Examining the Role of Essay Questions in OKCupid

Ofek Peres
Department of Mechanical Engineering
Princeton University
operes@princeton.edu

Divyanshu Pachisia
Department of Mechanical Engineering
Princeton University
divyanshupachisia@princeton.edu

Abstract

The biggest challenge for dating apps is to categorize their users into groups in order to match them. This clustering is heavily influenced by the questions asked during profile creation. OKCupid is unique in that it asks it's users lengthier questions, referred to as "essays questions". This paper examines how these essay questions influence clustering of the users, compared to the more traditional non-essay questions. For both K-Means clustering and Latent Dirichlet Allocation (LDA), we find that the non-essay questions achieve, on average, a 90.8% higher silhouette score but produce one large, isolated cluster that encompasses the vast majority of the users, skewing results. On the other hand, the essay questions produce a more uniform splitting of the users, but closer clusters. Overall, the essay questions are shown to be more useful to the dating app because of their ability to produce more uniformly split clusters that may be helpful in matching.

1 Introduction

The dating application market is anticipated to be worth $3.2 billion by 2020 [1]. The market is becoming increasingly crowded and a key factor in differentiating dating apps is their matching algorithms. One approach to matching people is to categorize them into clusters by personality and then match either within or between clusters. To produce meaningful clusters, dating apps have the ability to construct profiles based on the questions the apps ask users. OKCupid has approached this challenge by including more qualitative responses, including 10 "essay questions".

In this paper, we analyze the impact of these essays questions on clustering by comparing them to the non-essay (more traditional multiple choice) questions asked. We hypothesize that the essay questions help elucidate differences in people better than the traditional multiple-choice questions.

2 Related Work

Finding differences in personality within users in a dating app may be particularly useful as dating apps seem to attract outwardly similar people [4] [3]. Whitty describes how the people on dating apps are coming from similar working backgrounds with lots of travel time and little time to form interpersonal relationships. This could indicate that the multiple choice questions asked by dating websites could lead to poor clustering of these seemingly similar individuals. It is important to find the key differences between users to characterize them as unique individuals because complementary individuals are likely to form stronger interpersonal relationships [2]. Responses to longer essay questions delve deeper into personality and may provide this much needed information for OKCupid and other dating apps.

3 Methods

An initial exploration of the data was undertaken in order to better study the OKCupid dataset. It was noticed that there were a number of unique features in the profiles as compared to typical dating apps profiles. Most notably are the 10 essay questions that are asked of users. As such, it was decided to explore the unique aspect of OKCupid, the essays, separately from the typical data collected by dating apps. The data was preprocessed to make it usable in clustering analyses and then evaluation metrics were used to evaluate the clustering produced.

3.1 Data Processing: Non Essay Questions

1. **Location and Last Online:** These features may be useful in matching in practice based on convenience but were decided to not be representative of the user's personality, especially because all of the locations were in the Bay Area. Therefore, these two features were dropped in order to analyze and cluster based on personality.
2. **Star Sign:** Rather than store the specific star sign of every individual, the more telling characteristic that OKCupid collected was the level of importance of the star sign to the user. Therefore, in our analysis we used only this level of importance.
3. **Languages Spoken:** Similar to the star sign, rather then store each of the languages spoken by a given user, the important information that OKCupid collected was the *Number of Languages Spoken*. Therefore, the feature was converted to the count of languages spoken.
4. **One Hot Encoding** All non-numerical columns, such as education, job, and status, were converted into binary data through the use of one hot encoding, where each distinct category in a column is converted to its own binary column.

3.2 Data Processing: Essay Questions

The essays were first cleaned by removing punctuation, noisy characters and stop words, to produce a list of the important words in the essay. Normally, a bag of words approach is used to then encode these words; however, due to computational costs, it was decided to instead gather important characteristics of the essays. The following characteristics were used:

1. **Sentiment:** The sentiment of each essay was calculated using python's natural language processing toolkit, Sentiment Intensity Analyzer. [8] Sentiment is a very important feature as the way an essay is phrased is indicative of how a person thinks (e.g. in positive, neutral or negative terminology).
2. **Term Frequency - Inverse Document Frequency (TF-IDF):** TF-IDF scores measure the importance of a word in a datset normalized by how often it is used. This was decided to be an important characteristic as a person who used unique words in their essays relative to the the rest of the users' essays was going to stand out to other users.
3. **Length:** The length of each essay was determined as an important characteristic as it is informative about how invested in sharing information about themselves to find a match a user was.

