A Match Made in Twitter: Exploring Sentiment on the Residency Matching Process with Twitter

Richelle LeMay
CIS 731 Fall 2021
Kanas State University

Abstract

The medical residency application process in 2022 will be complicated by a change in the associated standardized test and also by the impact of COVID. The purpose of this paper is to determine the overall opinion of the current process by performing sentiment analysis on potentially related and unlabeled twitter data. Twitter text data poses a unique challenge for machine learning, given the short length.

1 Introduction

Medical students and residency programs are paired by a match process. Medical students apply for residency by providing a ranked list of programs, and programs provide a ranked list of applicants; however, the number of applicants typically exceeds the available residency positions at a given program. The highest-ranked student and program pairs are matched for these available positions. This information is made available to the student in March on what is known as Match Day. Students began submitting applications for Match Day 2022 on September 1, 2021. Residency programs began reviewing submissions on September 29.

In the past, a 3-digit score on the United States Medical Licensing Examination (USMLE) test was used by many residency programs to filter out students. More prestigious specialties and schools set higher internal thresholds for those scores. This was criticized due to known issues with standardized testing and a handful of other concerns. The USMLE is now moving to a pass/fail system, beginning with the 2022 match.

In addition to that change, COVID has limited opportunities for students to participate in clinical rotations at programs they wish to attend. This combination of factors has complicated Match Day 2022, in part, by encouraging students to diversify risk by applying to more residency programs. This increase has been called "Application Fever."

The goal of this project is to glean insight on opinions associated with these changes by performing a descriptive task using the pre-trained Vader sentiment analyzer on related twitter data, which will require removal of noise from the collected data as well partially annotating classification by direct observation.

2 Background and Related Work

Two works have been found that discuss different methodologies to remove noise from twitter data. Wijeratne and Heravi (2014) proposed a solution to filtering noise by removing any tweets with a context that falls outside the set of desired word senses from a list of keywords. Godfrey et al (2014) removed noise from tweets by first comparing the results of k-means clustering, DBSCAN, and a combination of those two methods. They then filtered out tweets that did not meet inclusion standards of at least two of the three methods.

Farooq et al (2015) used word sense disambiguation for sentiment analysis on product reviews. Gupta et al (2017) surveyed many available methods in Python for sentiment analysis on twitter data. These included extensive preprocessing, feature extraction, and eight different models. Whereas Band (2021) provided an example of using the Vader package in Python to perform sentiment analysis on twitter data.

3 Dataset

A total of 158,395 tweets from October 3, 2021 through October 9, 2021 were compiled by filtering those that contained "aamc," "applicationfever," "bcarmody," "eras," "medicalresidency," "medstudents," "medtwitter," "match2021," "match2022," "usmle," "residency," and "resident." The highest volume of tweets was collected on October 4. The justifications for the words contained in the filter list are found in Appendix I.

The data was compiled in 14 JSON documents, which were then collected into a large list to allow for selection of elements in the nested JSON structure. These elements include the tweet id, tweet timestamp, tweet text, user screen name, user follower count, user friend count, user full name, and user language. The original intent was to use the additional elements in feature extraction; however, only the tweet text was used in this analysis.

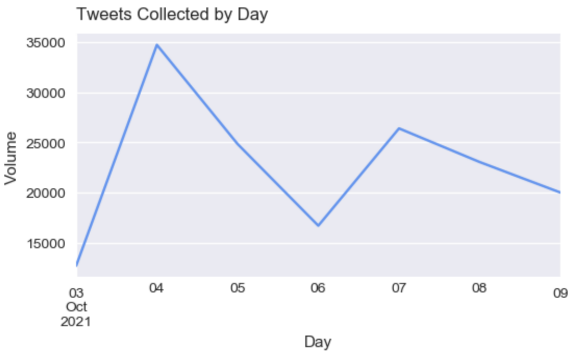


Figure 1: Total tweets collected by day

Several events occurred over the week of October 3, 2021 that impacted this dataset due to the list of words used to filter tweets:

- The trailer for the new *Resident Evil* film came out on October 7.
- An actor on *The Resident* discussed her departure from the show with Deadline on October 5.
- In the first week of October, it was rumored that Adele was considering locations for a Las Vegas residency.

Tweets regarding the topics above were prevalent in this dataset.

