (Getting Started with Python) https://www.coursera.org/learn/python				
n-gram	accuracy	usability	remark	
dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
learn python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
great course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
good course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
easy understand	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
course beginner	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
basic python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
charles severance	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
excellent course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
course really	0	0	No useful insight can be extracted from this bigram	
want learn python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
learn basic python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
get start python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
start learn python	1	1	This course indeed aims to teach everyone the basics of programming computers using Python.	
good course beginner	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
dr charles severance	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
highly recommend course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
really enjoy course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.6/5.0	
charles russell severance	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
course dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	

Python Data Structures https://www.coursera.org/learn/python-data				
n-gram	accuracy	usability	remark	
dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
data science	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
great course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	
good course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	
learn python	0	1	This course is about Python data structure. Not general Python	
python data	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
excellent course	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	
course dr chuck	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	
easy understand	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	
course learn	0	0	No useful insight can be extracted from this bigram	
python data structure	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
data structure python	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
course dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
thank dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
dr chuck great	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
thinks dr chuck	1	1	The course indeed is conducted by Dr Charles Russell Severance (aka (a.k.a. Dr. Chuck))	
learn python data	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
learn data structure	1	1	This course indeed aims to introduce the core data structures of the Python programming language	
good course beginner	1	1	The bigram is positive sentiment which is consistent with the course's high actual rating of 4.8/5.0	

Figure 10. Evaluation of accuracy and usability of n-gram

The evaluation result shows that our n-gram analysis performs well by scoring 52/60 in accuracy and 55/60 in usability respectively. Overall, the system can portray the overall insight of the shortlisted course correctly and informatively. However, there is some concern about repetitive or redundant information is generated from n-gram analysis. For example, trigram “data science python” and “python data science” both actually carry the same insight but only different in their term sequence.

The sentiment polarity score generated using Textblob/Vader was evaluated by comparing them with the actual rating given by the previous students.

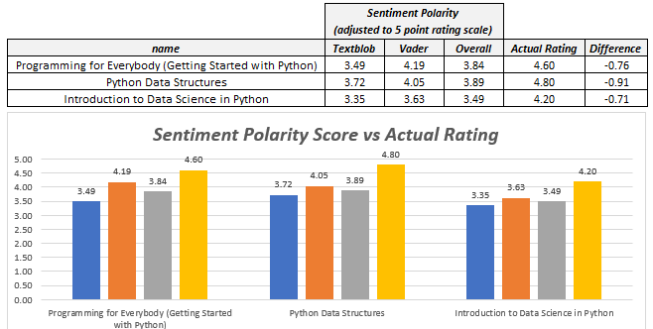


Figure 11. Evaluation of sentiment polarity score

As we can see from Figure 11, the overall sentiment polarity score generated by Textblob/Vader averagely lower

than the actual rating by about 0.8. Although the error is quite big there (-16%) however the system still able to get the ranking right which is the course “Python Data Structures” > “Programming for Everybody” > “Introduction to Data Science in Python”.

The sentiment classification performance of the Textblob/Vader also was evaluated by comparing it with the actual rating. For those actual rating with value > 3, we labeled them as “positive” and vice versa. The model performance of Textblob/Vader was then evaluated using a 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

Figure 12. Confusion matrix and classification report

As we can see from Figure 12, the model shows a 96% accuracy. Accuracy is the number of correct predictions made by the model by the total number of records. However, for an imbalanced dataset, accuracy is not a valid measure of model performance as the positive reviews of our dataset already consist of 97% of the total review. So accuracy is not the right measure for model performance in this scenario. The model has a high true positive rate of 97% which indicated it performs well when coming to predict positive sentiment. However, it is quite poor when comes to predict negative sentiment as the true negative rate is only 46%.

The major challenging issues in sentiment analysis is highly related to the correct interpretation of the context in which certain words are used. It is still difficult for a vast majority of tools such as Textblob and Vader to precisely evaluate what truly is a negative and a positive statement. Both the Textblob and Vader simply not advanced enough to successfully deal with the sarcasm or context of some of the discussions. Think of an example of someone being sarcastic in their reviews. Have a look at the simple review below which is picked from our dataset:

<p>★ ★ ★ ★ ★ By Aayush D • Dec 7, 2018</p> <p>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.</p>

Figure 13. Example review from the dataset

This review went through all the text preprocessing/normalization steps and had been reduced to:

“easy course complete day without much effort really get much think would”

Both the Textblob and Vader labeled it as positive with a polarity score of 0.17 and 0.44 respectively. The reason for it is the word “easy” which is positive in its essence. However, in this particular case, it was used to express disappointment

and isn’t a good thing for the student’s expectation. Still, it’s a perfect example that in some instances, a pair of eyes of a person is essential to properly evaluate the sentiment of a piece of social media content.

VII. CONCLUSION

In conclusion, the proposed text analytics pipeline solution able to achieve our research objective. It is able to help course seekers to gain a quick insight including the positive and negative reviews into the courses as well as enables them to make a quick comparison between multiple courses. The quick insight can serve as a good reference for the course seekers and helps them to save time during the course selection process. The solution also can serve as a good platform for the course provider to study and understand their course strength and weakness. In this work, 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.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contribution of Dr. Gan Keng Hoon upon providing necessary guidance, knowledge, and support throughout this work.

REFERENCES

- [1] <https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017/>
- [2] https://spiegel.medill.northwestern.edu/pdf/Spiegel_Online%20Review_eBook_Jun2017_FINAL.pdf
- [3] <https://smallbiztrends.com/2017/04/importance-of-online-reviews.html>
- [4] <https://monkeylearn.com/sentiment-analysis/>
- [5] <https://textblob.readthedocs.io/en/dev/>
- [6] <https://pypi.org/project/vaderSentiment/>