3.3 Clustering Methods

1. **Latent Dirichlet Allocation (LDA):** Described in section 3.5

   Hyperparameters: Number of Clusters (K), Prior of topic distribution (α) which was 0.1 after tuning with cross validation to maximize silhouette score.
2. **K-Means Clustering**: Finds k centroids in the dataset that best maximizes distance between centroids.

   Hyperparameter: Number of Clusters (k)

In the dataset 20% of data was held out to finally test the performance of our clustering. Using the 80% training set, the hyperparameters were tuned using scikit learn's **GridSearchCV**. GridSearchCV cross-validated our data set (using 5 folds) and reported the best hyperparameter(s) based on the evaluation metrics described in 3.4. This cross-validated model was then used to cluster the held-out data.

3.4 Evaluation Metrics

1. **Silhouette Score**: This measures homogeneity within a cluster and heterogeneity between clusters. A positive silhouette score indicates that the points within a cluster are close together but are far from other clusters, while a negative score indicates overlap. The silhouette score is calculated for each data point using Equation 1, where a measures average distance within a cluster and b measures average distance to all points in any *other* cluster.

$$S = \frac{b-a}{max(a,b)} \tag{1}$$

2. **Log Likelihood as Score**: For LDA, scikit learn has a score function that measures the score as log likelihood. This is described in Section 3.5.

3.5 One Method in Detail: Latent Dirichlet Allocation

LDA is part of a class of algorithms in unsupervised learning called **topic models**, which find topics, or groups, in unlabeled data. In this paper, we utilize LDA to find the user groups that each user belongs to based on their features.

Figure 1, from Hoffman et. al. (2013) [5], shows the graphical model governing LDA, where, for our purposes:

1. D : collection of D users
2. N : sequence of N features (either essay related or non-essay related)
3. K : The number of user groups, or clusters, which was tuned using cross validation to produce best performance. This is specified in advance.

The only data that is observed is the shaded circle, $w_{d,n}$ which is the features (traditionally words) that we observe.

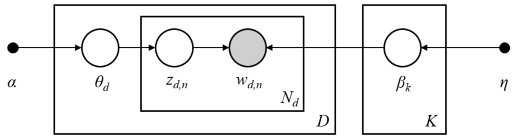


Figure 1: LDA Theory - Block Diagram

The algorithm utilized in LDA is as follows: [6]:

1. For each user group, $k \epsilon K$, draw $\beta_k \sim Dirichlet(\eta)$, where η was set to the default value of $1/D$. This describes the probability distribution with which each feature, $w_{d,n}$ may belong to a particular user group, k.
2. For each user, $d \epsilon D$, draw $\theta_d \sim Dirichlet(\alpha)$, where α was tuned using GridSearchCV to produce the best performance. This describes the probability distribution with which each user, d may belong to a particular user group, k.
3. For each feature, $w_{d,n}$, in user d:

   Draw the assignment to the user group, $z_{d,i}$ based on step 2. As the Dirichlet is a **conditional prior** to the multinomial distribution, this will be $z_{d,i} \sim Multinomial(\theta_d)$

   Based on the assignment to the user group, we then draw an observed feature. Based on step 1, this is characterized as $w_{d,n} \sim Multinomial(\beta_k)$

The three steps above correspond to finding the joint distribution of hidden and observed variables, as shown in Equation 2 [7].

$$p(\beta 1 : K, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \tag{2}$$

Using the generative process described above, we find that conditional distribution of the user group structure (the *posterior*) is as in Equation 3 [6].

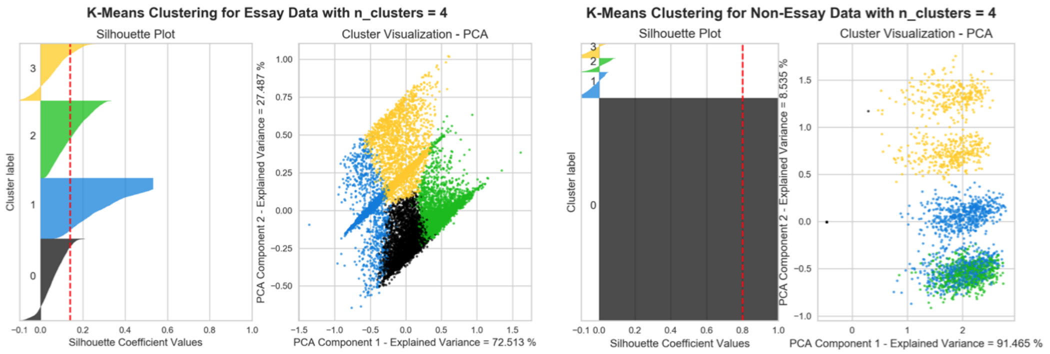
$$p(z, \theta, \beta | w, \alpha, \eta) = \frac{p(z, \theta, \beta | \alpha, \eta)}{p(w|\alpha, \eta)} \tag{3}$$

This posterior is intractable as the possible number of user group structures is exponentially large [7]. We utilize the inbuilt variational Bayesian method in scikit learn [6] to estimate the posterior. The **log likelihood** of this estimate on unseen users was used to score the performance of the distribution. Finally, to cluster users into groups, the group with the maximum likelihood was used for each user.

