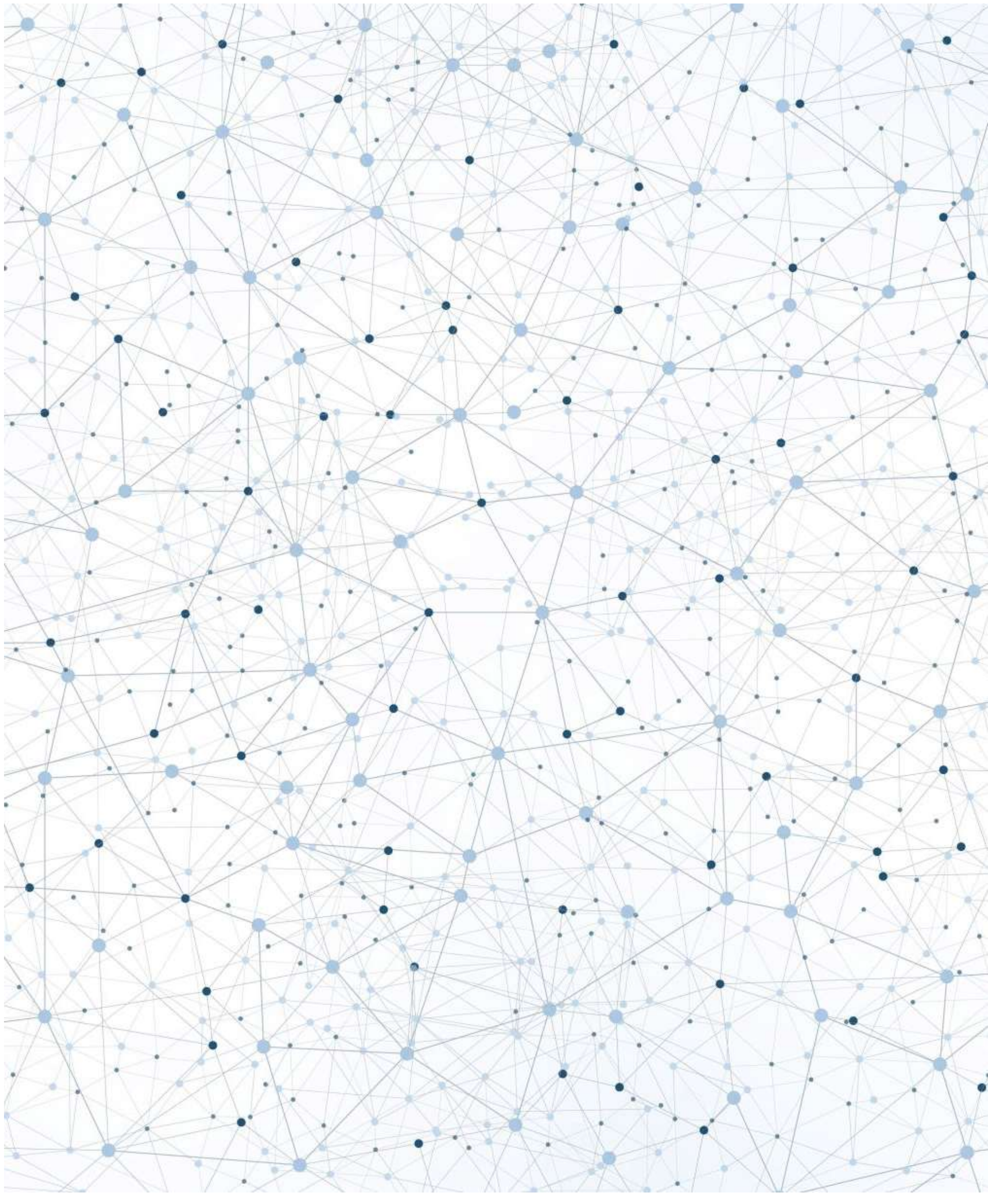
DATA MINING –SENTIMENT ANALYSIS

Tami Farber & Trang Do

Spring 2022

DSC 411: Data Mining



RESEARCH AREA

Sentiment analysis…

…used to analyze the text-based output of domains like:

- Social media
- Online reviews
- Research papers and other e-Text

Goals

- Identify trends
- Gather public-opinion
- Expose misinformation

RESEARCH QUESTIONS

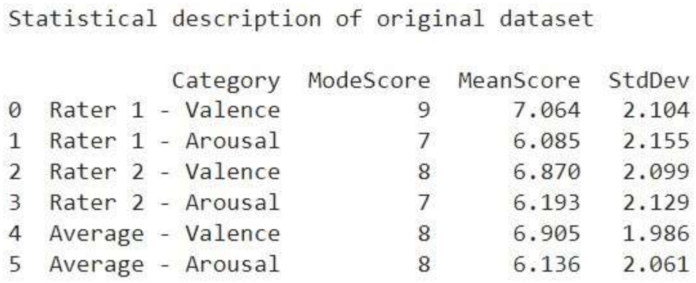
Which type of machine learning model is best suited for the task?

Is there an ML model that can reliably match the human rating system?

Can we create an ML model that will provide insight into the reviews that might have influenced the human rating system?

