A Preliminary Sentiment Analysis Study on the GMAT subreddit

Giovanni Angelini
nglgnn@gmail.com

ABSTRACT

A sentiment analysis has been performed on posts from Reddit.com and related to the GMAT exam. The comments with a positive connotation are identified to be the majority. While positive comments are arguably related to success in the score and bring less information on the challenges users are having with the exam. For this reason, I have studied in more detail the negative comments. Results shown a large interested in preparation material and an active market for preparation material and mock exams.

1 DATA SELECTION AND DATA MINING

I have performed the analysis in Python using the Reddit API to extract the information from the subreddit relative to the GMAT exam. The data set is composed by 1000 threads and about 5000 comments. The data sample cap is limited to the capability of the official Reddit API. Being able to overcome this limitation is possible but it is a time consuming activity that goes above the purpose of this simple exercise.

I have sort and cleared out the posts by removing external likes, images and non English comments.

For the analysis I have used the Natural Language Package NLTK for performing the sentiment analysis; specifically, I have used thee Vader algorithm [1]. In order to reduce type-II errors, I have decided to increase the threshold suggested for the algorithm from 5 percent to 20 percent. Doing so reduce the sample size but the sample is more likely to have less miss-identification.

I counted the frequency of all the words contained in the neutral and negative posts and computed their importance with a neural network so produce a map of words. I used this map to look into specific comments containing words that for me - as a non expert of the GMAT exam- were of some interest. Specifically, I have used the doc2vec algorithm available in the genesim package [2], to realize a word embedding from the comments. The algorithm trains a shallow neural network to assign a numerical value to words based on the context they appears. On addition to the embedded text, I vectorized the words in each comment and used a TF-IDF (Term Frequency- Inverse Document Frequency) metric to compute the number a word appears as well its relative importance in the comments setting a threshold of at least appearing in 10 different comments. I have then used these info as features in order to train a random forest with the goal of predicting the most relevant features that characterized the negative comments.By using the label produce by the Vader algorithm, I am introducing a strong bias in the random forest prediction. This issue in the current analysis can be easily overcome by manually read and assign a label to a large sample size.

[1] https://www.nltk.org/*modules/nltk/sentiment/vader.html*
[2] https://radimrehurek.com/gensim/

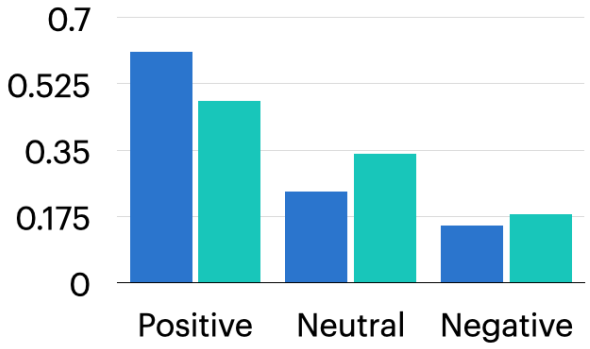

Figure 1: Dark blue bars: Percentage of posts labeled as Positive/Neutral/Negative. Light blue bars: Percentage of the comments to negative posts, labelled by the algorithm as: Positive/Neutral/Negative.

2 RESULTS

By using the sentiment analysis and plotting the amount of posts with a positive, neutral and negative comments it is clear how the negative comments compose a small fraction of the overall Reddit activity related to the GMAT, as well as most of the answers to negative comments have a positive connotation too (Fig.1). In order to have a nicer visualization of the most relevant word appearing in the posts, I have produced a word-cloud representation. The results for the positive, neutral and negative posts are shown respectively in Fig.2, Fig.3, and Fig.4. The positive and neutral comment don't show any particular word being recurrent, the cloud contains mostly not interesting words and it is clear to see the ID of some of the most active users appearing this is because the frequency of most of the words is mostly the same. However, the negative comments show a more interesting pattern: no id of users appears, meaning that there are questions having larger frequency that the number of posts itself, as well as there is a more emphasis of words related to the test itself as : questions, moc, test, prep, exam,score, quant, etc..

I have then looked into the comments to these negative posts. The most relevant words used in the answer to these posts are shown in Fig.5.

2.1 Negative comments

In addition to the negative posts I have looked into all the replies to posts considered negative and produced a word cloud for those comments (Fig.5) Finally, on the negative post and the relative replied, I have trained a random forest model using 80 percent of the sample size and used the remaining 20 as testing sample. The top 15 relevant features of the trained random forest are shown in table 1. In order to check the quality of the classification I have

Table 1: The 15 most important feature of the random forest trained model.

