BTP PRESENTATION – PHASE II

RELIABILITY DIMENSION IN REVIEW PLATFORM YELP

Submitted By:
Tarun Genwa (150123043)
Akash Mahalik (150123004)

Under Guidance of :
Dr. Ayon Ganguly

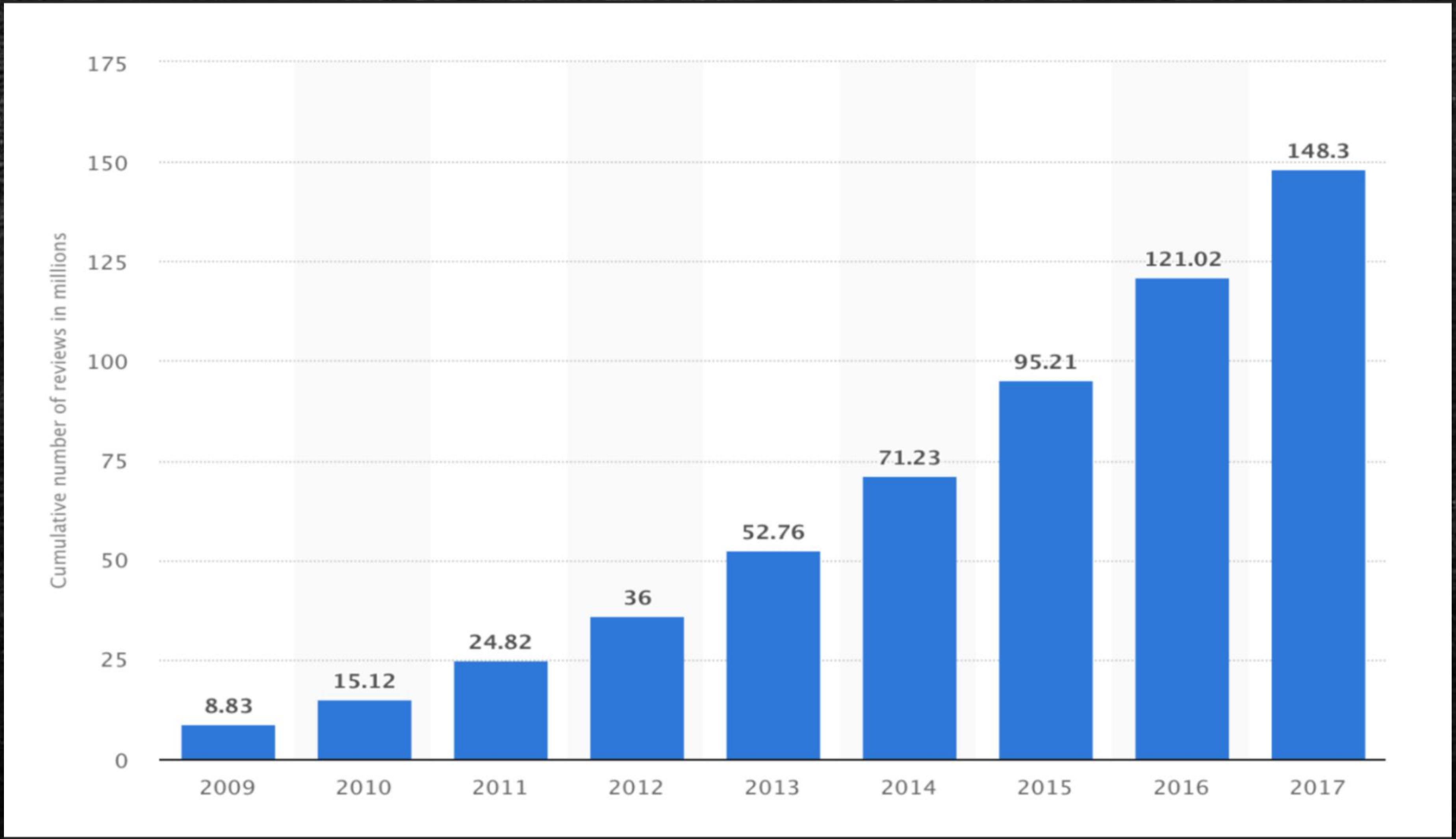
What if Reliability Analysis of Big Data Systems ?

- Reliability is the degree of belief as to what extent a system will work as it is expected to.
- With the advent of big data systems, containing massive volumes of data which are piling up at a high velocity, a lot of meaningful insights can be driven from the same. But an important remains as to how reliable these systems really are in order to use them for all the heavy analysis is being done these days.

Choosing Yelp as the Big Data source

- Yelp is crowd sourced review platform listing hundreds of thousands of local and chain businesses.
- The motivation for choosing Yelp dataset as our big data source was due to the effect that it carries on mass opinion and hence comes out to be a role player in business growth.
- The availability of massive datasets open sourced by Yelp itself for running various dataset challenges.

Total Reviews on Yelp (Plot)



Constituents Of Reliability Of Rating Model Of Business

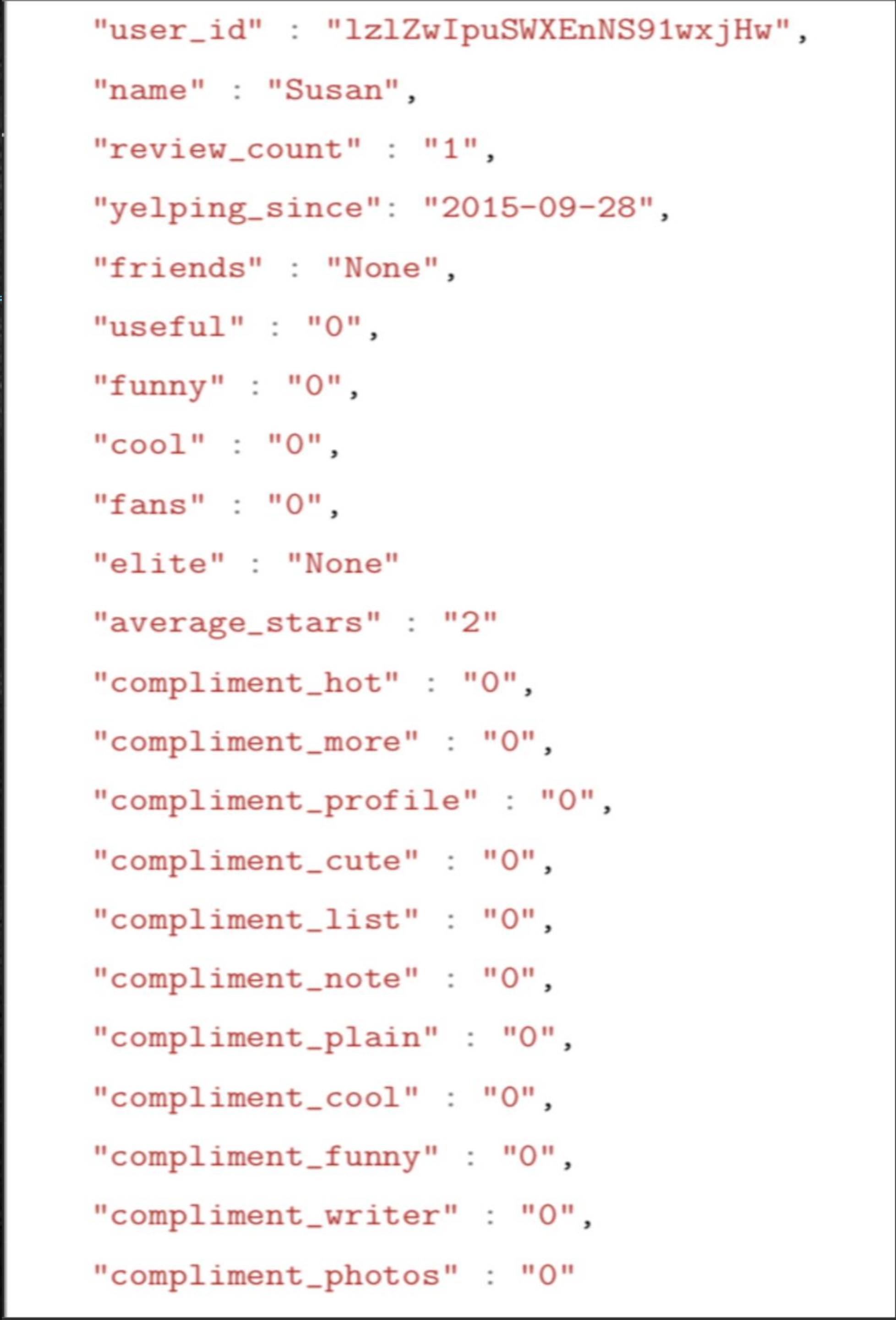
- User Credibility
- Topic Modelling of Review Text
- Sentimental Analysis of a Review and its Rating

Creating a Credibility Score of a User

- Credibility score of a user will help in deciding how relevant a user’s review is.
- It would help in taking decisions while filtering fake and biased reviews.
- It also helps in understanding the eliteness levels of different users.