4 Methodology

4.1 Data Preprocessing

The heaviest lift with this dataset was the cleaning and filtering associated with preprocessing. Given the wide range of words used to capture relevant tweets, as mentioned above, a large portion of the dataset was not applicable to the topics associated with the medical residency application process. In order to perform sentiment analysis, this unrelated data needed to be removed, which was the core task of this project.

Prior to any modeling, the data was compiled in a PySpark DataFrame. Non-English tweets and retweets were removed using a filter, and the tweet text was converted to lower case. Then any tweets that contained “resident evil” were removed also using a filter. The exclusion of non-English tweets, retweets, and tweets that contained “resident evil” decreased the dataset from 158,395 rows to 22,763 rows.

Monotonically increasing IDs were then associated with each tweet, and binary flags were created for the existence of words in Appendix I as well as the words “student,” “doctor,” “match22,” “interview,” “applicant,” and “program” in each tweet. The additional words were added based on exploratory analysis, as is seen in section 4.2.

Because “resident” and “residency” were the most prevalent words that were least clear in definition, the Lesk algorithm was used to find their word sense within the tweet text. A binary flag was created to indicate if either word was used in relation to medical residents or medical residency.

The cleaning process, addition of monotonically increasing IDs, and application of the Lesk algorithm took 0.678 seconds to complete using PySpark.

The PySpark DataFrame was then converted to a Pandas DataFrame due to challenges with cleaning the tweet text to be used in additional exploratory analysis and feature extraction. From the tweet text, punctuation was removed, stop words were removed, words less than three characters were removed, “@username” mentions were removed, websites were removed, and the remaining words were lemmatized. Alphanumeric words were not excluded in feature extraction given the importance of the hashtag “match2022” in this dataset; however, these words were removed in exploratory data analysis.

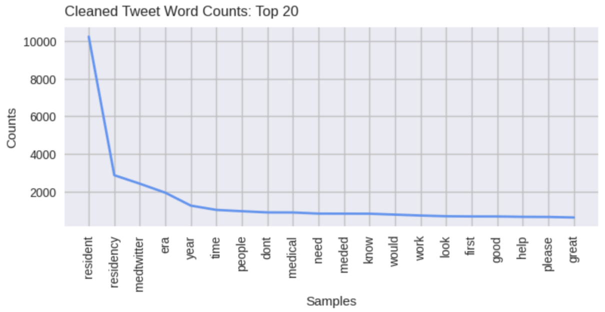


Figure 2: Word Frequency

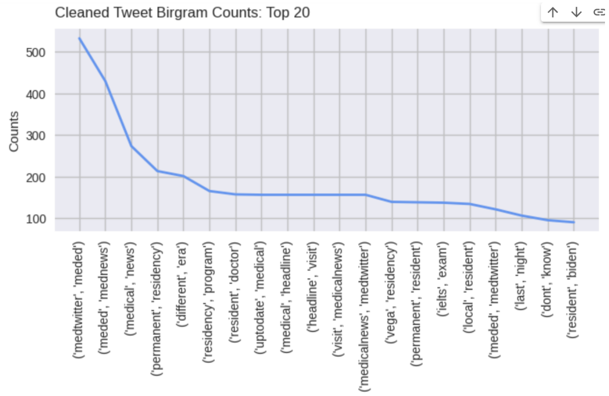


Figure 3: Bigram Frequency

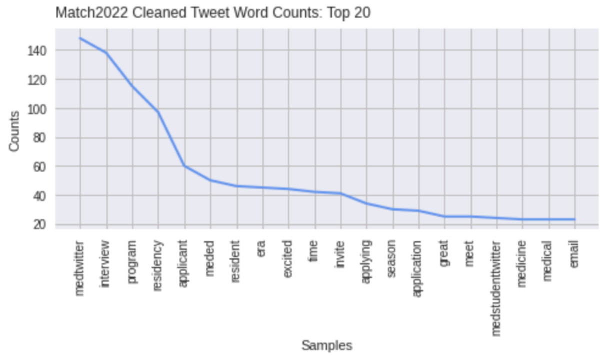


Figure 4: Match2022 Word Frequency

4.2 Exploratory Analysis

Frequency distributions of single words and bigrams were created on the cleaned tweet text. As expected, the words "resident" and "residency" were the most prominent. Although the single word frequency distribution does not lend much insight into the overall topics contained in our data, the bigram frequency shows expected medical topics.