ABOUT THE DATASET

- Collected from Yelp! reviews from the Chicago area
- Rated for valence and arousal by professional raters (humans)
- Made up of 2000 instances with 9 features



```
Statistical description of original dataset

            Category  ModeScore  MeanScore  StdDev
0  Rater 1 - Valence          9      7.064   2.104
1  Rater 1 - Arousal          7      6.085   2.155
2  Rater 2 - Valence          8      6.870   2.099
3  Rater 2 - Arousal          7      6.193   2.129
4  Average - Valence          8      6.905   1.986
5  Average - Arousal          8      6.136   2.061
```

DATA MINING TECHNIQUES

- Data Preprocessing

Step 1: Transform 1-9 ranking scale into five possible categories

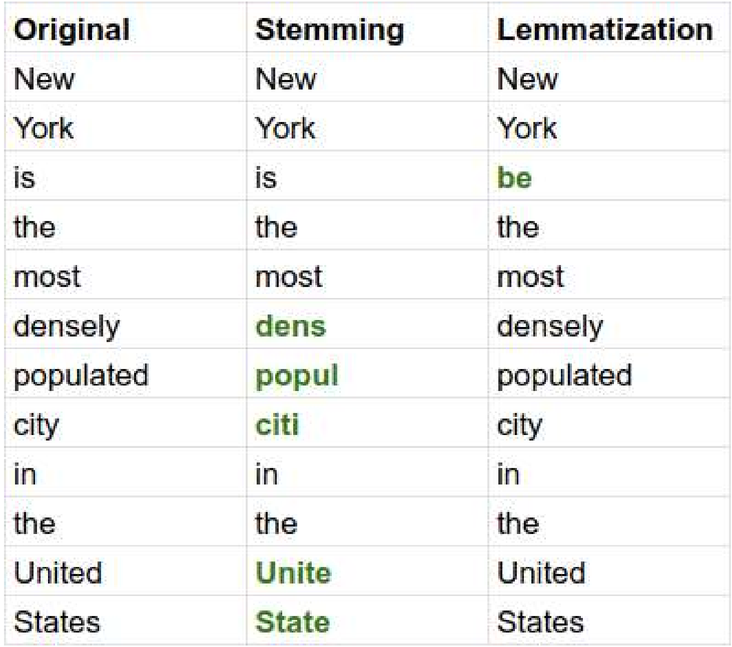
- negative [1-2]
- somewhat negative [3-4]
- neutral [5]
- somewhat positive [6-7]
- positive [8-9]

DATA MINING TECHNIQUES

- Data Preprocessing

Step 2: Clean the reviews

- Adjust misspelled words
  - Ex: gooddddd -> good
- Remove urls, stop words, punctuation
- Lemmatize text



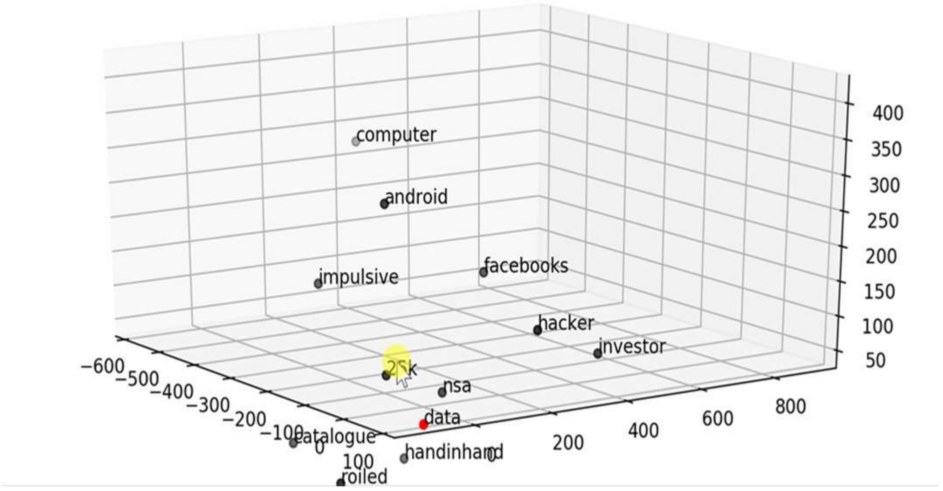
Original	Stemming	Lemmatization
New	New	New
York	York	York
is	is	**be**
the	the	the
most	most	most
densely	**dens**	densely
populated	**popul**	populated
city	**citi**	city
in	in	in
the	the	the
United	**Unite**	United
States	**State**	States

Data mining techniques

- Data Preprocessing

<u>Step 3:</u> Transform reviews with TF-IDF Vectorizer

- Transform the cleaned text to vectors that can be used to analyze the text
- Transform data for models requiring boolean values (0 and 1)