Approach to Credibility ??

- Credibility score is defined by various features in user dataset.
- The dataset already labels a user as elite or not.
- We can leverage machine learning techniques to assign feature importance and decide eliteness of a user.

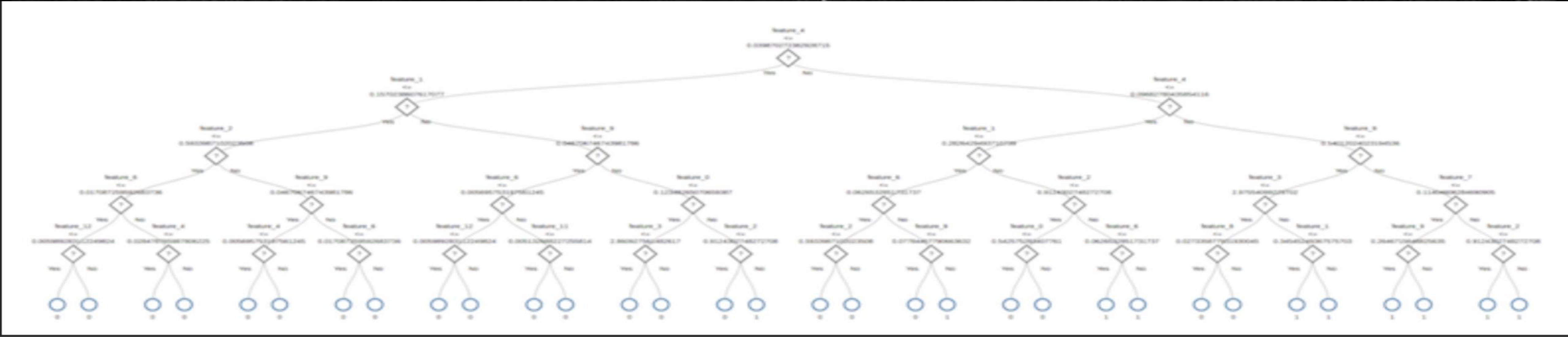


```
"user_id" : "lzlZwIpuSWXEnNS91wxjHw",
"name" : "Susan",
"review_count" : "1",
"yelping_since": "2015-09-28",
"friends" : "None",
"useful" : "0",
"funny" : "0",
"cool" : "0",
"fans" : "0",
"elite" : "None"
"average_stars" : "2"
"compliment_hot" : "0",
"compliment_more" : "0",
"compliment_profile" : "0",
"compliment_cute" : "0",
"compliment_list" : "0",
"compliment_note" : "0",
"compliment_plain" : "0",
"compliment_cool" : "0",
"compliment_funny" : "0",
"compliment_writer" : "0",
"compliment_photos" : "0"
```

Random Forest Classifier

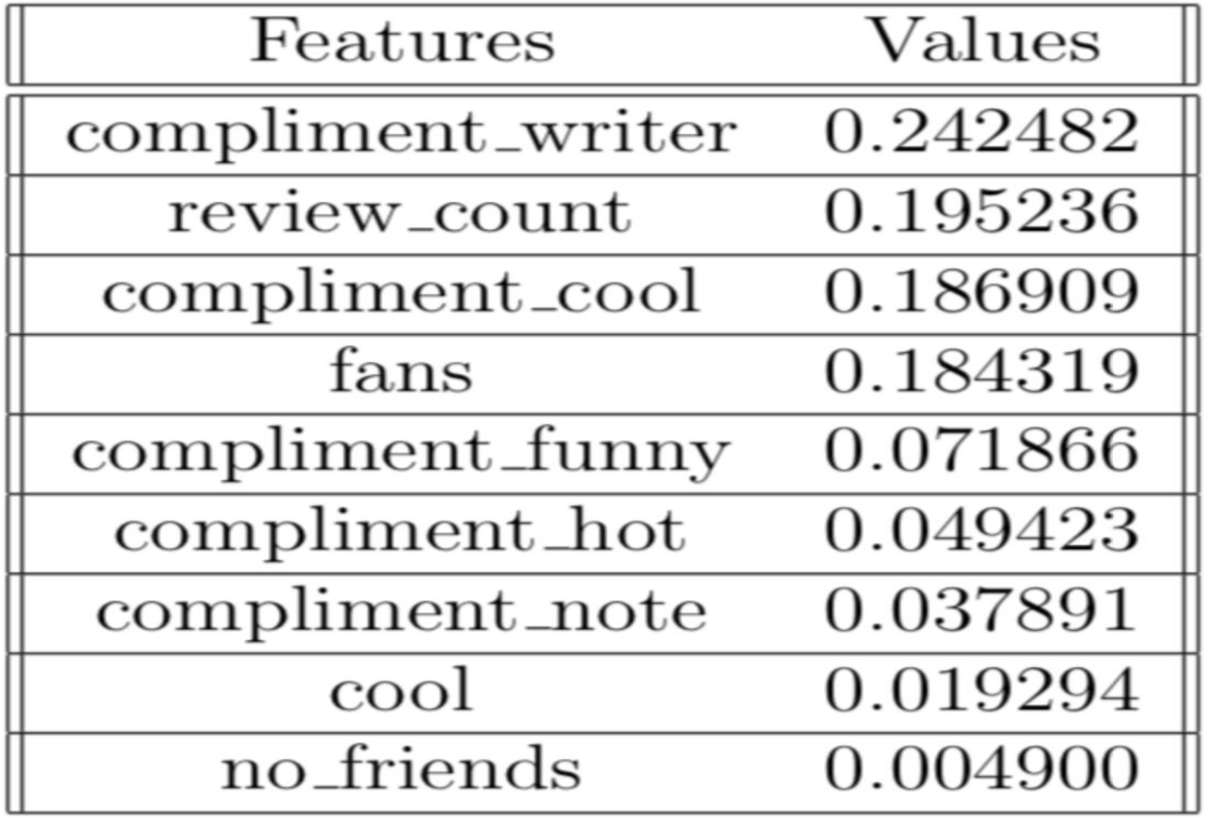
- Random Forest Classifier is basically an ensemble machine learning which combines weaker predictive models into stronger one by constructing a multitude of decision trees and generalising the model.

- We use Random Forest to obtain importance of the different features in the user dataset and thus calculate the eliteness score.

