Classifying Amazon Reviews Based on Customer Ratings using Natural Language Processing

Capstone Project II for Springboard

Louie Balderrama
January 2019 Cohort

Background

- Reviews provide feedback to a product and are therefore inherently useful
- Every Amazon product review is summarized by a numerical rating
- But the heart of the feedback is in the text itself, not the rating

The Goal:

- To build a classifier that would understand the essence of a piece of review and assign it the most appropriate rating

Background

- The rating associated with every review is an integer from one to five stars
- Ratings serve as supervised, multi-class labels for classifier
- Review texts themselves are the core predictor

Exploration of Natural Language Processing (NLP) including:

- Word Embedding
- Topic Modeling
- Dimension Reduction

Amazon Dataset

- Contains customer reviews for all listed Electronics products from May 1996 up to July 2014
- 1,689,188 reviews by 192,403 customers on 63,001 unique products

	asin	helpful	overall	reviewText	reviewTime	reviewerID	reviewerName	summary	unixReviewTime
0	0528881469	[0, 0]	5	We got this GPS for my husband who is an (OTR)...	06 2, 2013	AO94DHGC771SJ	amazdnu	Gotta have GPS!	1370131200
1	0528881469	[12, 15]	1	I'm a professional OTR truck driver, and I bou...	11 25, 2010	AMO214LNFCEI4	Amazon Customer	Very Disappointed	1290643200
2	0528881469	[43, 45]	3	Well, what can I say. I've had this unit in m...	09 9, 2010	A3N7T0DY83Y4IG	C. A. Freeman	1st impression	1283990400
3	0528881469	[9, 10]	2	Not going to write a long review, even thought...	11 24, 2010	A1H8PY3QHMQQA0	Dave M. Shaw "mack dave"	Great grafics, POOR GPS	1290556800

Unique ID of the product, *str*

Number of users that voted helpful; Total number of users that voted on the review, *list*

Reviewer's rating of product, *int*

Review text itself, *str*

Time when review was posted, *str*

Unique ID of the reviewer, *str*

Specified name of reviewer, *str*

Headline summary of review, *str*

Unix time when review was posted, *str*

Data Wrangling – NLP Pre-processing

- The final dataframe for the model will be drawn from the `reviewText` column
- The `overall` column will serve as the ground truth labels

Sample review text:

I’m a big fan of the Brainwavz S1 (actually all of their headphones – have yet to be disappointed with any of their products). The S1 has been my main set for active use (e.g., workouts, runs, etc.) since the flat cable is very durable and resistant to tangles. The S5 keeps all the good features of the S1 and adds to it – the sound quality is richer and better defined.That’s not to say the S1 sounds poor – they are quite good, in fact. But the S5 are better. The highs are better defined and the midrange has more punch to it. The bass comes through clearly without moving into the harsh territory when the volume is pushed (as the S1s can do). The overall sound quality is very pleasing.The build quality seems solid – as solid as the S1 or better. I love the flat cable! I know that’s something that is not appreciated by everyone, but for me it’s been working out wonderfully. Although this (as most other Brainwavz headsets) comes with an excellent hard shell case, I usually tote my earbuds wrapped around my mp3 player in my pocket. Easy to carry; very stressful on the cables and can lead to tangles with round wires. Flat wires, especially those with a thick jacket such as these, survive that abuse with zero problems.The earbuds themselves are more sleekly shaped than the “can” style of the S1.Comfort is in line with the customary Brainwavz style, which is to say it’s outstanding. It comes with a wide range of tips to fit pretty much any ear, plus the Comply foam tips (which are my favorite). If fitted properly you end up with zero ear irritation plus excellent sound isolation and bass response.These are an over-the-ear design much like the S1. I never used that design prior to the S1 and it did take me a little time to get accustomed to it. It became second nature quickly, and the design is a lot more stable when exercising than the conventional in-ear design.These are more expensive than the S1, but you can hear the difference in price. Still, if you’re looking to keep the cost down a bit, the S1 is an excellent performer as well.Great sound, great comfort, wonderful cable design, and it comes with a solidly made case and lots of eartips. Highly recommend.[Sample provided for review]

NLP Pre-processing – HTML Entities

- Dataset includes data that predated the universal UTF-8 standard
- Some special characters like the apostrophe (`) are expressed as numbers in between `&#` and `;` due to HTML parsing
- RegEx is used to drop tokens that fit the `\&\#[0-9]+\;` pattern:

I’m a big fan of the Brainwavz S1 (actually all of their headphones – have yet to be disappointed with any of their products). The S1 has been my main set for active use (e.g., workouts, runs, etc.) since the flat cable is very durable and resistant to tangles. The S5 keeps all the good features of the S1 and adds to it – the sound quality is richer and better defined.That’s not to say the S1 sounds poor – they are quite good, in fact. But the S5 are better. The highs are better defined and the midrange has more punch to it. The bass comes through clearly without moving into the harsh territory when the volume is pushed (as the S1s can do). The overall sound quality is very pleasing.The build quality seems solid – as solid as the S1 or better. I love the flat cable! I know that’s something that is not appreciated by everyone, but for me it’s been working out wonderfully. Although this (as most other Brainwavz headsets) comes with an excellent hard shell case, I usually tote my earbuds wrapped around my mp3 player in my pocket. Easy to carry; very stressful on the cables and can lead to tangles with round wires. Flat wires, especially those with a thick jacket such as these, survive that abuse with zero problems.The earbuds themselves are more sleekly shaped than the “can” style of the S1.Comfort is in line with the customary Brainwavz style, which is to say it’s outstanding. It comes with a wide range of tips to fit pretty much any ear, plus the Comply foam tips (which are my favorite). If fitted properly you end up with zero ear irritation plus excellent sound isolation and bass response.These are an over-the-ear design much like the S1. I never used that design prior to the S1 and it did take me a little time to get accustomed to it. It became second nature quickly, and the design is a lot more stable when exercising than the conventional in-ear design.These are more expensive than …

NLP Pre-processing – Lemmatization

- Words are reduced to their root form to ensure word usage consistency
- If *learning* is differentiated from its base version *learn*, we lose relational context between two documents that use either word
- Lemmatization is a technique that takes into account context similarity according to part-of-speech anatomy
- Stemming is another common approach although stemming only performs truncation

NLP Pre-processing – Lemmatization

- *WordNetLemmatizer* from NLTK library is used:

Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of their product The S1 have be my main set for active use e g workouts run etc since the flat cable be very durable and resistant to tangle The S5 keep all the good feature of the S1 and add to it the sound quality be rich and well define Thats not to say the S1 sound poor they be quite good in fact But the S5 be well The high be well define and the midrange have more punch to it The bass come through clearly without move into the harsh territory when the volume be push as the S1s can do The overall sound quality be very please The build quality seem solid as solid as the S1 or good I love the flat cable I know thats something that be not appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz headset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pocket Easy to carry ; very stressful on the cable and can lead to tangle with round wire Flat wire especially those with a thick jacket such as these survive that abuse with zero problem The earbuds themselves be more sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which be to say its outstanding It come with a wide range of tip to fit pretty much any ear plus the Comply foam tip which be my favorite If fit properly you end up with zero ear irritation plus excellent sound isolation and bass response These be an over-the-ear design much like the S1 I never use that design prior to the S1 and it do take me a little time to get accustom to it It become second nature quickly and the design be a lot more stable when exercise than the conventional in-ear design These be more expensive than the S1 but you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excellent performer as well Great sound great comfort wonderful cable design and it come with a solidly make case and lot of eartips Highly recommend [Sample provide for review]