LDA seemed appropriate to use for this task, as it makes the assumption that the order of the users and features does not matter, which is true for our dataset. However, a key flaw in this approach is that LDA categorizes users into how well they fit into every group. However, for clustering purposes the group with maximum fit was picked which may not accurately represent the reality of users who may fit into more than one group.

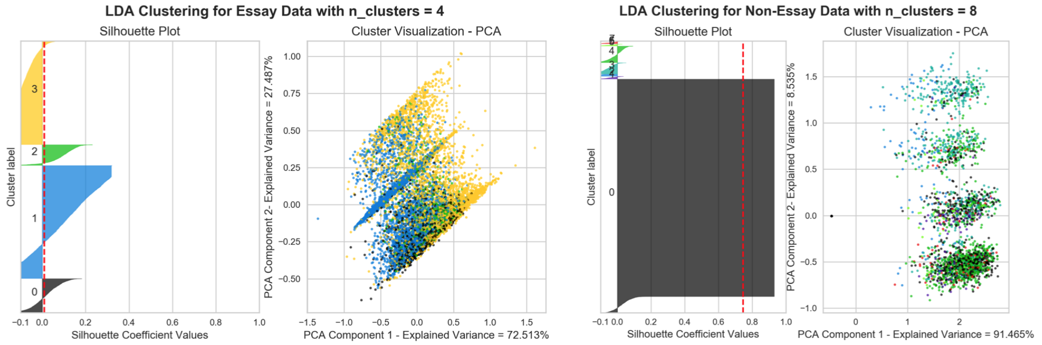
4 Results

The results show the clusters of users in the held-out data. Figures 2 and 3 compare the clusters for the users based on essay data and non-essay data, using K-Means clustering and LDA respectively, along with a visualization of the silhouette score for each data point. Principal Component Analysis was used to project the data down to the 2 dimensions of largest variance for visualization. These two dimensions are weighted linear combinations of features, as Figure 4 shows.



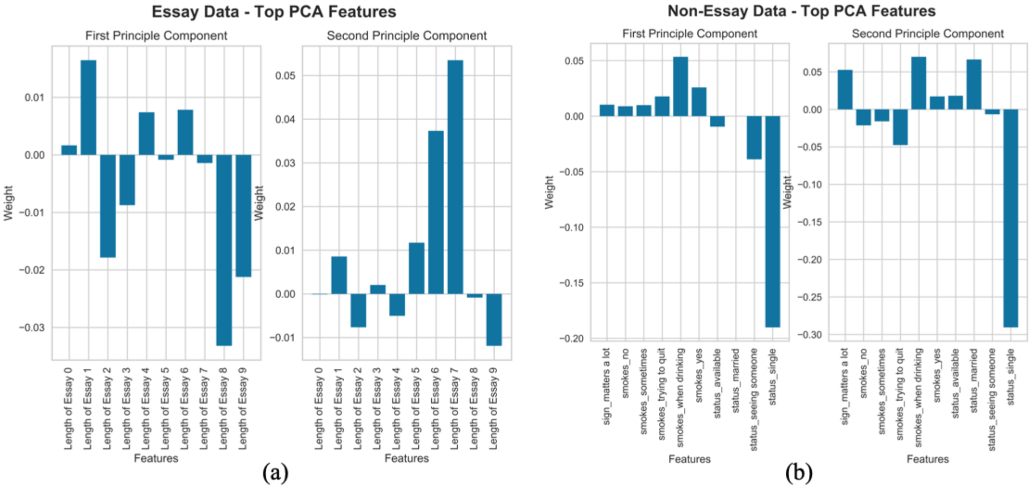
(a) (b)

Figure 2: K-Means Clustering Analysis



(a) (b)

Figure 3: LDA Clustering Analysis



(a) (b)

Figure 4: PCA Analysis

5 Discussion and Conclusion

5.1 Analyzing Clustering Usefulness

As Figures 2 and 3 in Section 4 illustrate, there was a large difference in clustering between the essay data and the non-essay data for both K-Means clustering and LDA. An interesting note is that LDA did a worse job finding distinct clusters overall, which can be explained by the fact that it attributes each user to more than one cluster, as discussed in Section 3.5.