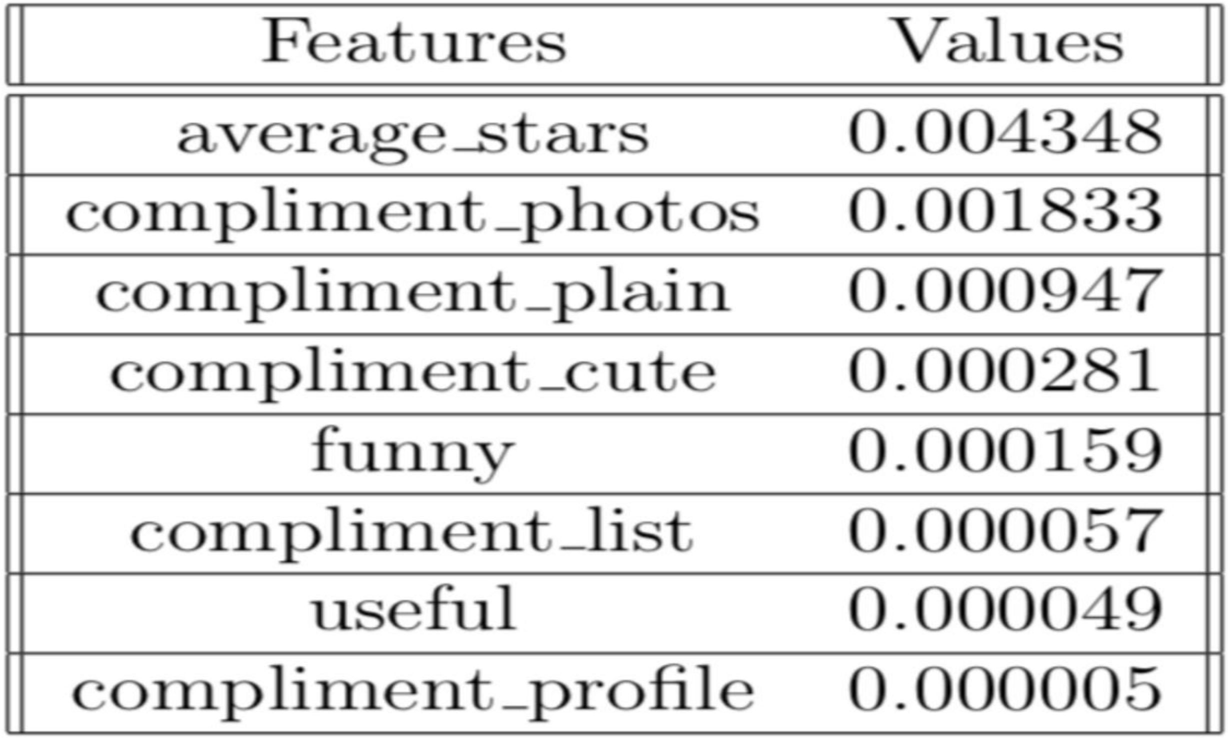


Calculating Eliteness Score

- Given a user it's eliteness score will be dot product of corresponding feature value with its importance.



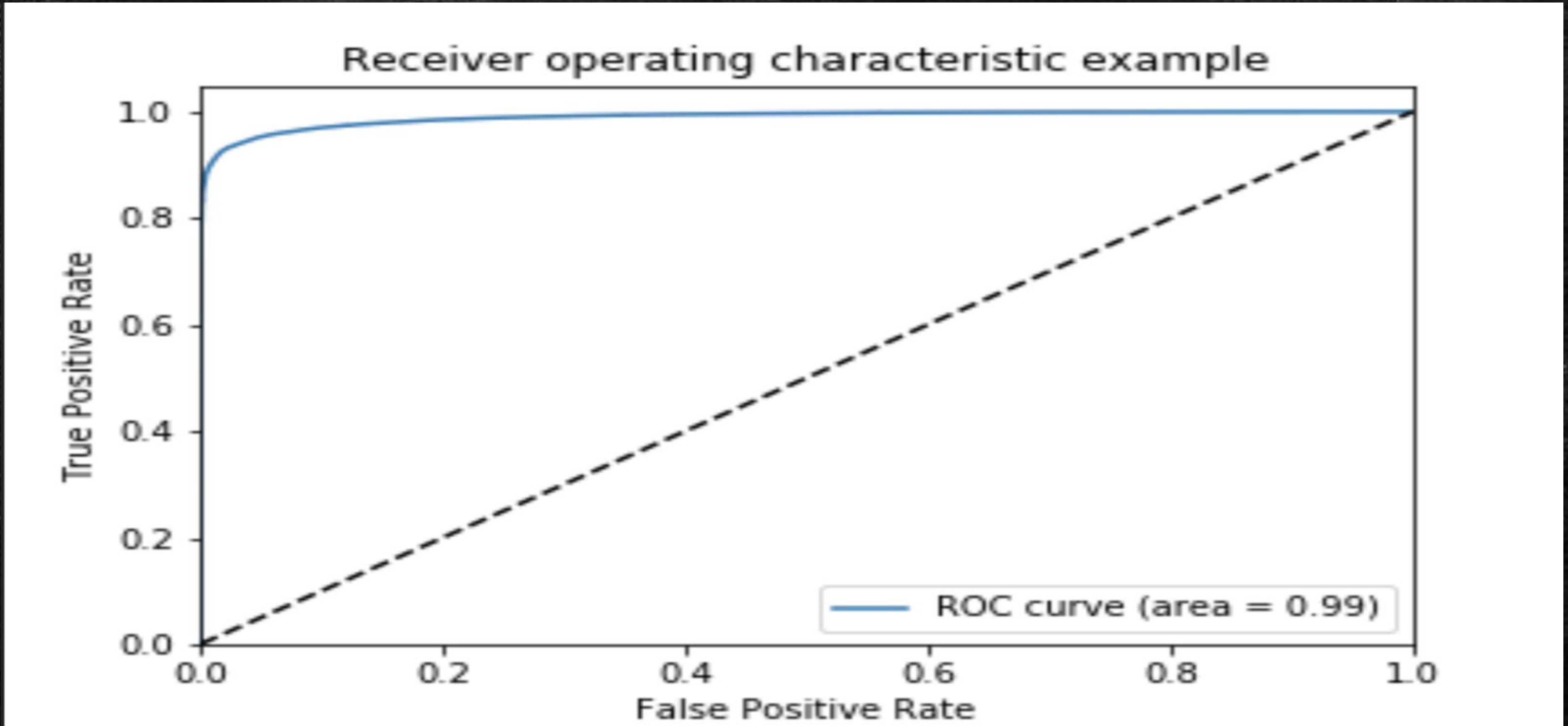
Features	Values
compliment_writer	0.242482
review_count	0.195236
compliment_cool	0.186909
fans	0.184319
compliment_funny	0.071866
compliment_hot	0.049423
compliment_note	0.037891
cool	0.019294
no_friends	0.004900



Features	Values
average_stars	0.004348
compliment_photos	0.001833
compliment_plain	0.000947
compliment_cute	0.000281
funny	0.000159
compliment_list	0.000057
useful	0.000049
compliment_profile	0.000005