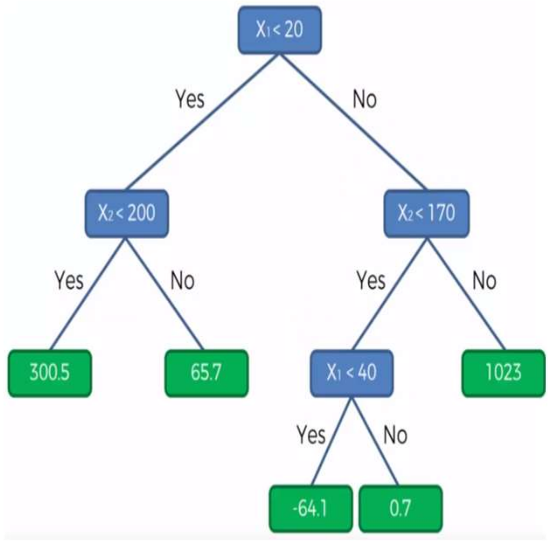


DATA MINING TECHNIQUES

- Basic Classifications

1. Decision Tree

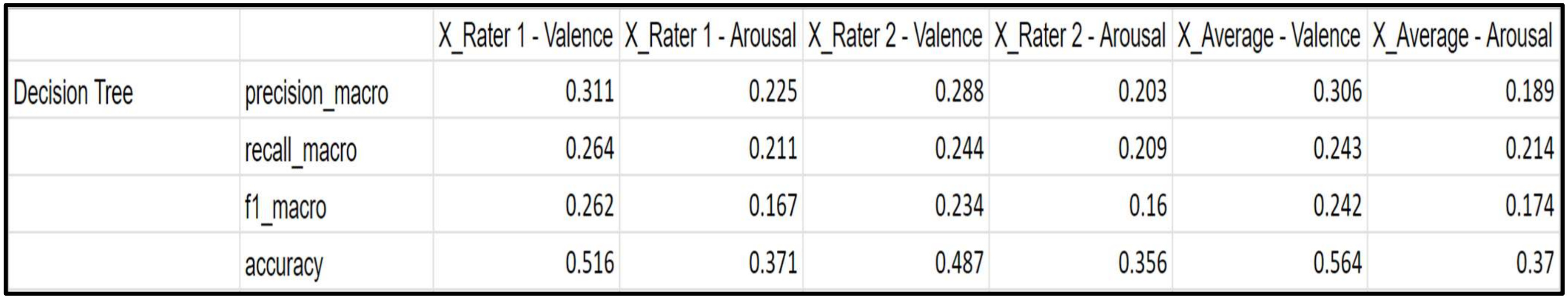
- Supervised learning technique
- Similar to a flowchart with tree structures of all the possible solutions to a decision, based on conditions
- Decides whether an instance satisfies a condition or not, continuing passing it to next condition until it reaches a decision



DATA MINING TECHNIQUES

- Basic Classifications

1. Decision Tree – Model Validation



		X_Rater 1 - Valence	X_Rater 1 - Arousal	X_Rater 2 - Valence	X_Rater 2 - Arousal	X_Average - Valence	X_Average - Arousal
Decision Tree	precision_macro	0.311	0.225	0.288	0.203	0.306	0.189
	recall_macro	0.264	0.211	0.244	0.209	0.243	0.214
	f1_macro	0.262	0.167	0.234	0.16	0.242	0.174
	accuracy	0.516	0.371	0.487	0.356	0.564	0.37

Took ~600 seconds

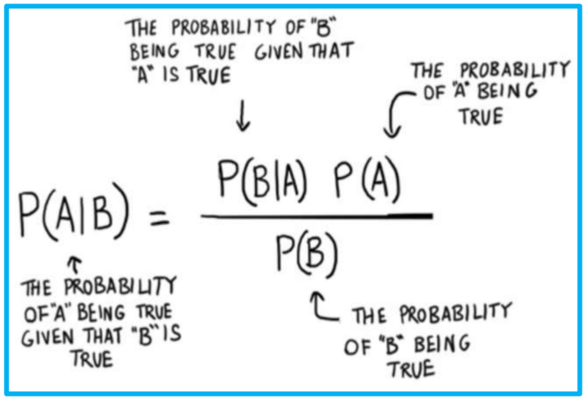
DATA MINING TECHNIQUES

- Basic Classifications

2. Multinomial Naïve Bayes

- One type of Naive Bayes Classification - a classification technique based on probability calculations of an event given a condition
- Considers term frequency

=> Commonly used for text mining

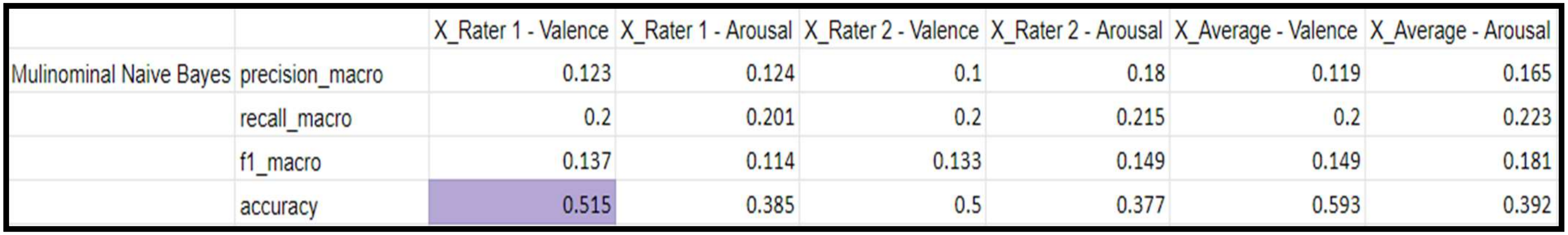


THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
THE PROBABILITY OF "A" BEING TRUE
P(A|B) = P(B|A) P(A) / P(B)
THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
THE PROBABILITY OF "B" BEING TRUE