NLP Pre-processing – Accents

- Each review is normalized from longform UTF-8 to ASCII encoding
- This removes accents in characters so words like *naïve* will simply be interpreted as *naive*:

Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of their product The S1 have be my main set for active use e g workouts run etc since the flat cable be very durable and resistant to tangle The S5 keep all the good feature of the S1 and add to it the sound quality be rich and well define Thats not to say the S1 sound poor they be quite good in fact But the S5 be well The high be well define and the midrange have more punch to it The bass come through clearly without move into the harsh territory when the volume be push as the S1s can do The overall sound quality be very please The build quality seem solid as solid as the S1 or good I love the flat cable I know thats something that be not appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz headset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pocket Easy to carry ; very stressful on the cable and can lead to tangle with round wire Flat wire especially those with a thick jacket such as these survive that abuse with zero problem The earbuds themselves be more sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which be to say its outstanding It come with a wide range of tip to fit pretty much any ear plus the Comply foam tip which be my favorite If fit properly you end up with zero ear irritation plus excellent sound isolation and bass response These be an over-the-ear design much like the S1 I never use that design prior to the S1 and it do take me a little time to get accustom to it It become second nature quickly and the design be a lot more stable when exercise than the conventional in-ear design These be more expensive than the S1 but you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excellent performer as well Great sound great comfort wonderful cable design and it come with a solidly make case and lot of eartips Highly recommend [Sample provide for review]

NLP Pre-processing – Punctuations

- The preprocessed reviews are further cleaned by dropping punctuations
- Only spaces and alphanumeric characters are kept by replacing all RegEx pattern `[^\w\s]` matches with a whitespace:

Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of their product The S1 have be my main set for active use e g workouts run etc since the flat cable be very durable and resistant to tangle The S5 keep all the good feature of the S1 and add to it the sound quality be rich and well define Thats not to say the S1 sound poor they be quite good in fact But the S5 be well The high be well define and the midrange have more punch to it The bass come through clearly without move into the harsh territory when the volume be push as the S1s can do The overall sound quality be very please The build quality seem solid as solid as the S1 or good I love the flat cable I know thats something that be not appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz headset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pocket Easy to carry ; very stressful on the cable and can lead to tangle with round wire Flat wire especially those with a thick jacket such as these survive that abuse with zero problem The earbuds themselves be more sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which be to say its outstanding It come with a wide range of tip to fit pretty much any ear plus the Comply foam tip which be my favorite If fit properly you end up with zero ear irritation plus excellent sound isolation and bass response These be an over-the-ear design much like the S1 I never use that design prior to the S1 and it do take me a little time to get accustom to it It become second nature quickly and the design be a lot more stable when exercise than the conventional in-ear design These be more expensive than the S1 but you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excellent performer as well Great sound great comfort wonderful cable design and it come with a solidly make case and lot of eartips Highly recommend [Sample provide for review]

NLP Pre-processing – Lowercasing

- Every letter is converted to lowercase
- This makes it so that *iPhone* will not be distinguishable from *iphone*

Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of their product The S1 have be my main set for active use e g workouts run etc since the flat cable be very durable and resistant to tangle The S5 keep all the good feature of the S1 and add to it the sound quality be rich and well define Thats not to say the S1 sound poor they be quite good in fact But the S5 be well The high be well define and the midrange have more punch to it The bass come through clearly without move into the harsh territory when the volume be push as the S1s can do The overall sound quality be very please The build quality seem solid as solid as the S1 or good I love the flat cable I know thats something that be not appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz headset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pocket Easy to carry very stressful on the cable and can lead to tangle with round wire Flat wire especially those with a thick jacket such as these survive that abuse with zero problem The earbuds themselves be more sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which be to say its outstanding It come with a wide range of tip to fit pretty much any ear plus the Comply foam tip which be my favorite If fit properly you end up with zero ear irritation plus excellent sound isolation and bass response These be an over-the-ear design much like the S1 I never use that design prior to the S1 and it do take me a little time to get accustom to it It become second nature quickly and the design be a lot more stable when exercise than the conventional in ear design These be more expensive than the S1 but you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excellent performer as well Great sound great comfort wonderful cable design and it come with a solidly make case and lot of eartips Highly recommend Sample provide for review

NLP Pre-processing – Stop Words

- Stop words consist of most commonly used words that include:
  - pronouns (*us*, *she*, *their*)
  - articles (*a*, *an*, *the*)
  - prepositions (*under*, *from*, *off*)
- Stop words are not helpful in distinguishing a document from another and are therefore dropped

im a big fan of the brainwavz s1 actually all of their headphone have yet to be disappoint with any of their product the s1 have be my main set for active use e g workouts run etc since the flat cable be very durable and resistant to tangle the s5 keep all the good feature of the s1 and add to it the sound quality be rich and well define thats not to say the s1 sound poor they be quite good in fact but the s5 be well the high be well define and the midrange have more punch to it the bass come through clearly without move into the harsh territory when the volume be push as the s1s can do the overall sound quality be very please the build quality seem solid as solid as the s1 or good i love the flat cable i know thats something that be not appreciate by everyone but for me its be work out wonderfully although this as most other brainwavz headset come with an excellent hard shell case i usually tote my earbuds wrap around my mp3 player in my pocket easy to carry very stressful on the cable and can lead to tangle with round wire flat wire especially those with a thick jacket such as these survive that abuse with zero problem the earbuds themselves be more sleekly shape than the can style of the s1 comfort be in line with the customary brainwavz style which be to say its outstanding it come with a wide range of tip to fit pretty much any ear plus the comply foam tip which be my favorite if fit properly you end up with zero ear irritation plus excellent sound isolation and bass response...

NLP Pre-processing – Single Whitespaces

- RegEx pattern `[\s]+` is used to ensure no more than a single whitespace separates words in sentences:

im big fan brainwavz s1 actually headphone yet disappoint product s1 main set active use e g workouts run etc since flat cable durable resistant tangle s5 keep good feature s1 add sound quality rich well define thats say s1 sound poor quite good fact s5 well high well define midrange punch bass come clearly without move harsh territory volume push s1s overall sound quality please build quality seem solid solid s1 good love flat cable know thats something appreciate everyone work wonderfully although brainwavz headset come excellent hard shell case usually tote earbuds wrap around mp3 player pocket easy carry stressful cable lead tangle round wire flat wire especially thick jacket survive abuse zero problem earbuds sleekly shape style s1 comfort line customary brainwavz style say outstanding come wide range tip fit pretty much ear plus comply foam tip favorite fit properly end zero ear irritation plus excellent sound isolation bass response ear design much like s1 never use design prior s1 take little time get accustom become second nature quickly design lot stable exercise conventional ear design expensive s1 hear difference price still look keep cost bit s1 excellent performer well great sound great comfort wonderful cable design come solidly make case lot eartips highly recommend sample provide review

Tokenization

- The pre-processed reviews make up our *corpora*, which is simply the collection of all our documents
- Each review is *tokenized* or transformed into an ordered list of words

Tokenized sample review text:

['im', 'big', 'fan', 'brainwavz', 's1', 'actually', 'headphone', 'yet', 'disappoint', 'product', 's1', 'main', 'set', 'active', 'use', 'e', 'g', 'workouts', 'run', 'etc', 'since', 'flat', 'cable', 'durable', 'resistant', 'tangle', 's5', 'keep', 'good', 'feature', 's1', 'add', 'sound', 'quality', 'rich', 'well', 'define', 'thats', 'say', 's1', 'sound', 'poor', 'quite', 'good', 'fact', 's5', 'well', 'high', 'well', 'define', 'midrange', 'punch', 'bass', 'come', 'clearly', 'without', 'move', 'harsh', 'territory', 'volume', 'push', 's1s', 'overall', 'sound', 'quality', 'please', 'build', 'quality', 'seem', 'solid', 'solid', 's1', 'good', 'love', 'flat', 'cable', 'know', 'thats', 'something', 'appreciate', 'everyone', 'work', 'wonderfully', 'although', 'brainwavz', 'headset', 'come', 'excellent', 'hard', 'shell', 'case', 'usually', 'tote', 'earbuds', 'wrap', 'around', 'mp3', 'player', 'pocket', 'easy', 'carry', 'stressful', 'cable', 'lead', 'tangle', 'round', 'wire', 'flat', 'wire', 'especially', 'thick', 'jacket', 'survive', 'abuse', 'zero', 'problem', 'earbuds', 'sleekly', 'shape', 'style', 's1', 'comfort', 'line', 'customary', 'brainwavz', 'style', 'say', 'outstanding', 'come', 'wide', 'range', 'tip', 'fit', 'pretty', 'much', 'ear', 'plus', 'comply', 'foam', 'tip', 'favorite', 'fit', 'properly', 'end', 'zero', 'ear', 'irritation', 'plus', 'excellent', 'sound', 'isolation', 'bass', 'response', 'ear', 'design', 'much', 'like', 's1', 'never', 'use', 'design', 'prior', 's1', 'take', 'little', 'time', 'get', 'accustom', 'become', 'second', 'nature', 'quickly', 'design', 'lot', 'stable', 'exercise', 'conventional', 'ear', 'design', 'expensive', 's1', 'hear', 'difference', 'price', 'still', 'look', 'keep', 'cost', 'bit', 's1', 'excellent', 'performer', 'well', 'great', 'sound', 'great', 'comfort', 'wonderful', 'cable', 'design', 'come', 'solidly', 'make', 'case', 'lot', 'eartips', 'highly', 'recommend', 'sample', 'provide', 'review']

Phrase Modeling – Bigrams

- *Phrases* are neighboring words that appear to convey one meaning as though they are a single word, like *smart TV*
- Gensim's built-in phraser is used
- Higher phraser threshold means the more often two words must appear adjacent to be grouped into a phrase

Sample bigram phrases:

['hdmi_dvi', 'lens_without', 'time_forget', 'like_return', '2_00', 'fast_run', 'make_convenient', 'point_think', 'matter_fact', 'although_make', 'actually_see', 'sure_problem', 'course_good', 'get_catch', 'take_find', 'include_product', 'problem_design', 'work_everything', 'standard_camera', '1080p_120hz', 'make_give', 'set_ipad', 'control_cable', 'nikon_brand', 'really_beat', 'game_also', 'tiny_size', 'tiny_camera', 'use_default', 'color_come', 'get_12', 'plug_network', 'piece_technology', 'light_fit', 'button_click', '4kb_qd', 'wheel_click', 'wish_purchase', 'hold_device', 'ipod_phone', 'might_break', 'work_need', 'big_small', 'tell_would', 'lot_high', 'noise_ratio', 'less_200', 'star_seem', 'design_camera', 'camera_function']

Phrase Modeling – Trigrams

- Trigrams are generated by applying another gensim phraser on top of a bigram phraser
- If *sd* and *card* appear together often enough per the set parameters, the phraser groups them together as *sd_card*
- If *sd_card* appears adjacent to the token *reader* in enough instances, the phraser further links them together as *sd_card_reader*

Sample trigram phrases:

['play_blu_ray', 'samsung_galaxy_s4', 'old_macbook_pro', 'quality_top_notch', 'b_w_filter', 'one_living_room', 'mac_os_x', 'far_exceed_expectation', 'nexus_7_2013', 'cell_phone_use', 'customer_service_great', '5d_mark_iii', 'cell_phone_camera', 'macbook_pro_work', 'first_blu_ray', 'case_nexus_7', 'double_sided_tape', 'price_highly_recommended', 'almost_non_existent', '2_4ghz_5ghz', 'macbook_pro_13', 'customer_service_rep', 'samsung_840_pro', 'blu_ray_disk', 'use_third_party', 'n_uuml_vi', 'home_theater_pc', 'complete_waste_money', 'small_form_factor', 'use_home_theater', 'fast_forward_rewind', 'wi_fi_connection', 'amazon_return_policy', 'new_kindle_fire', '192_168_1', 'aps_c_sensor', 'ear_bud_come', 'mp3_player_work', 'mp3_player_use', 'use_macbook_pro', 'run_os_x', 'canon_5d_mark', 'blu_ray_movie', 'western_digital_passport', 'dd_wrt_firmware', 'inch_macbook_pro', 'heart_rate_monitor', 'great_mp3_player', 'kindle_fire_hd', 'samsung_galaxy_tab']

Count-based Feature Engineering

- For a machine learning model to work with text input, documents must first be *vectorized* or converted into containers of numerical values
- *Bag of Words* is the classical approach of getting token frequency per document

Sample Bag of Words model:

Word: address	Frequency: 1	Word: earlier	Frequency: 1
Word: around	Frequency: 1	Word: ease	Frequency: 2
Word: arrive	Frequency: 1	Word: ect	Frequency: 1
Word: back	Frequency: 1	Word: email	Frequency: 2
Word: bad	Frequency: 1	Word: exception	Frequency: 1
Word: big	Frequency: 2	Word: exchange	Frequency: 1
Word: come	Frequency: 1	Word: expect	Frequency: 1
Word: contact	Frequency: 1	Word: freeze	Frequency: 2
Word: could	Frequency: 1	Word: get	Frequency: 1
Word: day	Frequency: 1	Word: glitch	Frequency: 1

Count-based Feature Engineering

- *Term Frequency-Inverse Document Frequency* (TF-IDF) is another approach where continuous values are assigned to tokens
- Words that appear frequently overall are weighted lower because they do not establish saliency in a document
- Words that are unique to some but not all documents are weighted higher because they help distinguish the documents from the others

Sample TF-IDF model:

```
Word: address          Weight: 0.113          Word: come             Weight: 0.046
Word: around           Weight: 0.060          Word: contact          Weight: 0.103
Word: arrive           Weight: 0.093          Word: could            Weight: 0.054
Word: back             Weight: 0.051          Word: day              Weight: 0.061
Word: bad              Weight: 0.068          Word: earlier          Weight: 0.141
Word: big              Weight: 0.126          Word: ease             Weight: 0.220
```