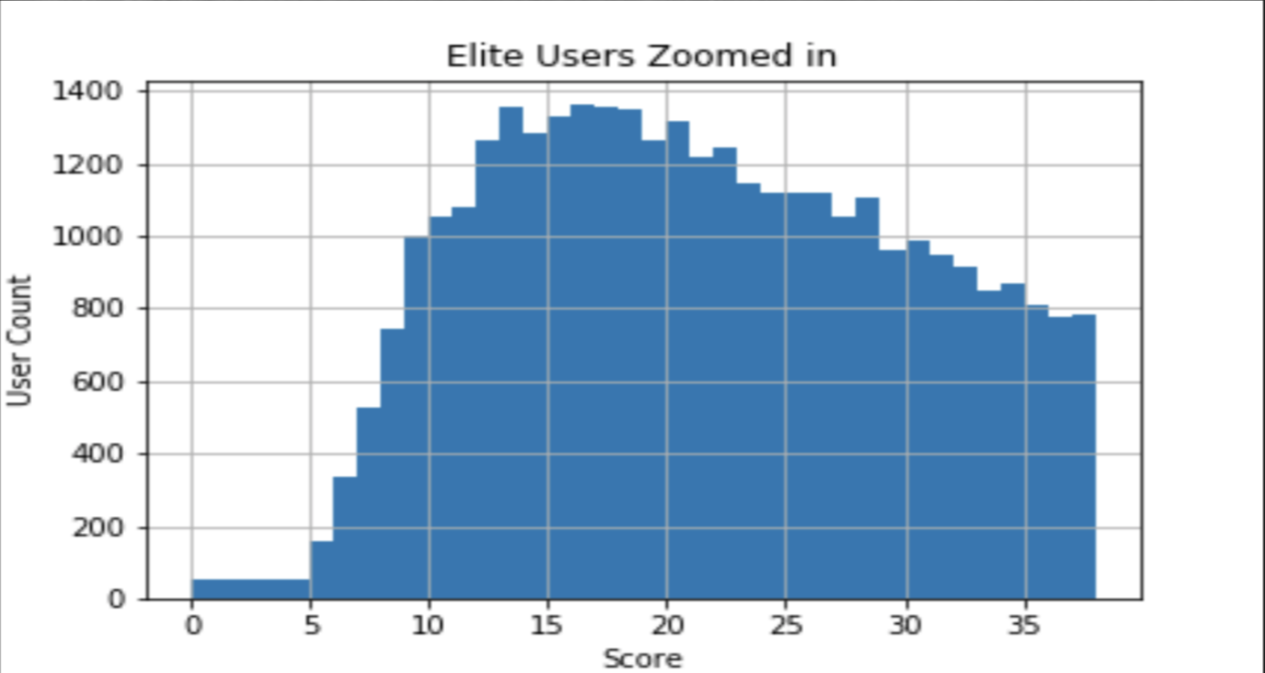
Reliability of Feature Importance

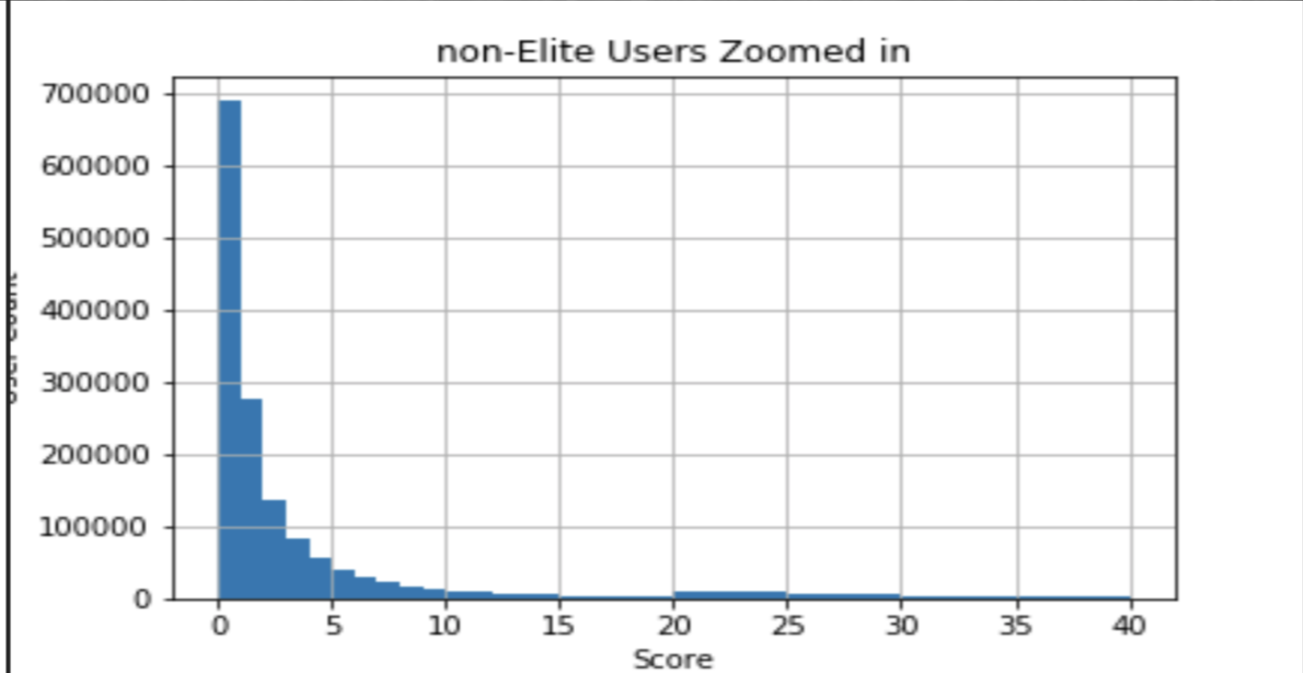
- The random forest model we trained has an accuracy of 98.21% and thus can be seen from the ROC curve.



Results (User Credibility)

- Only 4.62% of the total users have been labelled "elite" in the dataset.
- It can be clearly seen from the spike in non-Elite score distribution that majority of the users have low score.





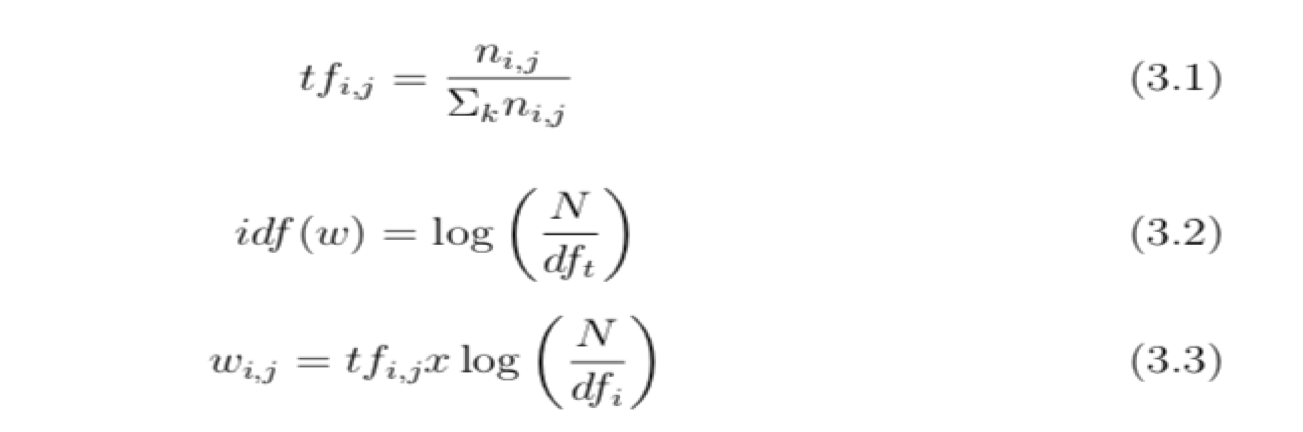
Topic Modelling

- Topic modelling corresponds to discovering abstract ‘topics’ from a collection of documents
- It can be used for information retrieval from unstructured texts in real world datasets using techniques like Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI), Non-Negative Matrix Factorization (NMF)
- But some data preprocessing is also required, so we use Term Frequency-Inverse Document Frequency (TF-IDF) to extract out the relevant information

Term Frequency - Inverse Document Frequency

- TF-IDF helps us in finding important words which give an idea about the document. It is product of Term Frequency (TF) and Inverse Document Frequency (IDF)

- Term Frequency measures the frequency of a word in a document divided by document length for normalization.



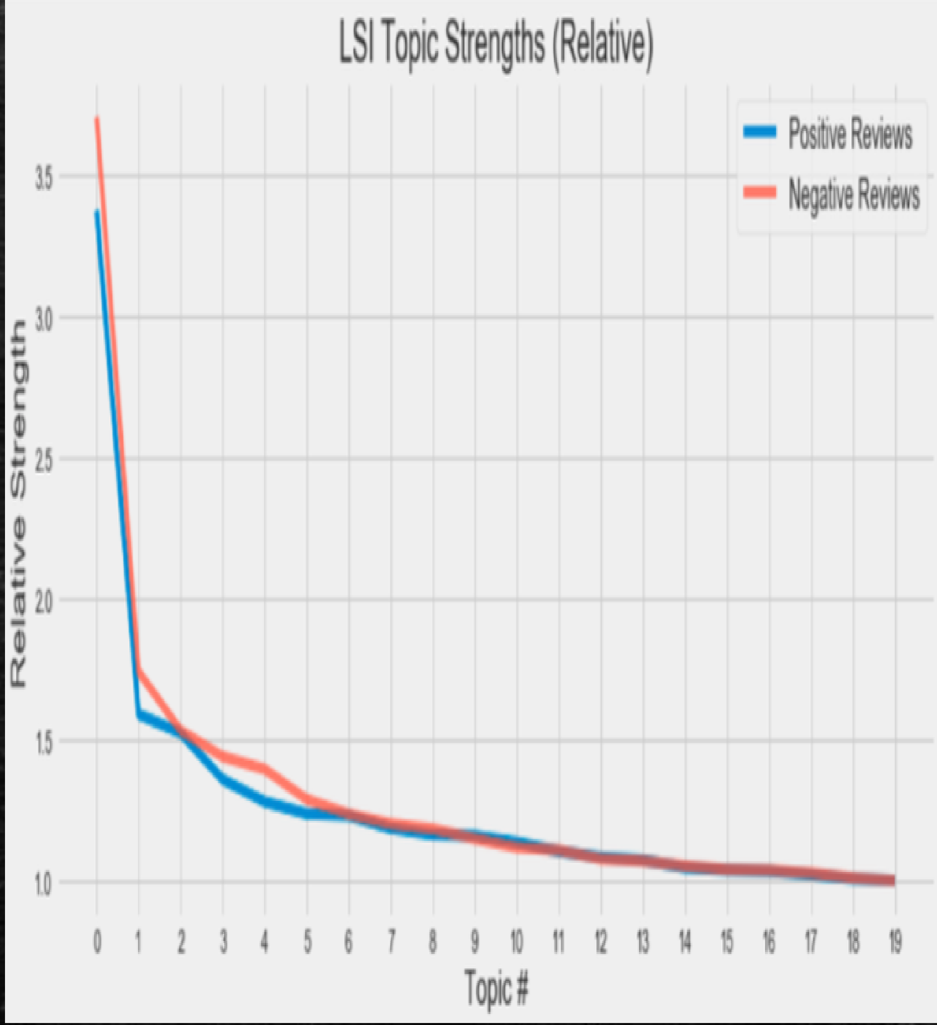
$$tf_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}} \tag{3.1}$$