Importance of the feature	Feature
1	Word: mock
2	Word: wrong
3	Word: worry
4	Number of characters
5	Doc2Vector Vector2
6	Doc2Vector Vector4
7	Number of words
8	Doc2Vector Vector 1
9	Doc2Vector Vector 3
10	Doc2Vector Vector 0
11	Word: experience
12	Word: specific
13	Word: online
14	Word: miss
15	Word: score
...	...

Figure 2: Words cloud obtained from the post labeled as positive from the sentiment analysis.

Figure 3: Words cloud obtained from the post labeled as neutral from the sentiment analysis.

produce a ROC (Receiver Operating Characteristic) curve. As can see in Fig.6 the Area under the curve and the shape of the curve is not optimal, mostly due to the relative small sample size and the need of a better cleaning up of the comments. The random forest shows similar features to the word cloud.

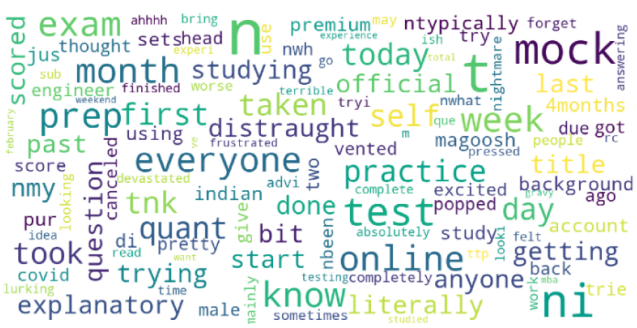

Figure 4: Words cloud obtained from the post labeled as negative from the sentiment analysis.

Figure 5: Word cloud obtained from the answers to negative post.

2.2 Results on negative posts

I have analyzed keywords of negative posts and comments that were attracting my attention. I printed the comments in which these words were appearing, and here is the summary of what I

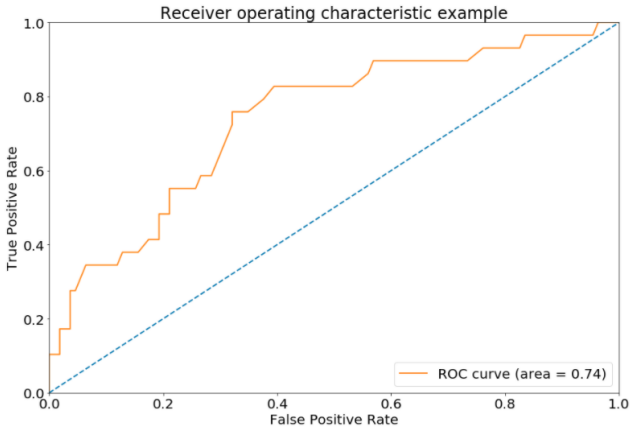

Figure 6: Receiver Operating Characteristic curve for the trained random forester classifier

have learnt:

TTP and Ninja:

These two words are highly correlated in the answers to negative feedback's to the GMAT. By printing some of the comments containing these keywords I have noticed that many Reddit users suggest people to utilize TTP GMATNinja as a successful strategy to preparation for the exam. The frequency of these keywords is significantly higher once we select post that regard low score results.

Mock:

The word mock doesn't appear to be in the cloud of the most frequent word of the negative comments, however it is a main word in the negative posts. The word appears mostly in posts of users complaining how mock tests are easier than the real exam.

Club:

The world Club refers to the GMAT Club in the majority of comment. All the comments I have looked are suggestion to users to use GMAT Club especially for Quant.

Magoosh:

Magoosh appear mostly as suggestion for Quant preparation, however most users suggest to pass to TTP.

Online:

The world online is always associated to the online version of the exam, it seems that many people have problem with the online version or for focusing reason or because they know less the rules related to it. They have noticed they score higher on the in person than online version.

A smaller fraction complained about technical issue on their laptop as freezing screen.

email:

About half of the world containing the word email refer to user asking email address of other users. However some comments refer to not nice email from the GMAC accomodation department. Some comments were saying to swhich to GRE because is not convenient to get scheduled with GMAC.

3 CONCLUSION

This simple preliminary analysis done on the sub-reddit GMAT forum, shows how the majority of the comments on the forum are positive. Analyzing the negative comments it is clear that the majority of those are composed by suggestions on how to improve scores using third-party tutoring and studying material. A small fraction of negative concern the performance of users on online version of the test.