DATA MINING TECHNIQUES

- Basic Classifications

2. Multinomial Naïve Bayes – Model Validation



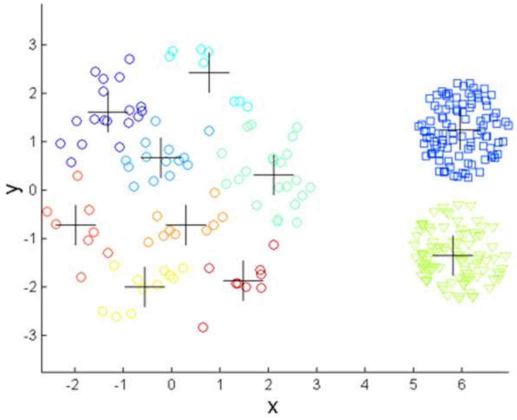
		X_Rater 1 - Valence	X_Rater 1 - Arousal	X_Rater 2 - Valence	X_Rater 2 - Arousal	X_Average - Valence	X_Average - Arousal
Mulinominal Naive Bayes	precision_macro	0.123	0.124	0.1	0.18	0.119	0.165
	recall_macro	0.2	0.201	0.2	0.215	0.2	0.223
	f1_macro	0.137	0.114	0.133	0.149	0.149	0.181
	accuracy	0.515	0.385	0.5	0.377	0.593	0.392

Took ~52 seconds

DATA MINING TECHNIQUES
- Basic classifications

3. K-means Clustering

- Unsupervised learning model
- Requires normalized or vectorized data
- Generates k clusters of each data point around a centroid
- Uses distance measures to create optimal clusters
- SSE: sum of least squares $SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$
- Cluster labels can be compared to known target labels to test model predictions

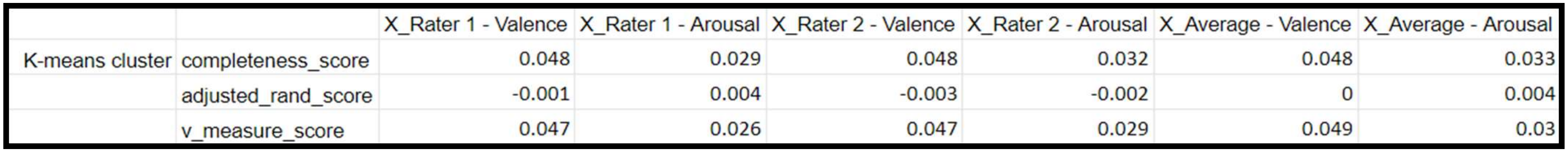


K-means Clusters

DATA MINING TECHNIQUES

- Basic Classifications

3. K-means Clustering – Model Validation

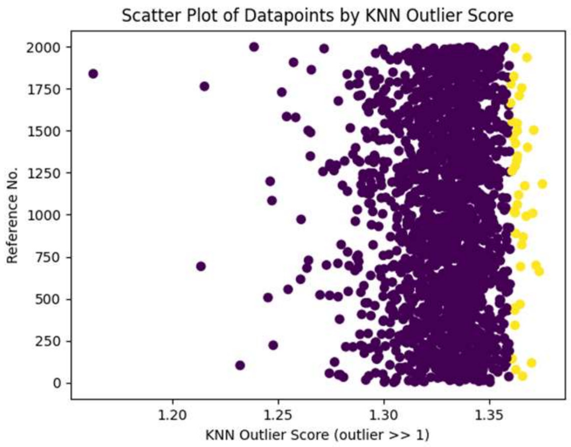


		X_Rater 1 - Valence	X_Rater 1 - Arousal	X_Rater 2 - Valence	X_Rater 2 - Arousal	X_Average - Valence	X_Average - Arousal
K-means cluster	completeness_score	0.048	0.029	0.048	0.032	0.048	0.033
	adjusted_rand_score	-0.001	0.004	-0.003	-0.002	0	0.004
	v_measure_score	0.047	0.026	0.047	0.029	0.049	0.03

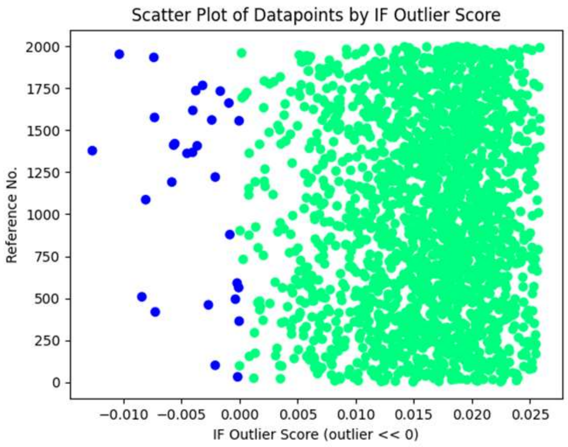
- Model was built with 5 clusters to approximate the number of class labels
- Labels were encoded to match format of cluster labels
- Single cluster data point is the normalized collection of words in a single review
- With 10-Fold cross-validation, takes ~ 1 hour to complete

DATA MINING TECHNIQUES
- Anomaly detection

An anomaly is data that stands out from the rest



- Error
- Insignificant noise
- Unexpected truth



In sentiment analysis and text mining, anomalies can add insight or complicate our model.