$$idf(w) = \log\left(\frac{N}{df_t}\right) \tag{3.2}$$

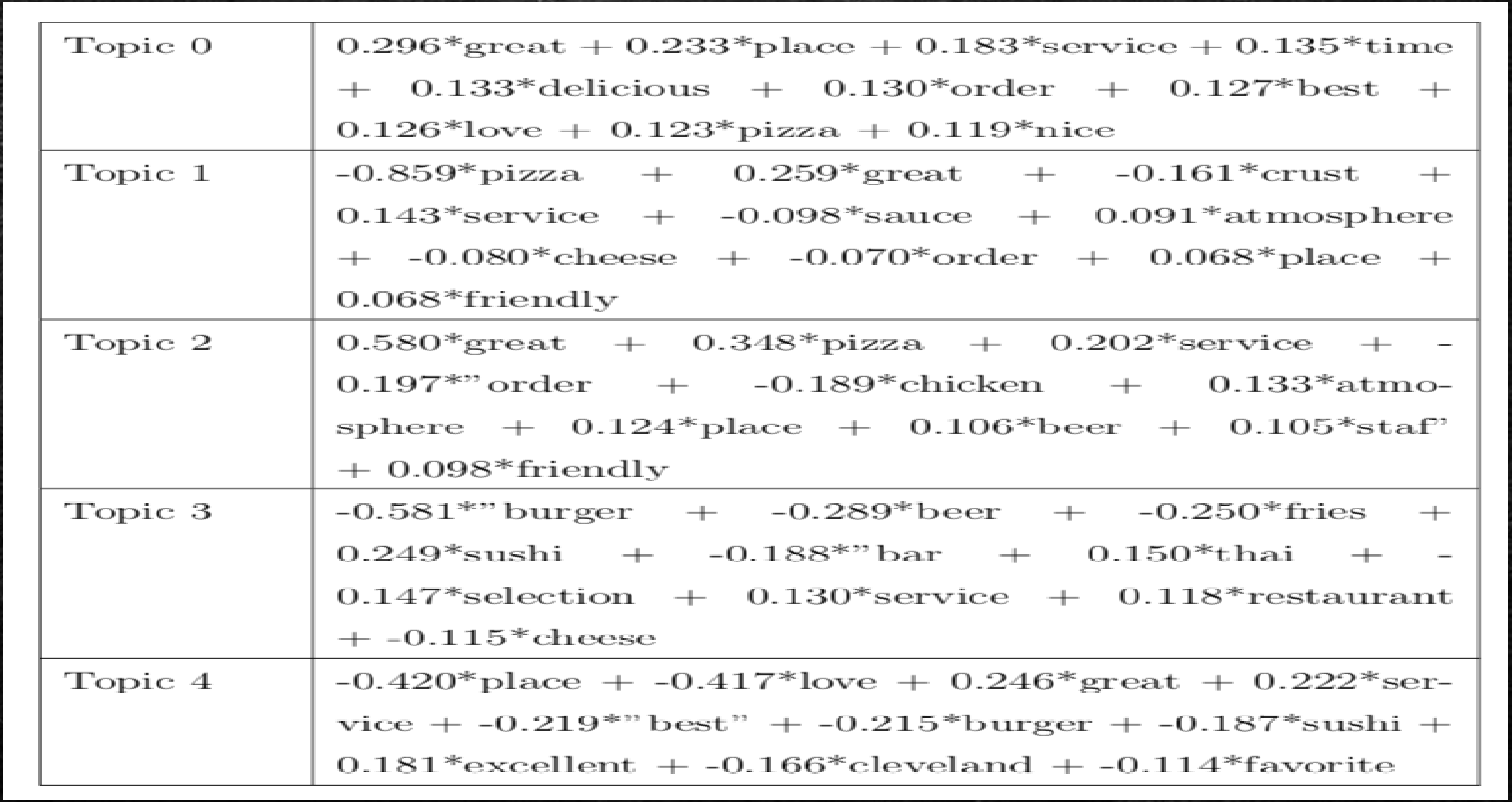
$$w_{i,j} = tf_{i,j} x \log\left(\frac{N}{df_i}\right) \tag{3.3}$$

Latent Semantic Indexing (LSI)

- LSI uses Singular Value Decomposition to scan unstructured and complex data within a document and thus helps in finding hidden (latent) relationships between words to improve upon information indexing
- Relative elbow plot is used to determine the required number of topics to be chosen