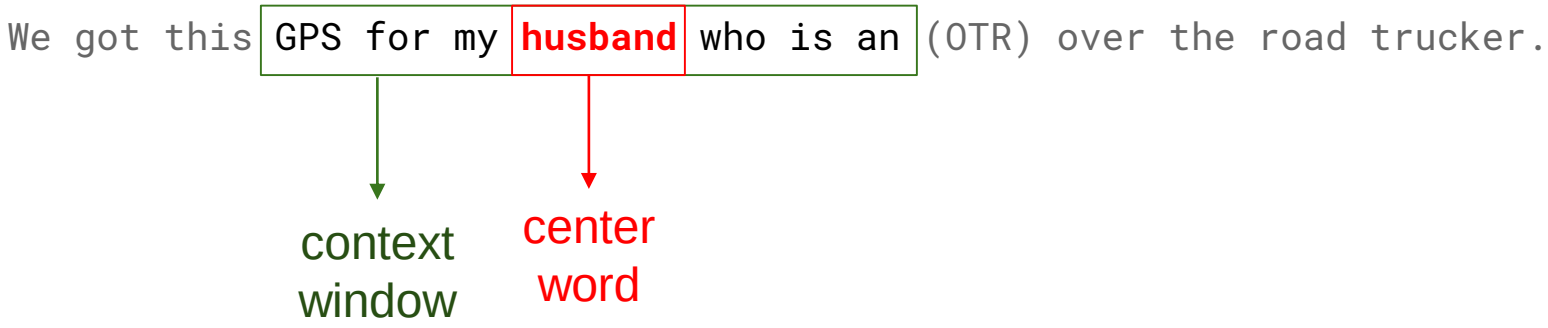
Word Embedding for Feature Engineering

- Count-based techniques, however, do not give regard to word sequence and sentence structure, and thus lose semantics
- The *Word2Vec* technique, on the other hand, actually embeds meaning in vectors by quantifying how often a word appears within the vicinity of a given set of other words.

token	0	1	2	...	97	98	99
get	2.027105	2.285539	0.325559	...	0.786811	3.543153	0.554558
gps	4.430076	-1.91234	9.017261	...	2.411811	0.086889	-9.82164
husband	1.160196	4.993967	1.216444	...	-1.11843	0.796972	-4.29458
otr	1.308931	-1.70058	2.168358	...	-0.60044	2.36648	-1.84929
road	3.249146	2.961001	10.25173	...	1.872263	-1.10259	0.958208

Feature Engineering – Word2Vec

- A *context window* slides across every document one token at a time
- In each step, the *center word* is described by its adjacent words and the probability that the token appears together with the others is expressed in multiple interacting dimensions



Final Dataframe

- All unique tokens in the entire corpora are vectorized in 100 dimensions making up the `word_vec_df` vocabulary
- The `word_vec_df` is sliced by the words that appear in a given review and the mean is taken along each of the 100 dimensions
- This singularizes multiple word embeddings into one observation for each product review
- Finally, the ground truth label is imposed from the original Amazon dataset's `overall` series to create the final `model_df` dataframe

INPUT REVIEW

We got this GPS for my husband who is an...

PRE-PROCESSED

["get", "gps", "husband", ...]

SLICED WORD_VEC_DF

token	0	1	2	...	97	98	99
get	2.027105	2.285539	0.325559	...	0.786811	3.543153	0.554558
gps	4.430076	-1.91234	9.017261	...	2.411811	0.086889	-9.82164
husband	1.160196	4.993967	1.216444	...	-1.11843	0.796972	-4.29458
...	...	...	...	...	...	...	...

MEAN

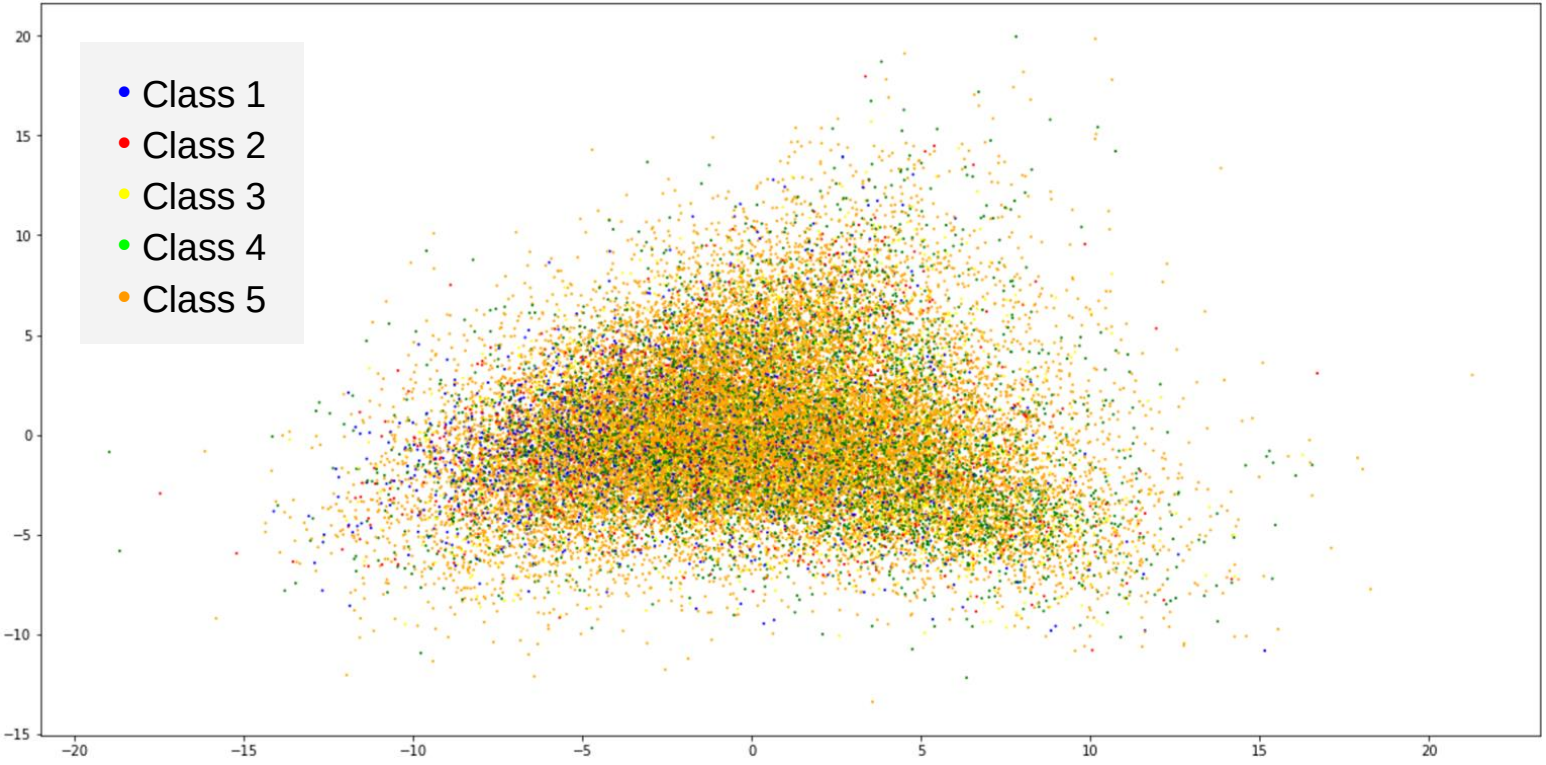
MODEL_DF

review_id	feature_0	feature_1	...	feature_98	feature_99	label
0	0.456036	-0.340525	...	1.381710	-0.931236	5