The bigram "vega residency" is a result from the tweets speculating Adele's Las Vegas residency. The IELTS exam is an exam for those who wish to work or study in a country where the primary language is English. This bigram is likely a result from tweets discussing exams required for residency in another country. The "reisdent biden" bigram frequency comes from tweets by insurrectionists referring to President Biden. The word frequency in tweets that contained "match2022" were also investigated, as these tweets were the most likely to cover our topic in question. In addition to the words used to initially filter tweets, interview, program, and applicant were also prevalent.

4.3 Models

The intent of this project was to develop a non-model based rigid filter and a clustering filter, then compare the sentiment results. The original plan was to construct a consensus matrix based on consecutive n clusters in a KNN model using the cosine distance between the tf-idf of the tweets in the data set as was done in Godfrey et al (2014). However, the dataset proved too large to construct a consensus matrix within a reasonable time. Godfrey et al (2014) also used the DBSCAN clustering algorithm in their analysis to remove noise, which then became the new model of focus in this project.

However, DBSCAN had several issues during this particular implementation. First, DBSCAN expects clusters to be similarly dense, which does not appear to be the case in this dataset. Second, of the 22,763 total tweets, only 426 contain "match2022" in the text. This indicates that tweets relating to the medical residency placement process make up approximately 2% of the gathered tweets. As a potential solution, DBSCAN was used on the rigidly filtered data, in the hopes this would flag any unrelated tweets as noise.

Given the issues with the application of DBSCAN on this dataset, a small portion of tweets were manually labeled to allow for the use of the KNN and SVM classification models as well.

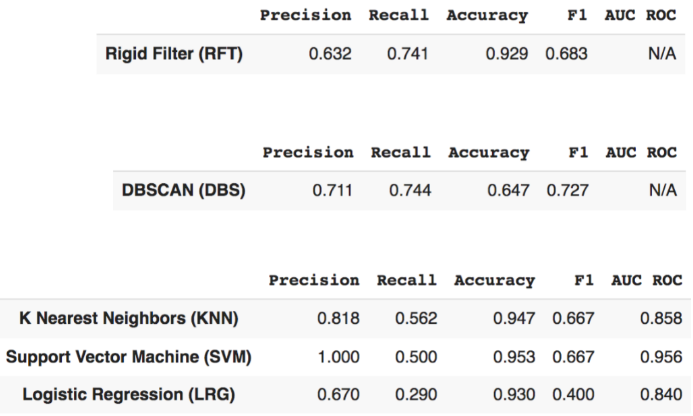
The DBSCAN pipeline used the cosine distance of the TF-IDF of a sample of half of the rigidly filtered and cleaned tweet texts to produce labels, whereas the KNN and SVM classification pipelines were trained on the TF-IDF of 70% of the cleaned and labeled tweets texts to produce binary labels.

5 Evaluation

As discussed previously, these data had a high imbalance of positive to negative labels; therefore, both precision and recall were important evaluation criteria. As such, the F1 metric was used to compare all filters. However, the classification based filters were also evaluated using the AUC ROC.

Tweets that contain “match2022” were used to measure F1 and AUC ROC against the unlabeled data, effectively becoming the positive label. This caused some issues with the metrics. Specifically, identified false positives may have been true positives if the tweet itself was about our topic in question, but did not contain the “match2022” positive label. This likely decreases precision.

The goal was to produce a filter with an F1 score greater than 0.60; if a classification model was used, the goal was to produce an AUC ROC greater than 0.80.



	Precision	Recall	Accuracy	F1	AUC ROC
Rigid Filter (RFT)	0.632	0.741	0.929	0.683	N/A