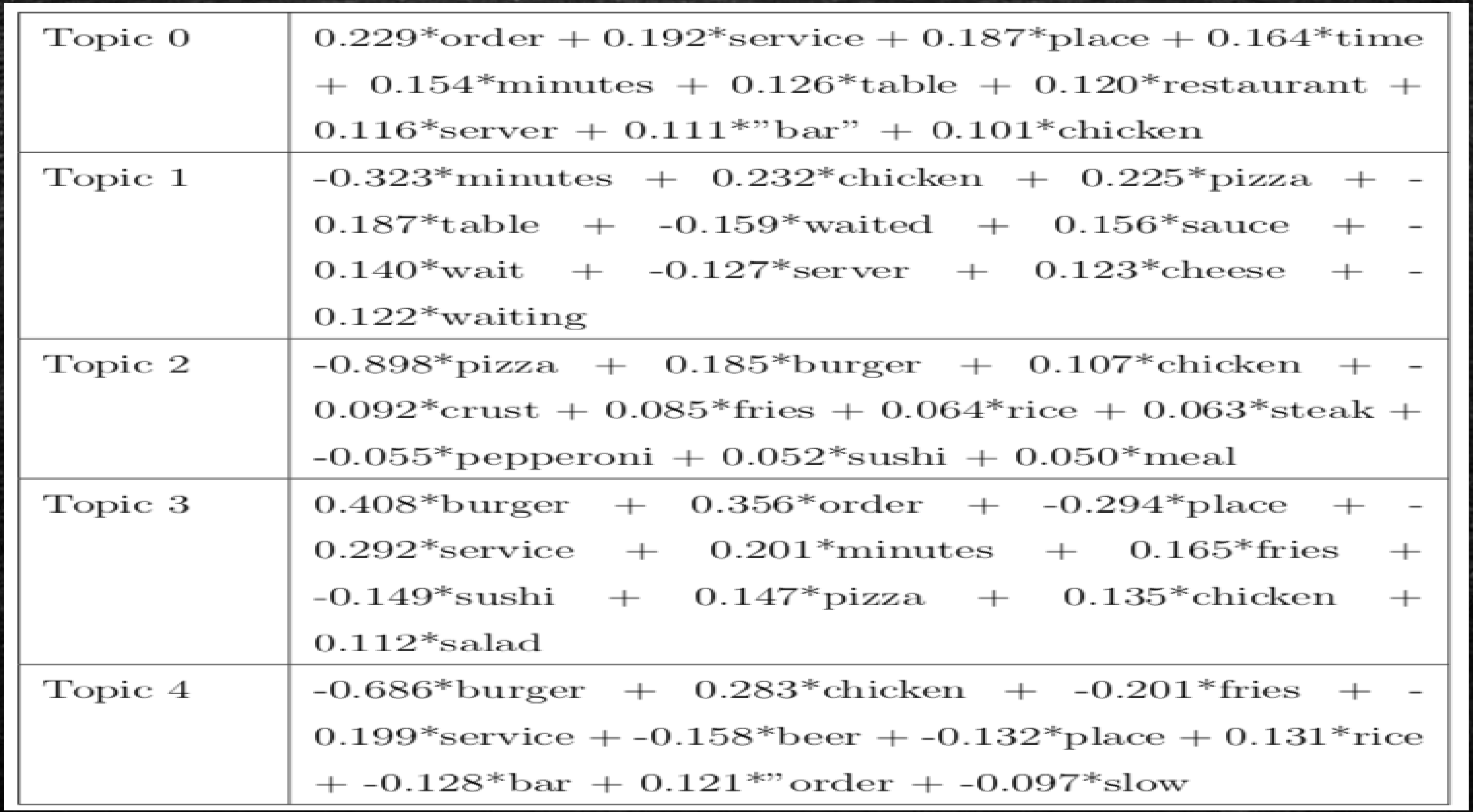


Top 5 Positive Topics using LSI



Topic 0	0.296*great + 0.233*place + 0.183*service + 0.135*time + 0.133*delicious + 0.130*order + 0.127*best + 0.126*love + 0.123*pizza + 0.119*nice
Topic 1	-0.859*pizza + 0.259*great + -0.161*crust + 0.143*service + -0.098*sauce + 0.091*atmosphere + -0.080*cheese + -0.070*order + 0.068*place + 0.068*friendly
Topic 2	0.580*great + 0.348*pizza + 0.202*service + -0.197*”order + -0.189*chicken + 0.133*atmosphere + 0.124*place + 0.106*beer + 0.105*staf” + 0.098*friendly
Topic 3	-0.581*”burger + -0.289*beer + -0.250*fries + 0.249*sushi + -0.188*”bar + 0.150*thai + -0.147*selection + 0.130*service + 0.118*restaurant + -0.115*cheese
Topic 4	-0.420*place + -0.417*love + 0.246*great + 0.222*service + -0.219*”best” + -0.215*burger + -0.187*sushi + 0.181*excellent + -0.166*cleveland + -0.114*favorite

Top 5 Negative Topics using LSI



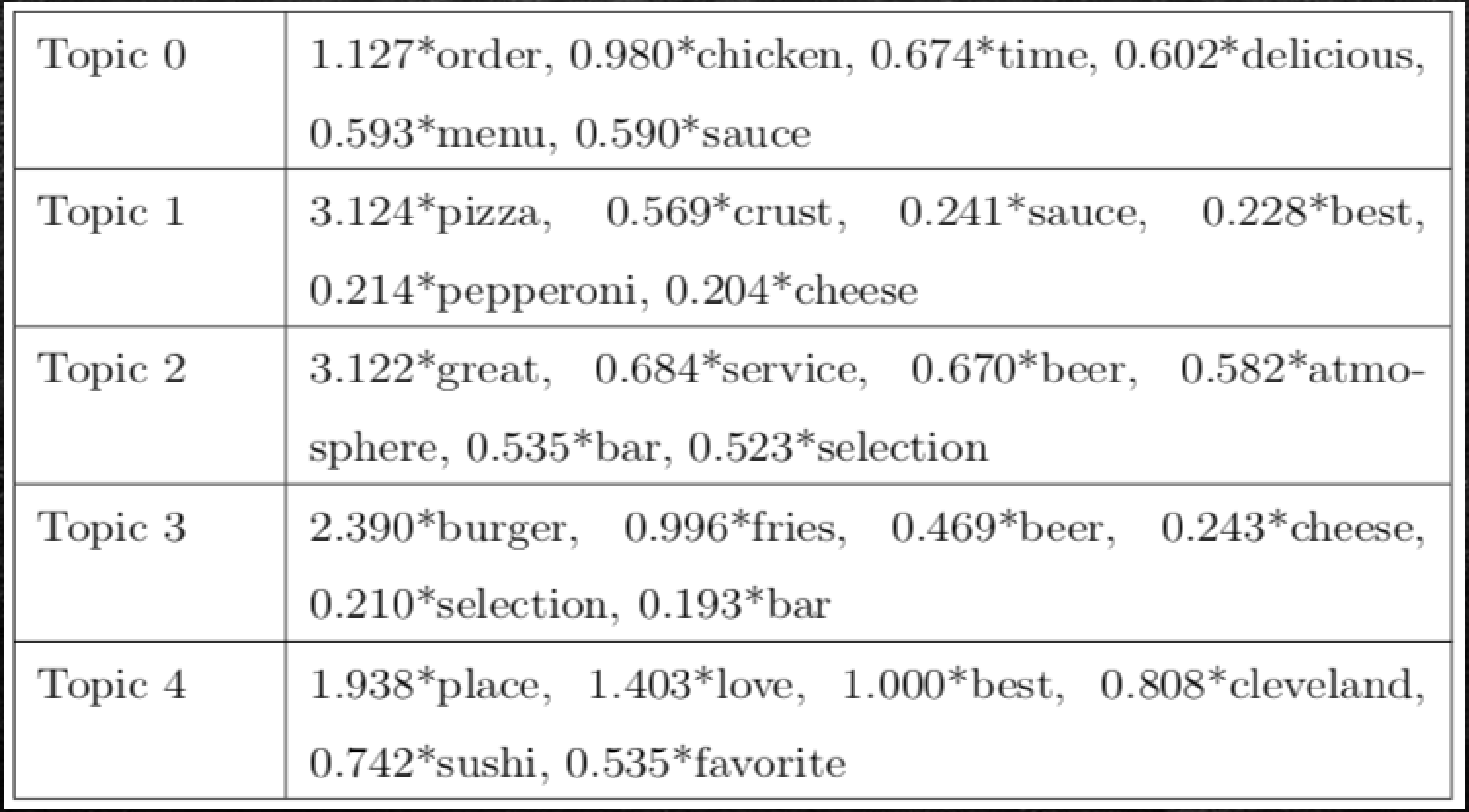
Topic 0	0.229*order + 0.192*service + 0.187*place + 0.164*time + 0.154*minutes + 0.126*table + 0.120*restaurant + 0.116*server + 0.111*”bar” + 0.101*chicken
Topic 1	-0.323*minutes + 0.232*chicken + 0.225*pizza + -0.187*table + -0.159*waited + 0.156*sauce + -0.140*wait + -0.127*server + 0.123*cheese + -0.122*waiting
Topic 2	-0.898*pizza + 0.185*burger + 0.107*chicken + -0.092*crust + 0.085*fries + 0.064*rice + 0.063*steak + -0.055*pepperoni + 0.052*sushi + 0.050*meal
Topic 3	0.408*burger + 0.356*order + -0.294*place + -0.292*service + 0.201*minutes + 0.165*fries + -0.149*sushi + 0.147*pizza + 0.135*chicken + 0.112*salad
Topic 4	-0.686*burger + 0.283*chicken + -0.201*fries + -0.199*service + -0.158*beer + -0.132*place + 0.131*rice + -0.128*bar + 0.121*”order + -0.097*slow

Non-Negative Matrix Factorization (NMF)

- NMF is a linear algebraic model which reduces high dimensional vectors into a low- dimensionality representation being similar to PCA