Principal Component Analysis

- *Principal Component Analysis* (PCA) is a dimensionality reduction technique that helps visualize if there is a clear decision boundary along the five overall rating classifications
- The more that the data appear clustered, the higher the likelihood that our machine learning model is more effective
- PCA was used to reduce our 100 dimensions in our model to just 2
- However, our 2D plot produced no evident decision boundary due to high complexity

Principal Component Analysis



Exploratory Data Analysis – Word Similarity

- Since Word2Vec quantifies tokens, we can derive *word similarity* by finding the distance between tokens across all dimensions
- The similarity comes from how often these tokens appear in the same window of words as their counterpart

```
nook:    ['kindle', 'ereader', 'nookcolor', 'kobo', 'paperwhite']
phone:   ['cellphone', 'smartphone', 'cell', 'droid', 'iphone']
tv:      ['television', 'hdtv', 'tvs', 'vizio', 'flatscreen']
good:    ['decent', 'great', 'wise', 'excellent', 'descent']
price:   ['pricing', 'cost', 'priced', 'pricetag', 'expensive']
```

Exploratory Data Analysis – Word Algebra

- Word vectors can also be added or subtracted
- To add is to combine the meaning of the components
- To subtract is to take out the context of one token from another

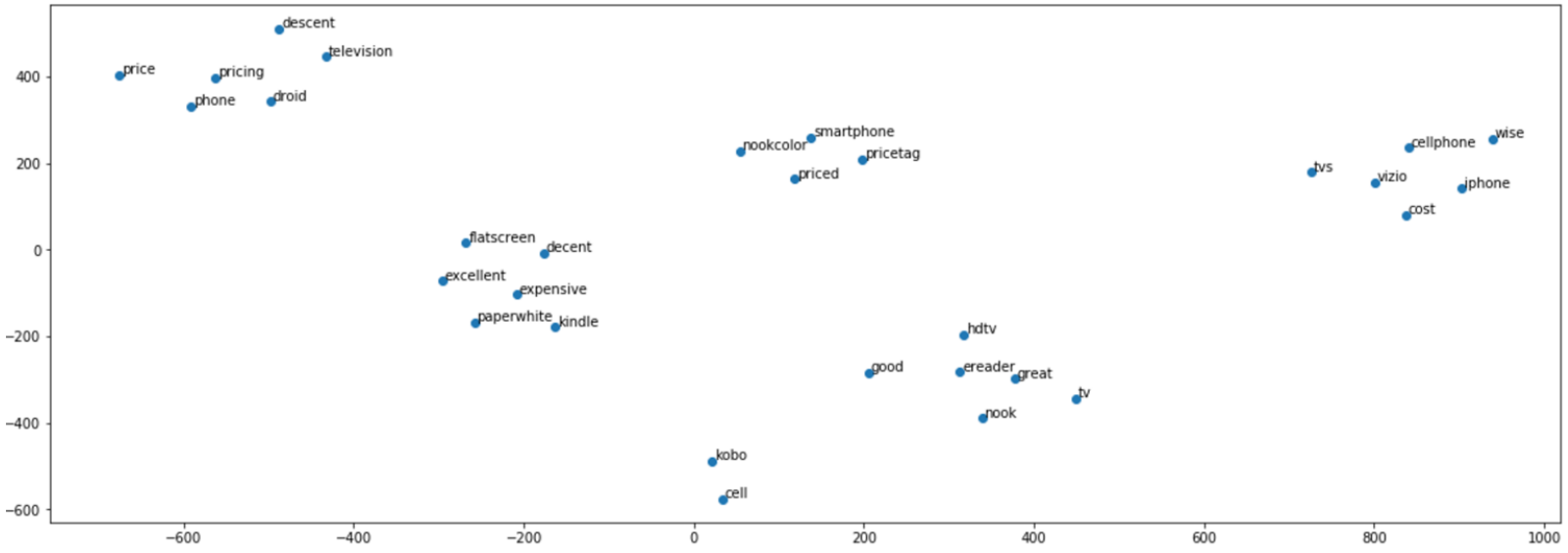
```
Books + Touchscreen = ebooks

Cheap – Quality = ebay

Tablet – Phone = netbooks
```

Exploratory Data Analysis – t-SNE

- Like PCA, *t-Distributed Stochastic Neighbor Embedding* (t-SNE) is another technique for reducing high-dimensional data for 2D visualization
- PCA is used to perceive word similarity in terms of spatial distance in a plot



Exploratory Data Analysis – NER

- While gensim can perform word tagging to identify part-of-speech, spaCy can go further and identify what nouns in the documents refer to
- *Named-Entity Recognition* (NER) classification tags include distinguishing persons, organizations, products, places, dates, etc.

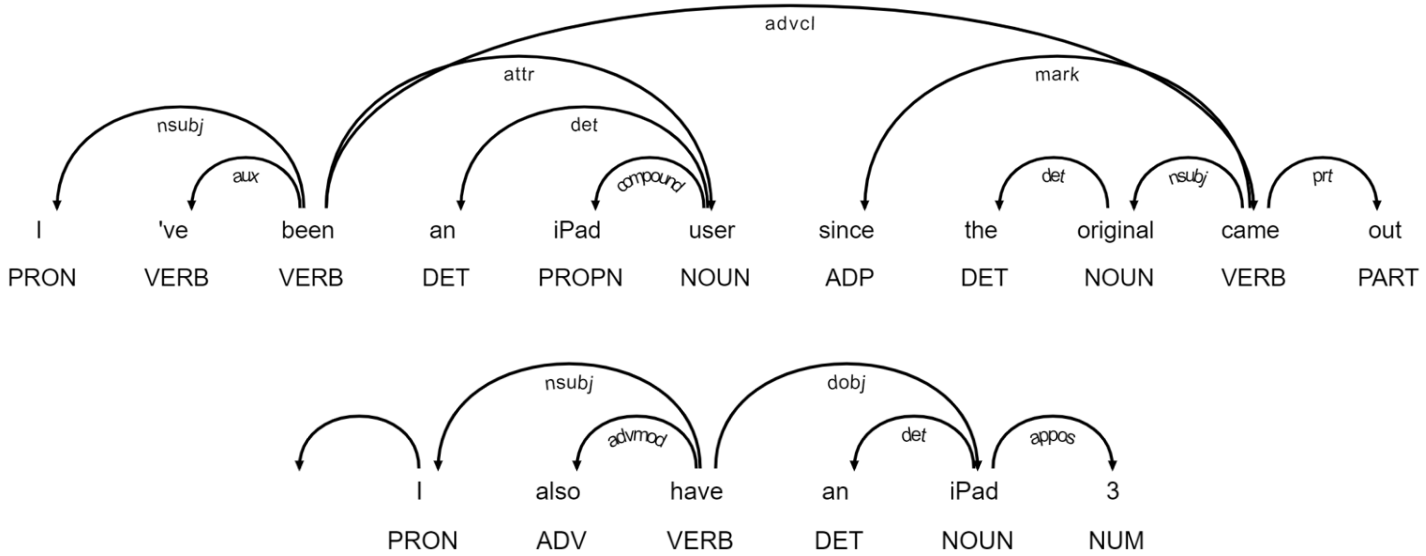
```
     ORG:  [iPad, OS, Amazon, Samsung Galaxy S3, iSyncr, Apple, Google]
    DATE:  [the past few years, the past 24, 9/20/12, Two days later]
   MONEY:  [500, 200, 15, another $200+]
  PERSON:  [Retina Display, Netflix, Screen Glare, Battery HD]
 ORDINAL:  [first]
CARDINAL:  [two, 46, two, 5, 4]
     GPE:  [Caltrain, Palo Alto, iTunes, Cloud, Bluray, mp4, LA, SF]
```