DATA MINING TECHNIQUES
- Anomaly detection

1. Isolation Forest

Starts with a single tree…

- Random data point is assigned to the root node
- Random partitions are chosen
- Decisions branch until a termination node is reached
- The roots with the shortest paths are scored and labeled anomalous
- Labels: outlier/anomaly = –1 inlier = 1
- Scores are values between -1 and 1, the closer to –1 a score, the more anomalous

DATA MINING TECHNIQUES
- Anomaly detection

1. Isolation Forest - Partial Result

Modeled on our dataset, Isolation Forest identified 29 anomalies.

It seemed to isolate the lengthier reviews [1,322 words to 4,988 words]:

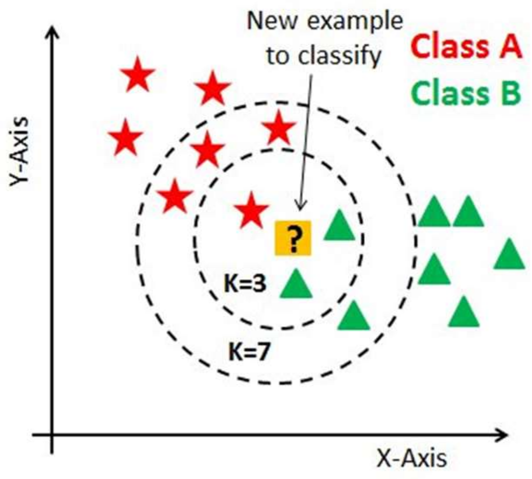
> *"Growing up on 10th Street in Park Slope, Brooklyn we had several outstanding Pizzeria's to choose from. So when taking my daughter and her fiance on a pizza quest I naturally had to stop in my old neighborhood and sample a slice from Anellio's, which wasn't there anymore. Instead Peppino's Brick Oven Pizzeria now occupies the location. I quickly looked up the comments on Yelp and saw that it was rated very high. Inside did not resemble anything from the old business. It does not resemble the familiar take style counter near the entrance, but rather a sit down set up. The Brick oven is now in the back in plane sight. We were greeted by a hostess and opted to take the Pizza to go. We ordered the Margarita with Meatballs. We sat at a table while we waited the the young girl brought us to go cups of water. Very nice touch. Within a few minutes our Pizza was ready. Brick ovens do cook very quickly. The Pizza looked and tasted fantastic. Sauce was sweet and fresh, the Mozzarella was fresh not the commercial shredded type, the meatballs were sliced homemade. The crust was thin, but not too thin. Not burned as most brick oven pizza can be. Texture of the crust was light, not chewy, with just the right amount of pull as you bite. Overall and excellent PIE. Good Service. My neighborhood still Rocks Great Pizza! "*

DATA MINING TECHNIQUES
- Anomaly Detection

2. K-nearest Neighbors (KNN)

- The unlabeled data point is likely to have the label of the major class of its k nearest neighbors
- The outlier score of an object is the distance to its k nearest neighbors

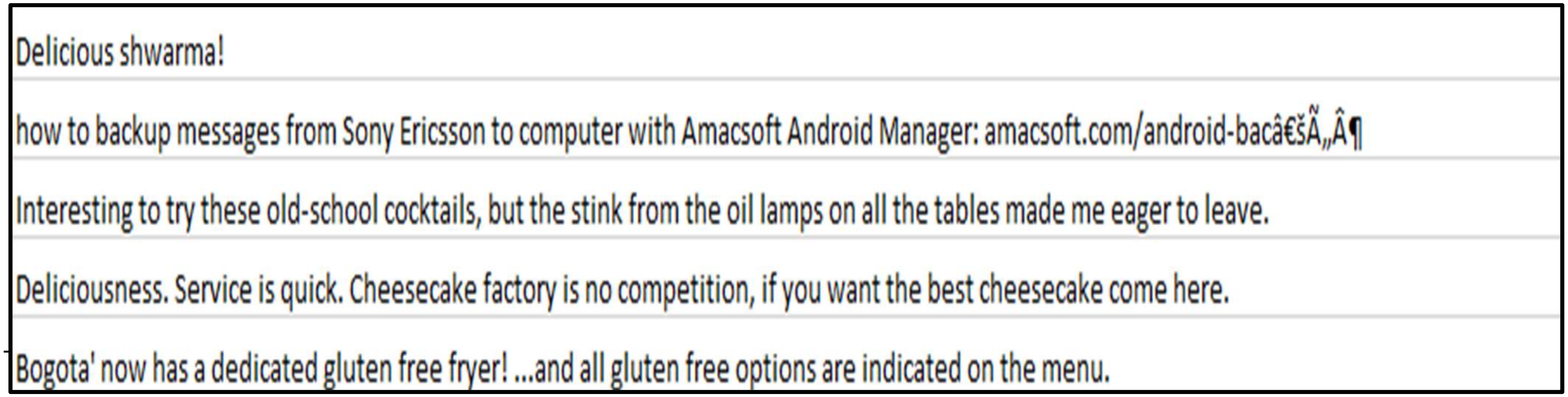
=> The higher the outlier score, the more likely a data record is an outlier



DATA MINING TECHNIQUES
- Anomaly Detection

2. K-nearest Neighbors (KNN) - Partial Result

- KNN model chose reviews with most unique words like “confetti”, “bon endroit”, “schwarma”, and “Bourdain”



Delicious shwarma!

how to backup messages from Sony Ericsson to computer with Amacsoft Android Manager: amacsoft.com/android-bacâ€šÃ„Â¶

Interesting to try these old-school cocktails, but the stink from the oil lamps on all the tables made me eager to leave.