	Precision	Recall	Accuracy	F1	AUC ROC
DBSCAN (DBS)	0.711	0.744	0.647	0.727	N/A

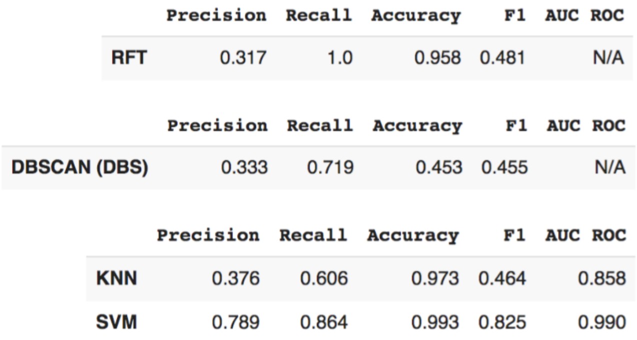
	Precision	Recall	Accuracy	F1	AUC ROC
K Nearest Neighbors (KNN)	0.818	0.562	0.947	0.667	0.858
Support Vector Machine (SVM)	1.000	0.500	0.953	0.667	0.956
Logistic Regression (LRG)	0.670	0.290	0.930	0.400	0.840

Figure 5: Evaluation Results of Labeled Data

6 Results

The results of the rigid filter (RFT), DBSCAN on rigid filtered data (DBS), KNN classification (KNN), SVM classification (SVM), and Logistic Regression (LRG) of the labeled data can be seen in figure 5.

All but the LRG filter meets the criteria of an F1 score greater than 0.60, and all of the classification filters have the same AUC ROC scores above our criteria of 0.80. The DBS clustering model and the two classification models were contenders for the final filter. However, the LG filter will no longer be considered given its poor F1 score.



	Precision	Recall	Accuracy	F1	AUC ROC
RFT	0.317	1.0	0.958	0.481	N/A

	Precision	Recall	Accuracy	F1	AUC ROC
DBSCAN (DBS)	0.333	0.719	0.453	0.455	N/A

	Precision	Recall	Accuracy	F1	AUC ROC
KNN	0.376	0.606	0.973	0.464	0.858
SVM	0.789	0.864	0.993	0.825	0.990

Figure 6: Evaluation Results of Unlabeled Data

The accuracy score for most of the filters is relatively high. This is likely a result of the dataset being largely imbalanced. The accuracy seen with the DBSCAN filters decrease as the data is pre-filtered, and as such, is less imbalanced.

These filters were next applied to half of the available tweets and compared against the existence of “match2022” in the text as a label. Half of the available tweets were used, as the DBSCAN algorithm struggled with the larger corpus size when using the full dataset.

In most cases, precision decreased, recall increased, and the F1 score decreased. The SVM model on unlabeled data fit both criteria of an F1 score greater than 0.60 and an AUC ROC greater than 0.80. The SVM model predicted the smallest volume of positive labels, whereas the rigid filter predicted the most.

The final step was comparing the Vader sentiment analysis of tweets that were associated with the medical residency placement process for unfiltered data, rigidly filtered data, and the trained SVM model.

Although the average Vader sentiment score for all filters was positive, Figure 8 demonstrates that the more refined the filter was, the higher the Vader sentiment score was. The unfiltered data contained the highest volume of data, whereas the SVM filter contained the least. It appears the