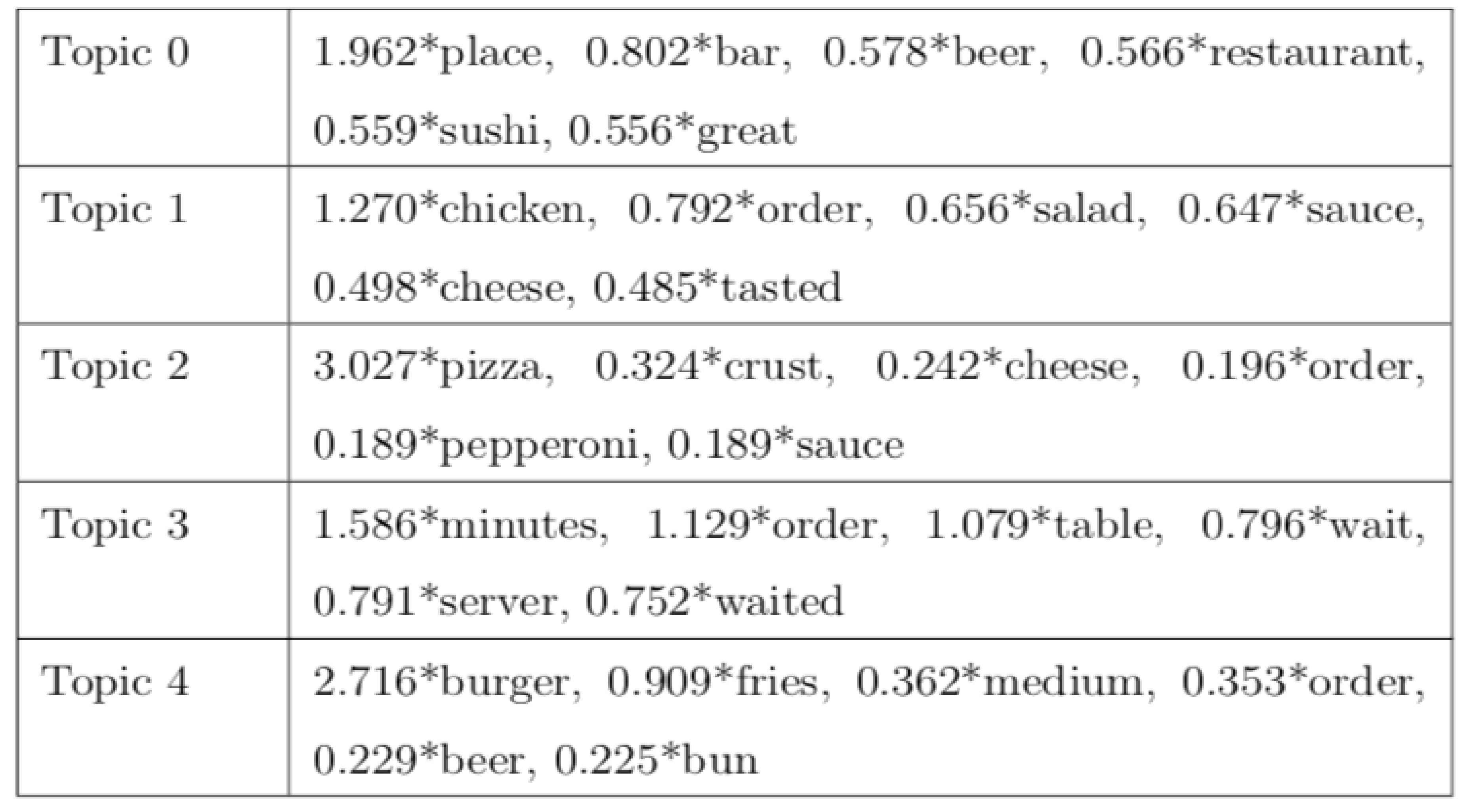
- It uses non negative vectors and forces the coefficients to also be non-negative.

Top 5 Positive Topics using NMF



Topic 0	1.127*order, 0.980*chicken, 0.674*time, 0.602*delicious, 0.593*menu, 0.590*sauce
Topic 1	3.124*pizza, 0.569*crust, 0.241*sauce, 0.228*best, 0.214*pepperoni, 0.204*cheese
Topic 2	3.122*great, 0.684*service, 0.670*beer, 0.582*atmosphere, 0.535*bar, 0.523*selection
Topic 3	2.390*burger, 0.996*fries, 0.469*beer, 0.243*cheese, 0.210*selection, 0.193*bar
Topic 4	1.938*place, 1.403*love, 1.000*best, 0.808*cleveland, 0.742*sushi, 0.535*favorite

Top 5 Negative Topics using NMF

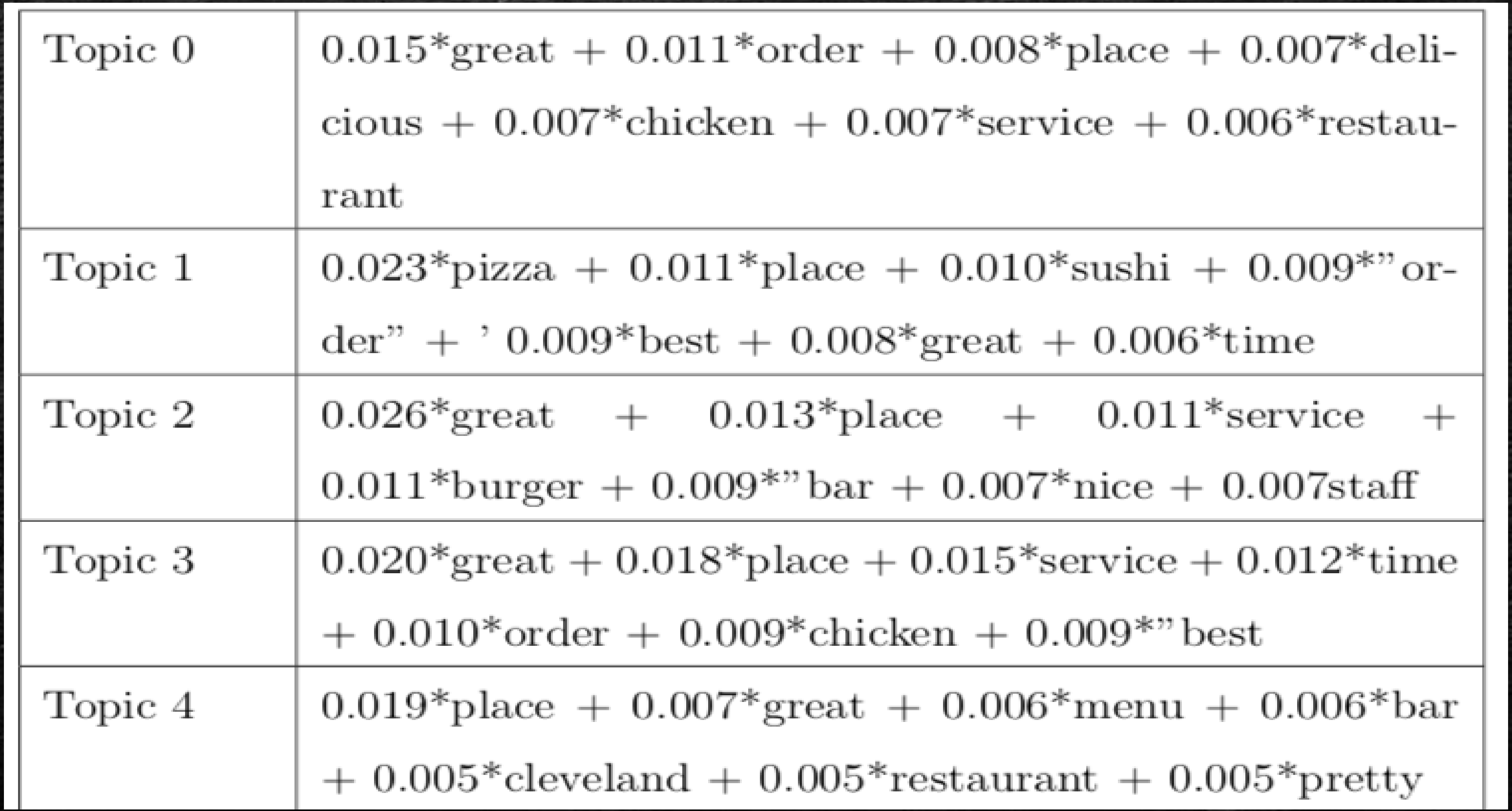


Topic 0	1.962*place, 0.802*bar, 0.578*beer, 0.566*restaurant, 0.559*sushi, 0.556*great
Topic 1	1.270*chicken, 0.792*order, 0.656*salad, 0.647*sauce, 0.498*cheese, 0.485*tasted
Topic 2	3.027*pizza, 0.324*crust, 0.242*cheese, 0.196*order, 0.189*pepperoni, 0.189*sauce
Topic 3	1.586*minutes, 1.129*order, 1.079*table, 0.796*wait, 0.791*server, 0.752*waited
Topic 4	2.716*burger, 0.909*fries, 0.362*medium, 0.353*order, 0.229*beer, 0.225*bun

Latent Dirichlet Allocation (LDA)