Exploratory Data Analysis – NER

I've been an iPad ORG user since the original came out. I also have an iPad 3. I have worked in IT for the past few years DATE so I would say I am pretty good with technology and fancy new devices. With that introduction out of the way, I will be reviewing key points that I have seen touched upon in other reviews. Here goes...BUILDThe device feels nice and solid. I'm a little surprised at how heavy it is, but that's not necessarily a bad thing. The rubberized backing is always nice for added grip. It's not as nice as say unibody aluminum, but it's not $ 500 MONEY either.SCREENThe screen is fantastic. But my problem is the same as when iPad ORG got Retina Display PERSON , other than the OS ORG , most apps look rather pixelated. A lot of the games I tried are not high definition, at least not high enough to look smooth on this screen. Hopefully apps get updated to higher resolutions. LOCK SCREEN ORG ADSYeah there are ads on my lock screen. I'm not sure why this is such a big deal. How much time do people really spend looking at the lock screen? The first ORDINAL thing I thought when I saw the ads is WOW the pictures are really crisp! The ads are there to subsidize some of the $ 200 MONEY price tag. I might pay the $ 15 MONEY to get rid of them so I can customize it, but I might not. I feel like this has been blown out of proportion by other customers.SOUNDThe sound from the speakers is great. Much better than you would get from more expensive devices, very crisp and clean. I have the official Amazon ORG case on and it has not affected the sound at all. Nothing much else to say, I doubt anyone will complain about this.CRASHINGI've had two CARDINAL apps crash on opening. I don't know if it is the app or the OS ORG . It's probably somewhere in the middle. Again, not a big deal for me. If it crashes, then I just tap it again and it works. I've also watched a few movies using the built in player as well as Netflix PERSON and Amazon ORG Prime. No crashes for me at all. I'm sure OS stability will be improved as time goes on.OVERALL SATISFACTIONCompared to my iPad 3, obviously the Fire HD is not as "good" so to speak. I mainly got it because I wanted something smaller. I also mainly used the iPad ORG to surf the web, watch videos, and play some simple games. The Fire HD accomplishes this and does so much more. If you are expecting an iPad ORG killer, or a desktop replacement, or a productivity machine, then you should look elsewhere.I bought this to be a media device, and I believe that is what Amazon ORG meant this to be. In this regard, I think this is a great device. In fact, I decided to keep this and sell my iPad ORG 3, which will give me another $200+ MONEY to spend on other things. Just remember, this device is not for everyone. If you want a media device, you will be happy with this. Do not expect an iPad ORG for $ 200.UPDATE MONEY 9/18/12Just wanted to add a few more things I have noticed over the past 24 DATE hours.- Power/Volume Buttons ORG : There are a bit hard to press, which is somewhat alleviated by having the official case. Maybe it's because I'm a longtime iPad ORG user, but this will definitely take some getting used to.- PERSON Screen Glare PERSON : It took me a little while to notice, but I was playing a Seek & Find ORG game while on Caltrain GPE , with the bright Palo Alto GPE sun shining right on me, and didn't have any trouble seeing the screen. I remembered that Amazon ORG mentioned how the screen was changed to reduce glare, and they did an amazing job.UPDATE 9/20/12Two days later DATE and I am still very happy

Exploratory Data Analysis – Dependency Tree

- The capability of spaCy's NER is based on deciphering sentence structure
- It breaks down how tokens interact with and influence each other



Exploratory Data Analysis – Topic Modeling

- Reviews can be grouped together according to the type of electronics product they correspond to via *Latent Dirichlet Allocation* (LDA)
- LDA is an unsupervised learning model that clusters documents together according to topic
- Product reviews will have weights assigned to each of the topic and the topics themselves will have weights on every token

```
            Words Salient to Topic 1:
TOKEN: sound      WEIGHT: 0.0357492938637733 46
TOKEN: speaker    WEIGHT: 0.0163549687713384 63
TOKEN: headphone  WEIGHT: 0.0152490502223372 46
TOKEN: good       WEIGHT: 0.0149301188066601 75
TOKEN: music      WEIGHT: 0.0124199781566858 30
```

Exploratory Data Analysis – Topic Modeling

sound, 0.035749293863773346
speaker, 0.016354968771338463
headphone, 0.015249050222337246
good, 0.014930118806660175
music, 0.01241997815668583

drive, 0.015999674797058105
use, 0.011478731408715248
router, 0.010722918435931206
work, 0.01045207493007183
get, 0.010170512832701206

case, 0.032898712903261185
cover, 0.012469933368265629
fit, 0.0111733078956604
use, 0.011004170402884483
like, 0.010999665595591068

work, 0.03235096111893654
cable, 0.0308877881616354
great, 0.020522940903902054
one, 0.019090095534920692
product, 0.018262816593050957

speaker, 0.01458591129630804
mount, 0.013087787665426731
monitor, 0.011773170903325081
light, 0.010449431836605072
system, 0.009331470355391502

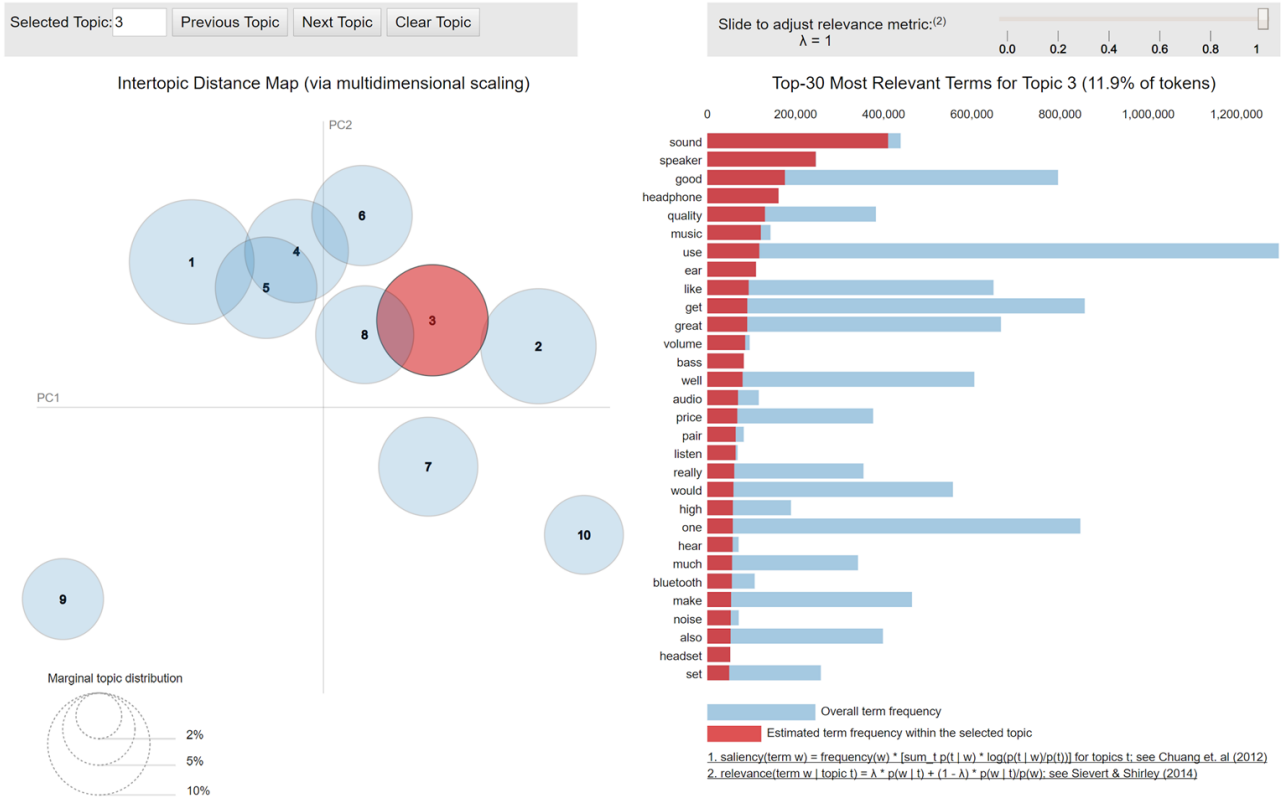
battery, 0.02889261767268181
charge, 0.026525508612394333
use, 0.017796754837036133
phone, 0.017193373292684555
charger, 0.011224666610360146