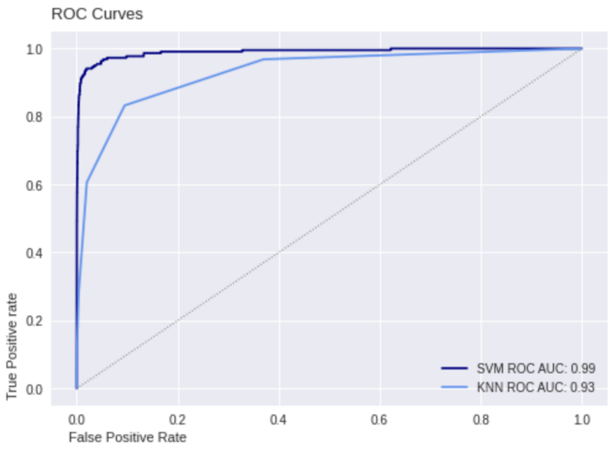


Figure 7: ROC AUC of KNN and SVM Models

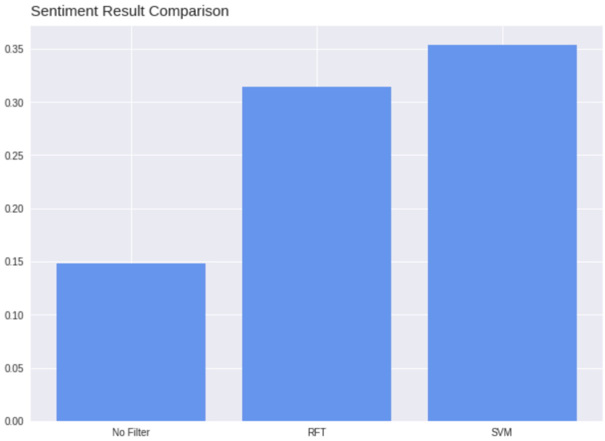


Figure 8: Sentiment Comparison of Filters

overall sentiment seen in tweets regarding the medical residency placement process were positive; however, there exists no baseline for comparison. This result in sentiment may or may not be a change from prior years; however, this process could be repeated in upcoming years to continue to measure the impact of the changes.

It is possible the increase in sentiment seen with an increase in filter refinement is a result of the types of tweets that are within the topic in question. Many of the identified tweets were from either student introductions or program introductions. Both students and programs are, in a sense, advertising to each other. Such tweets would use positive language.

7 Conclusion

I had high hopes for this project that were challenged by my limited experience and time constraints. However, I appreciated the exposure to many new topics in my search for an appropriate model for unlabeled short-text analysis methods. Attempts at new-to-me methods, such as the clustering consensus matrix and label propagation, were equally fun to discover and disappointing to implement. As with any good class, I have learned that there is so much more to learn.

It might have been beneficial to explore the team option, had I felt more comfortable in approaching partners. The opportunity to have other classmates as a sounding board may have alleviated some of the struggles with discovering new methods or talking through the issue presented by the lack of comparative data to assess the sentiment analysis results.

If given the opportunity, I would explore using additional fields as features. The sentiment results appear to indicate that the Vader score could be used as a feature, and additional twitter data may well be useful in features as well.

Future work for this project would also include tracking the sentiment score as Match Day 2022 approaches. There may be trends in sentiment during weeks with more student interviews or program open houses. It may also be insightful to see if there are any changes in sentiment associated with student specialty as more challenging program specialties have had higher USMLE thresholds.

8 References

Band, A. (2021, April 11). T*witter Sentiment Analysis using Vader & Tweepy | Python in Plain English.* Medium. https://python.plainenglish.io/twitter-sentiment-analysis-using-vader-tweepy-b2a62fba151e

Bedi, G. (2019, November 9). *A guide to Text Classification(NLP) using SVM and Naive Bayes with Python.* Medium. https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

Farooq, U., Dhamala, T. P. , Nongaillard, A., Ouzrout, Y., & Qadir, M. A. (2015, December) *A*

Word Sense Disambiguation Method for Feature Level Sentiment Analysis. 9th IEEE International Conference on Software, Knowledge, Information Management & Applications, Kathmandu, Nepal. 10.1109/SKIMA.2015.7399988.

Godfrey, D., Johns, C., Meyer, C.D., Race, S., & Sadek, C. (2014). *A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets*. ArXiv, abs/1408.5427

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). *Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python.* International Journal of Computer Applications, 165(9), 29–34. 10.5120/ijca2017914022

Moon, Kristen. "USMLE Step 1 is Now Pass/Fail - Who Benefits From This Big Change?" *Forbes*, Forbes, 7 April 2020, https://www.forbes.com/sites/kristenmoon/2020/04/07/usmle-step-1-is-now-passfailwho-benefits-from-this-big-change/?sh=4f5ff61b4873.

Wijeratne, S., & Heravi, B. (2014, September). *A Keyword Sense Disambiguation Based Approach for Noise Filtering in Twitter*. 1st Insight Student Conference, Dublin, Ireland. 10.13140/2.1.4197.0565.

Appendix I. Hashtag and Mention Justification

aamc - The Association of American of Medical Colleges

applicationfever - Hashtag used to describe increased medical residency applications

bcarmody - One of the foremost experts/critics on the Match process

eras - Electronic Residency Application Service

medicalresidency - Hashtag used by prospective residents

medstudents - Hashtag used by medical students

medtwitter - Hashtag used by medical professionals

match2021 - Hashtag used for Match Day 2021

match2022 - Hashtag used for Match Day 2022

usmle - United States Medical Licensing Examination

residency - Hashtag may pull in tweets associated with medical residency

resident - Hashtag may pull in tweets associated with medical residency