When only **non-essay** data was considered, 81% of users in the test section fell into cluster 0, as seen in row one of Table 1. The silhouette score analysis for the non-essay clustering demonstrates a large positive score of 0.8, meaning that these users fit into the cluster they were assigned to well and are relatively far from fitting into the other available clusters. The majority of users fitting into one cluster highlights a concern discussed in a number of related works: that many of the users coming to dating sites will seem similar on the surface level, as captured by the non-essay data. Therefore, the high silhouette score may be misleading given the goal of the dating site in finding fine-tuned differences within similar groups for matching.

For the essay questions we see the opposite: a more uniform splitting of users that comes at the cost of less separation between clusters. This result shows that the essay questions *do* successfully promote user uniqueness and diversity, characteristics that would create a more uniform splitting of users into clusters. This comes with a trade-off of having more fluid clusters that do not have as sharp boundaries. This could result in some users being mischaracterized. However, the fluid nature of these clusters is perhaps a more accurate representation of people and so may be more useful compared to having one isolated cluster that most people fall into.

5.2 Characterizing the clusters

In order to find the characteristics of each cluster, the top ten weighted features in the two principle components were taken, displayed in Figure 4. Based on the position of the cluster on the plot, some inferences were made on the characteristics of each cluster, shown in Table 1.

Overall, the clusters in the non-essay data are not only concentrated in one cluster, but also have generic characteristics. For example, the largest weighted feature in the non-essay data set was being single. This feature is not a very interesting or effective way to cluster potential dating partners into finer categories to produce the complementary partners examined in Section 2. On the contrary, the essay data provided more uniform clustering with more sophisticated characteristics. For example, Cluster 1 corresponds to those people willing to share information on Essay 8 that asks them about an embarrassing secret. This tells us a lot more about the person and so matches based on this feature may allow for better matching and therefore a more successful dating application.

It must be noted that these clusters are characterized just based on the top 10 weighted features in the PCA components, but in reality may have many more characteristics not examined due to their lower importance (smaller weights).

	Non-Essay Data				Essay Data			
	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Proportion	**81%**	12%	4%	3%	21%	23%	30%	26%
PCA Component 1	Negative	Positive	Positive	Positive	Neutral	Negative	Positive	Neutral
PCA Component 2	Neutral	Neutral	Negative	Positive	Negative	Spread	Neutral	Positive
Description	Single	Smokes	Smokes, Single	Smokes, Star Sign Matters	Long Essay 9, "message me if..."	Long Essay 8, willing to share secret	Long Essay 1, what I'm doing with my life...	Long Essay 6, spends time thinking about...

Table 1: K-Means Clustering PCA Explanation

5.3 Concluding Remarks

The essay questions asked by OKCupid do a better job in making more evenly distributed clusters by examining more personal characteristics, but lead to less distinct clusters. This is especially useful in differentiating users who appear similar in the non-essay data and therefore could lead to better matches. To further explore this finding, the essay data could be characterized in a more detailed manner by utilizing a bag of words approach as well as finding other characteristics of essay data. Additionally, a model could be built by combining the essay and non-essay data to improve the poor silhouette scores generated by the essay data.

Acknowledgments

Professor Barbara Engelhardt
Jonathan Lu
Diana Cai
The SciKit Learn library which was used for all the code for this assignment

References

[1] https://blog.marketresearch.com/american-singles-fuel-the-2.5-billion-dating-market

[2] Dryer, D. C., & Horowitz, L. M. (1997). When do opposites attract? Interpersonal complementarity versus similarity. Journal of Personality and Social Psychology, 72(3), 592-603.

[3] Brym, Robert & L. Lenton, Rhonda. (2001). Love Online: A Report on Digital Dating in Canada.

[4] Whitty, Monica, The Joys of Online Dating,Chapter 12: Mediated Interpersonal Communication

[5] Hoffman, Matthew et. al. (2013) Stochastic Variational Inference

[6] https://scikit-learn.org/stable/modules/decomposition.html#latentdirichletallocation

[7] David M. Blei. 2012. Probabilistic topic models. Commun. ACM 55, 4 (April 2012), 77-84. DOI: https://doi.org/10.1145/2133806.2133826

[8] https://www.nltk.org/api/nltk.sentiment.html