Deliciousness. Service is quick. Cheesecake factory is no competition, if you want the best cheesecake come here.

Bogota' now has a dedicated gluten free fryer! ...and all gluten free options are indicated on the menu.

DATA MINING TECHNIQUES
- Advanced Classifications

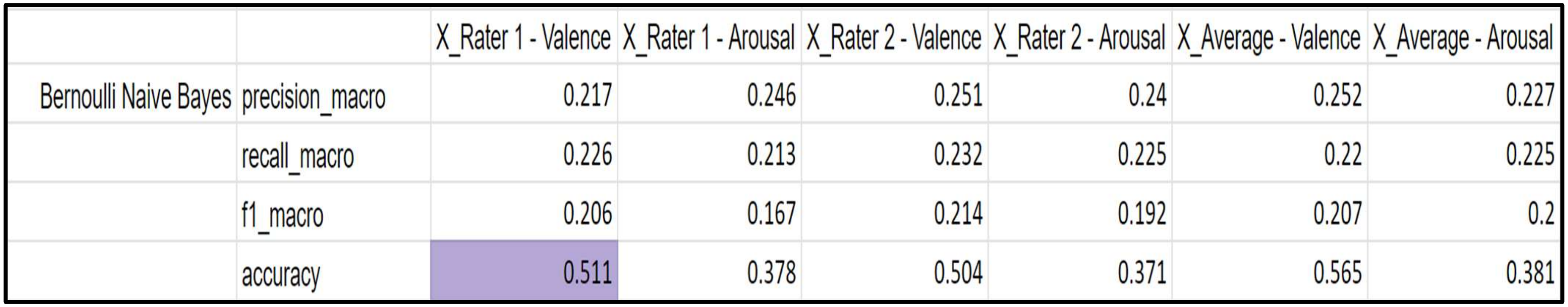
1. Bernoulli Naïve Bayes

- Another type of Naive Bayes Classification - a classification technique based on probability calculations of an event given a condition
- Only considers presence of a word
- Also used in text mining

DATA MINING TECHNIQUES

- Advanced Classifications

1. Bernoulli Naïve Bayes – Model Validation



		X_Rater 1 - Valence	X_Rater 1 - Arousal	X_Rater 2 - Valence	X_Rater 2 - Arousal	X_Average - Valence	X_Average - Arousal
Bernoulli Naive Bayes	precision_macro	0.217	0.246	0.251	0.24	0.252	0.227
	recall_macro	0.226	0.213	0.232	0.225	0.22	0.225
	f1_macro	0.206	0.167	0.214	0.192	0.207	0.2
	accuracy	0.511	0.378	0.504	0.371	0.565	0.381

Took ~62 seconds

DATA MINING TECHNIQUES
- Advanced Classifications

2. Rules-based Classification

- Classifies records by using a collection of “if...then...” rules
- IF clause – antecedent & THEN clause – consequent
- We used an indirect method to generate rules
  - ➢ Extract rules from Decision Tree
  - ➢ Only considers presence of words, not frequency

DATA MINING TECHNIQUES
- Advanced Classifications

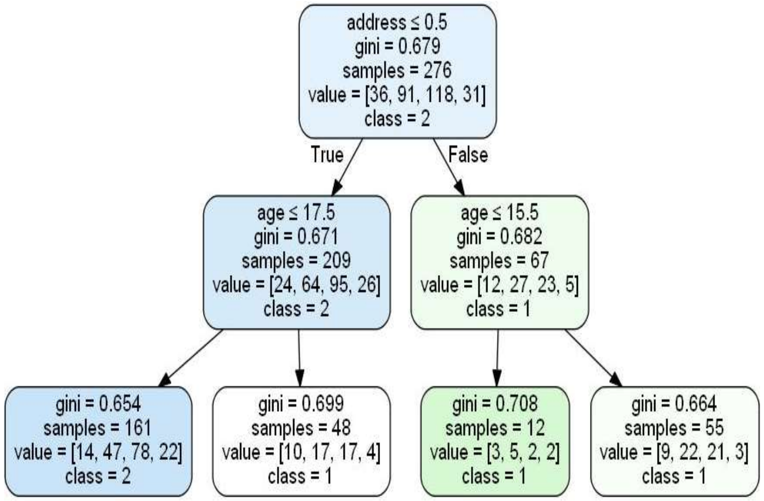
2. Rules-based Classification

- Traverse decision tree
- Nodes store threshold/condition, class label and count of each label in that decision path