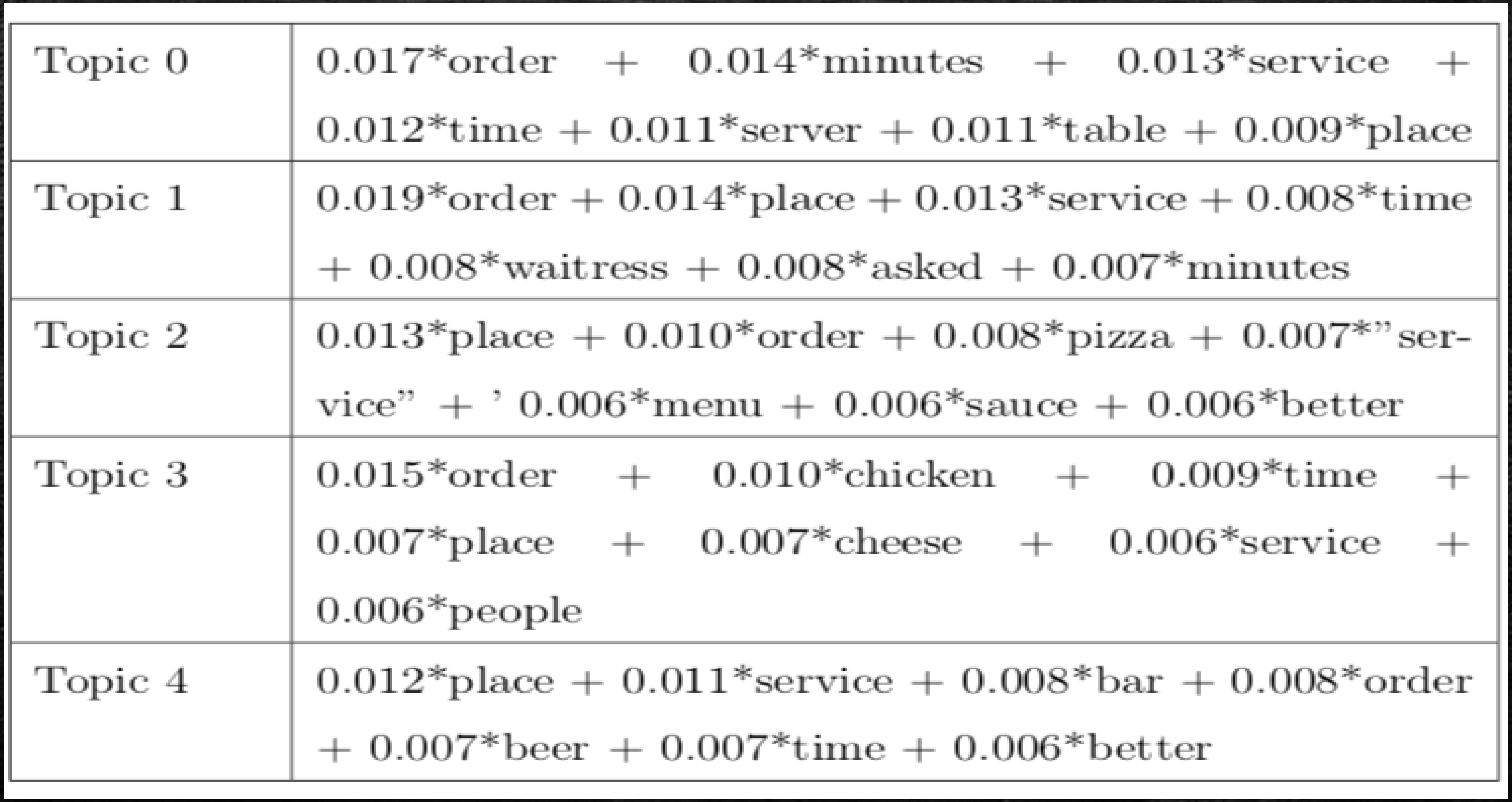
- In LDA technique, each document is used to describe distribution of topics and each topic described by distribution of words.
- It assumes a fixed number of topics in which each topic represents a set of words. A mapping is established from all documents to each topic in such a way that most of the words in each document are captured by the corresponding topics.

Top 5 Positive Topics using LDA



Topic 0	0.015*great + 0.011*order + 0.008*place + 0.007*delicious + 0.007*chicken + 0.007*service + 0.006*restaurant
Topic 1	0.023*pizza + 0.011*place + 0.010*sushi + 0.009*”order” + ’ 0.009*best + 0.008*great + 0.006*time
Topic 2	0.026*great + 0.013*place + 0.011*service + 0.011*burger + 0.009*”bar + 0.007*nice + 0.007staff
Topic 3	0.020*great + 0.018*place + 0.015*service + 0.012*time + 0.010*order + 0.009*chicken + 0.009*”best
Topic 4	0.019*place + 0.007*great + 0.006*menu + 0.006*bar + 0.005*cleveland + 0.005*restaurant + 0.005*pretty

Top 5 Negative Topics using LDA



Topic 0	0.017*order + 0.014*minutes + 0.013*service + 0.012*time + 0.011*server + 0.011*table + 0.009*place
Topic 1	0.019*order + 0.014*place + 0.013*service + 0.008*time + 0.008*waitress + 0.008*asked + 0.007*minutes
Topic 2	0.013*place + 0.010*order + 0.008*pizza + 0.007*”service” + ’ 0.006*menu + 0.006*sauce + 0.006*better
Topic 3	0.015*order + 0.010*chicken + 0.009*time + 0.007*place + 0.007*cheese + 0.006*service + 0.006*people
Topic 4	0.012*place + 0.011*service + 0.008*bar + 0.008*order + 0.007*beer + 0.007*time + 0.006*better

Comparison between LSI, LDA, and NMF

- LSI gave us topics but had given weights to the words in the range of [-1,1] from which we cant infer the relative importance of words in a topic
- NMF and LDA gives positive weights to the words in a topic
- Topics given by LDA has a lot of overlapping of words but NMF on the other hand has almost no overlapping thus providing more discrete topics

Sentimental Analysis of a Review and its Rating

- Sentimental Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral which in our case is predominantly review text.
- Our key aim with sentimental analysis is to pick out the emotion of the review text and analyse with its corresponding rating as if to analyse the bias between both of them.
- It works as a validation to our model of building a pseudo rating system.

Results (Sentiment Analysis)

- Naïve Bayes for prediction of positive or negative reviews gave us a f1-score of 0.97.
- Naïve Bayes for prediction of star rating when processed only upon the review text gave us a f1-score of 0.63 which is evident from the fact that stars are just not dependent on the review text rather than on many parameters like cool, useful, geographical locations and other important variables.

Conclusion

- Now overall, after conducting our analysis in three different directions, we would like to conclude our findings and try to draw some meaning out of all of them conjunction.
- It's obvious that an increased percentage of reviews by reliable users will benefit the businesses which are actually doing good work and people would be less likely to be disguised by the false reviews.
- The businesses which would focus on the topics found out via Topic Modelling can achieve major gains by improving in those scenarios, thus the platform would prove helpful for those type of businesses.
- At last, the method of sentimental analysis to give a pseudo rating and cross checking the positivity or negativity of a review with its' indexed rating will definitely keep the sanity check of the reviews published on the platform.

Conclusion

- Since proving that whether the crowd-sourcing review platform Yelp is reliable or not is not a binary, it is mostly a distribution of pros and cons given to it in various sections.
- The ability of malicious and ill-intended reviews to creep in the system is cruel but the measures to keep it sanitized has been rigorous.
- It can be inferred that if in together users are more reliable on the platform, with the businesses getting a feedback of their performance on a daily basis and also the actions of users are kept a check on, a pretty efficient and not so vulnerable system of crowd-sourced review and ratings can be maintained for a prolonged period.
- In conclusion, we would like to say that there are a hundred different ways this huge dataset can be utilized and analyzed to ascertain more vulnerabilities and focal points of the platform which will help in keeping it a reliable source of making decisions.

References

- Kyong Jin Shim Koo Ping Shung Maruthi Prithivirajan, Vivian Lai. Analysis of star ratings in consumer reviews. IEEE International Conference on Big Data, 2015.
- Y. Paule Sun. Spatial analysis of users-generated ratings of yelp venues. J.D.G. Open geospatial data, softw. stand., 2017.
- William Q. Meeker Yili Hong, Man Zhang. Big data and reliability applications: The complexity dimension. Journal of Quality Technology, 2018.
- James Huang, Stephanie Rogers, and Eunkwang Joo. Improving restaurants by extracting subtopics from Yelp reviews. iConference 2014 (Social Media Expo), 2014.
- Chen Wen Li and Jin Zhang. Prediction of Yelp review star rating using sentiment analysis. 2014.