=> Allows us to calculate accuracy and coverage

For our dataset

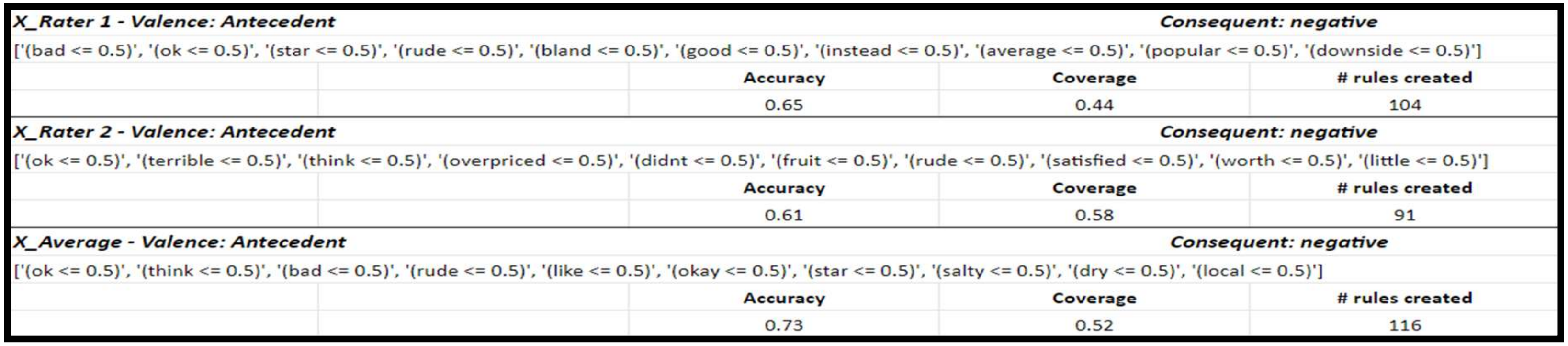
- Node threshold is 0.5
  - ➢ If a word in antecedent is >= 0.5, the word is present in a sentence
  - ➢ If a word is <0.5, the word is not present in a sentence



DATA MINING TECHNIQUES

- Advanced Classifications

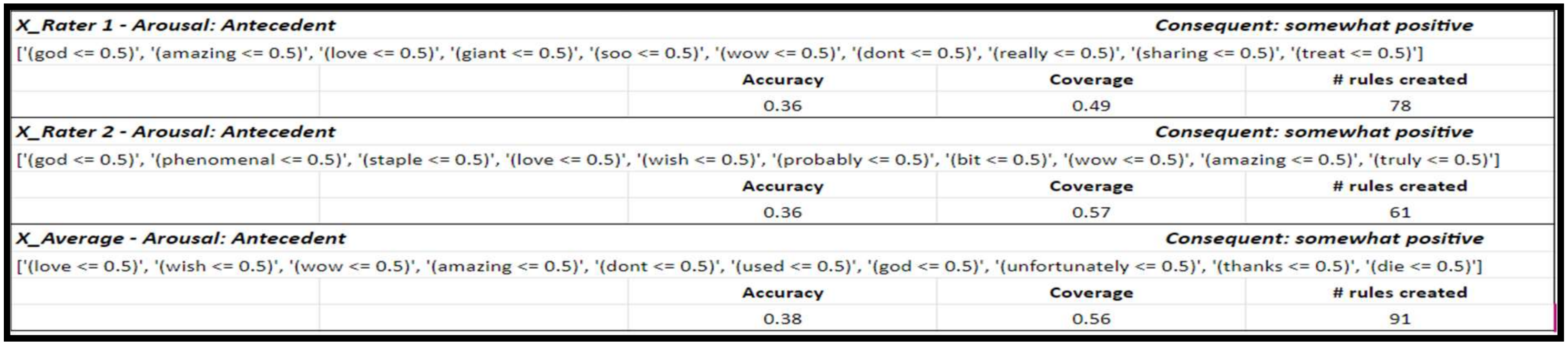
2. Rules-based Classification – Top rule for each target



X_Rater 1 - Valence: Antecedent			***Consequent: negative***	
['(bad <= 0.5)', '(ok <= 0.5)', '(star <= 0.5)', '(rude <= 0.5)', '(bland <= 0.5)', '(good <= 0.5)', '(instead <= 0.5)', '(average <= 0.5)', '(popular <= 0.5)', '(downside <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.65	0.44	104
X_Rater 2 - Valence: Antecedent			***Consequent: negative***	
['(ok <= 0.5)', '(terrible <= 0.5)', '(think <= 0.5)', '(overpriced <= 0.5)', '(didnt <= 0.5)', '(fruit <= 0.5)', '(rude <= 0.5)', '(satisfied <= 0.5)', '(worth <= 0.5)', '(little <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.61	0.58	91
X_Average - Valence: Antecedent			***Consequent: negative***	
['(ok <= 0.5)', '(think <= 0.5)', '(bad <= 0.5)', '(rude <= 0.5)', '(like <= 0.5)', '(okay <= 0.5)', '(star <= 0.5)', '(salty <= 0.5)', '(dry <= 0.5)', '(local <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.73	0.52	116

DATA MINING TECHNIQUES
- Advanced Classifications

2. Rules-based Classification – Top rule for each target



X_Rater 1 - Arousal: Antecedent				***Consequent: somewhat positive***
['(god <= 0.5)', '(amazing <= 0.5)', '(love <= 0.5)', '(giant <= 0.5)', '(soo <= 0.5)', '(wow <= 0.5)', '(dont <= 0.5)', '(really <= 0.5)', '(sharing <= 0.5)', '(treat <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.36	0.49	78
X_Rater 2 - Arousal: Antecedent				***Consequent: somewhat positive***
['(god <= 0.5)', '(phenomenal <= 0.5)', '(staple <= 0.5)', '(love <= 0.5)', '(wish <= 0.5)', '(probably <= 0.5)', '(bit <= 0.5)', '(wow <= 0.5)', '(amazing <= 0.5)', '(truly <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.36	0.57	61
X_Average - Arousal: Antecedent				***Consequent: somewhat positive***
['(love <= 0.5)', '(wish <= 0.5)', '(wow <= 0.5)', '(amazing <= 0.5)', '(dont <= 0.5)', '(used <= 0.5)', '(god <= 0.5)', '(unfortunately <= 0.5)', '(thanks <= 0.5)', '(die <= 0.5)']				
		Accuracy	**Coverage**	**# rules created**
		0.38	0.56	91

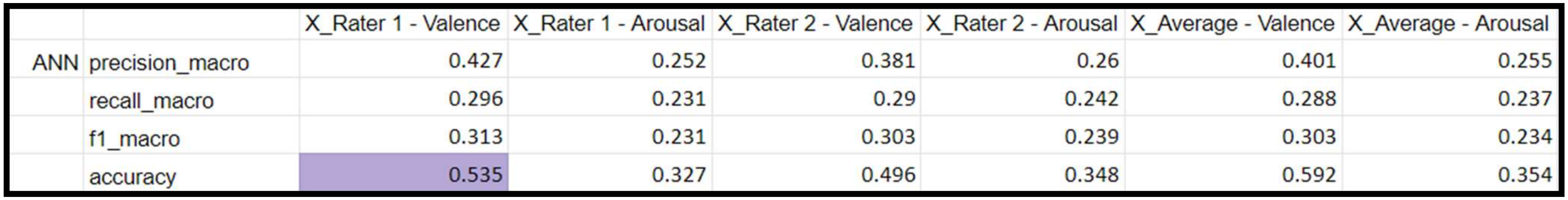
DATA MINING TECHNIQUES
- Advanced Classifications

3. Artificial Neural Networks (ANN)

- Train models using layers of nodes with "neurons"
- Iterative tuning process
- Creates unseen internal rules
- Takes normalized, vectorized, and encoded data

DATA MINING TECHNIQUES
- Advanced Classifications

3. Artificial Neural Networks (ANN) – Model Validation



		X_Rater 1 - Valence	X_Rater 1 - Arousal	X_Rater 2 - Valence	X_Rater 2 - Arousal	X_Average - Valence	X_Average - Arousal
ANN	precision_macro	0.427	0.252	0.381	0.26	0.401	0.255
	recall_macro	0.296	0.231	0.29	0.242	0.288	0.237
	f1_macro	0.313	0.231	0.303	0.239	0.303	0.234
	accuracy	0.535	0.327	0.496	0.348	0.592	0.354

- MLP Classifier Model was built with 3 layers - input, hidden, output
- Used "relu" activation function
- The hidden layer trained on 50 neurons
- Labels were encoded to match format of cluster labels
- With 5-Fold cross-validation, takes ~ 2+ hours to complete

RESULT

Which type of machine learning model is best suited for the task?

- Multinomial Naïve Bayes: accuracy better on average , favors valence rating
    - All valence models had accuracy score of 50% or better
- Bernoulli Naïve Bayes: average accuracy of 45%
    - All valence models had accuracy score of 50% or better
- Artificial Neural Network (MLP Classifier): average accuracy of 44%
    - Similar success with valence over arousal
    - Computationally expensive, evident in runtimes

RESULT

Is there an ML model that can reliably match the human rating system?

- None of the models perform well when learning the arousal sentiment
  - Best arousal prediction was Multinomial Naïve Bayes at 39.1%
- Precision, recall, and F1 scores are all low
- Human language is complex, varied, and full of slang

RESULT

Can we create an ML model that will provide insight into the reviews that might have influenced the human rating system?

- Rule-based organized frequent itemsets of words in logical and easy to follow manner
- Rules show a trend toward contextual matching of sentiment
- Relevance and relationship between words is easy to spot
- Not computationally expensive to create rules

SUMMARY

- We preprocess data by removing unnecessary words and characters, lemmatizing text, transform word to vectors
- We perform different Classification techniques to predict class labels of our 6 target attributes
  - ➢ We found that Multinomial Naïve Bayes was most suitable for our dataset
  - ➢ None of the techniques we implemented can reliably match the human rating system
  - ➢ Rule-based organized the frequent itemsets of words that can provide insights into the reviews
- We also perform anomaly detection to detect text anomalous to the dataset
- We learn to utilize documentation to implement different models
- We learn to preprocess and analyze textual data with different models

REFERENCES

Stemming vs Lemmatization | Baeldung on Computer Science

python - Traversal of sklearn decision tree - Stack Overflow

Text Classification with NLP: Tf-Idf vs Word2Vec vs BERT | by Mauro Di Pietro | Towards Data Science

Como funciona o algoritmo KNN » Computer Science Master

Machine Learning Series Day 3 (Naive Bayes) | by Alex Guanga | Becoming Human: Artificial Intelligence Magazine

Decision Tree Regression in 6 Steps with Python | by Samet Girgin | PursuitData | Medium

K-means images: Introduction to Data Mining 2nd Edition, Cluster Analysis: Basic Concepts and Algorithms