camera, 0.04835493117570877
lens, 0.018904486671090126
use, 0.01494103018194437
good, 0.010452806949615479
take, 0.010180431418120861

use, 0.022391512989997864
keyboard, 0.021374180912971497
tablet, 0.014760561287403107
mouse, 0.011335919611155987
like, 0.010733792558312416

tv, 0.030175620689988136
remote, 0.01235356554389
use, 0.011232638731598854
get, 0.009809658862650394
video, 0.009135127998888493

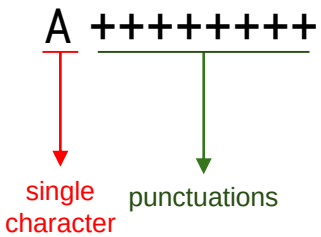
Exploratory Data Analysis – Topic Modeling



Machine Learning – Dealing with NaNs

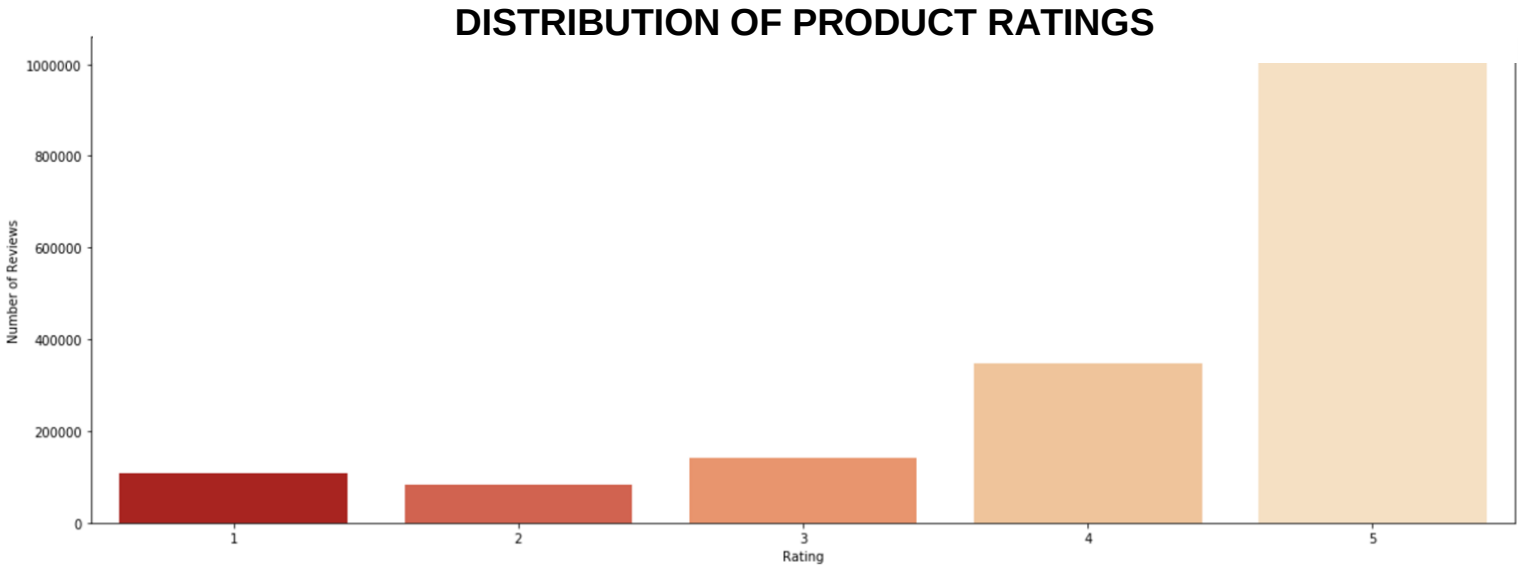
- NaN values are dropped from `model_df`
- NaN values occur when a tokenized document is empty
- The review below is invalidated because only alphanumeric characters were kept during pre-processing, leaving us with just *A*
- However, we've chosen that single characters do not get tokenized in our process thus creating a NaN value

Sample review text that will become NaN:



Machine Learning – Dealing with Class Imbalance

- The distribution of ratings shows that, in general, users highly approve of products bought on Amazon
- This, however, gives us a highly imbalanced dataset:

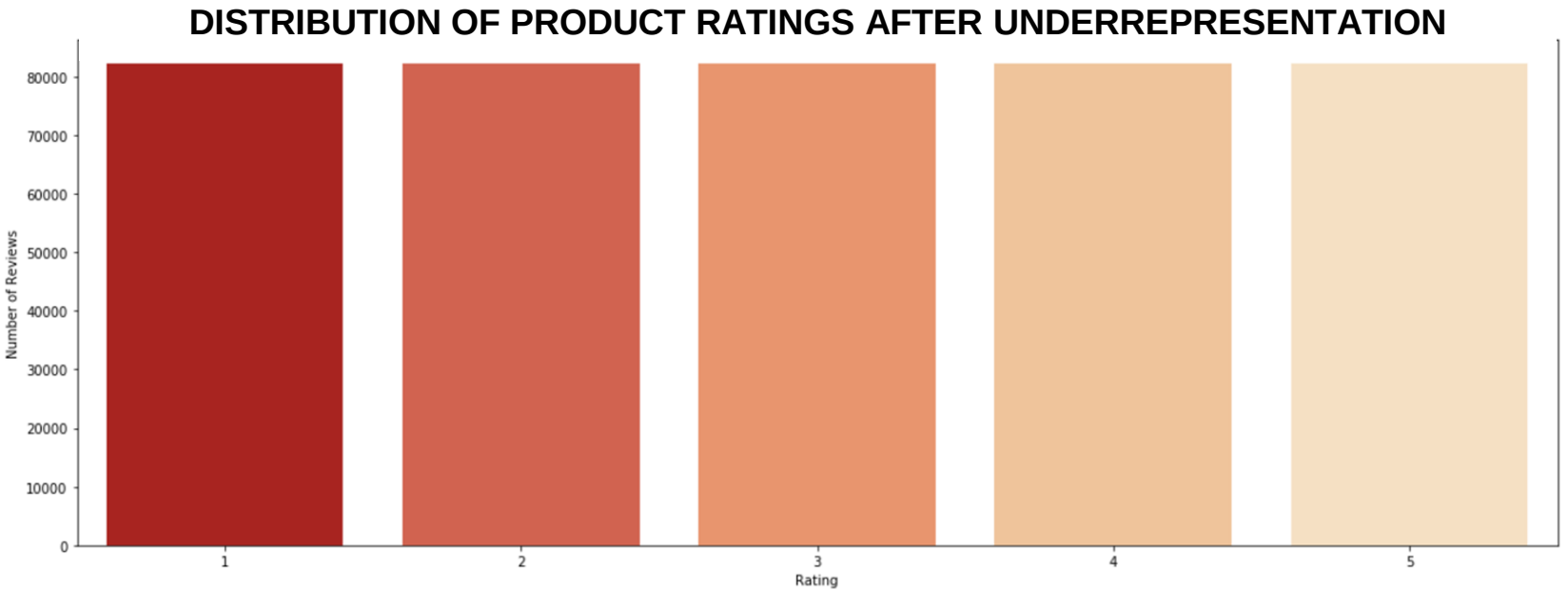


Machine Learning – Dealing with Class Imbalance

- *Synthetic Minority Over-sampling Technique* (SMOTE) can over-represent by bootstrapping the minority classes to match the size of the majority classes
- SMOTE clusters classes then performs bootstrapping to produce random samples
- However, this is computationally very expensive due to the massive class imbalance in our data
- But since our dataset is huge, we can afford to sample every class and still have significant amount of data for the model
- We therefore opt to under-represent the majority class according to our most minority class

Machine Learning – Dealing with Class Imbalance

- Trimming every other classes to the smallest class balances the data for the training and the testing phases:



Machine Learning – Metrics and Baseline

- *Accuracy* will identify how many reviews are correctly labeled by the model
  - There are five ratings and thus five classes
  - No review can have two or more ratings
  - The probability that a correct prediction is made from pure guesswork is 20%
- *F1 Score* is taking precision and recall into consideration
- The baseline is for when a model only randomly guesses the output labels – in this case, if every prediction is the same class

Baseline Scores:

```
Baseline Accuracy:   20.000%
 Baseline F1 Score:   0.200
```

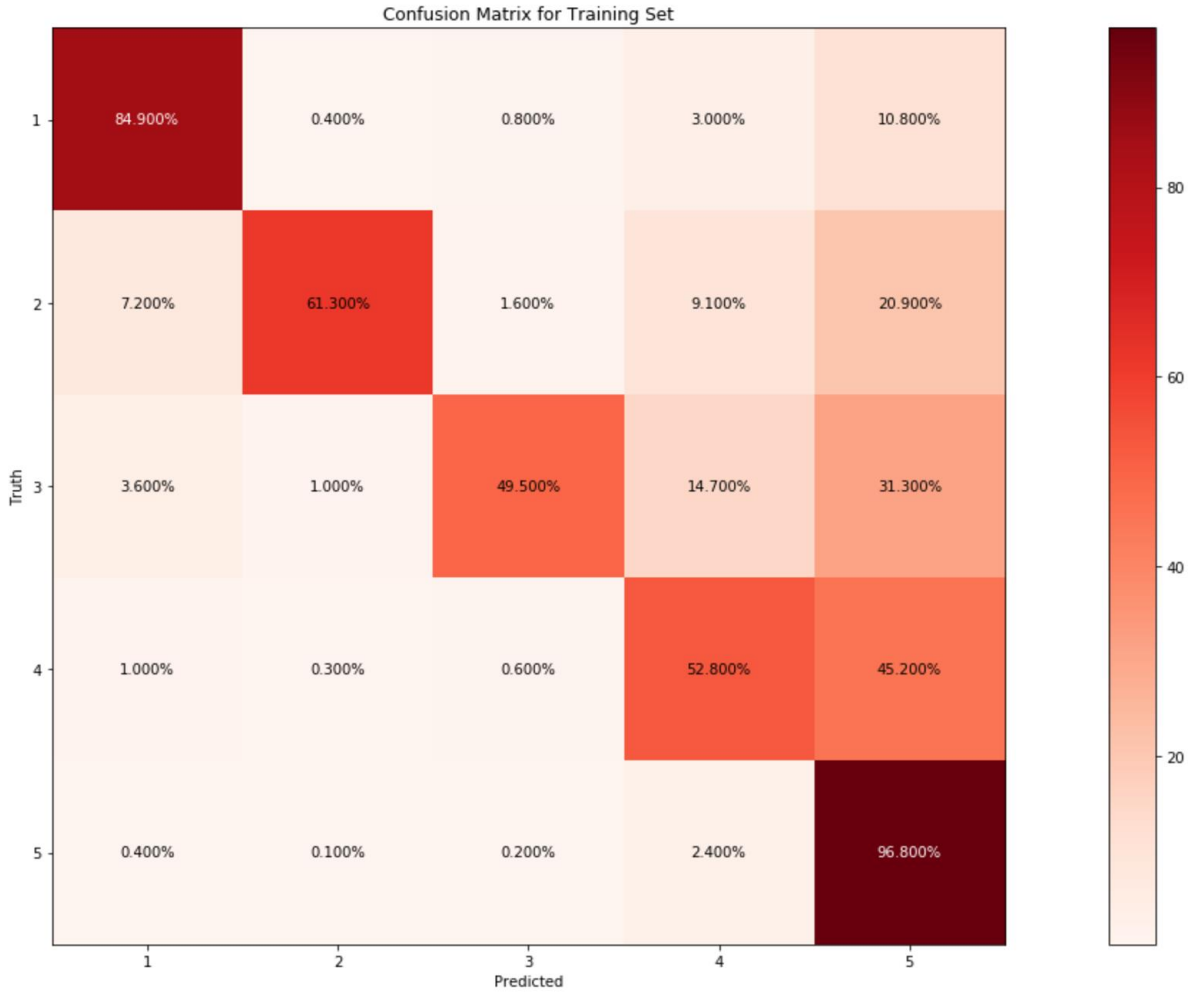
Machine Learning – XGBoost

- Boosting models outperformed Logistic Regression and Random Forest approaches using default parameters
- Tuned, multi-class XGBoost model was ultimately used for the study
- F1 scores were *micro-averaged* in the training set results which means the false positives, true positives, and false negatives were taken into account across all classes as opposed to averaging each class independently

Tuned XGBoost Scores:

```
Training Set Accuracy:    81.303%
 Training Set F1 Score:    0.813

Cross Validation Accuracy:        64.617%
```


Confusion Matrix for Training Set
84.900%
0.400%
0.800%
3.000%
10.800%
7.200%
61.300%
1.600%
9.100%
20.900%
3.600%
1.000%
49.500%
14.700%
31.300%
1.000%
0.300%
0.600%
52.800%
45.200%
0.400%
0.100%
0.200%
2.400%
96.800%
Truth
1
2
3
4
5
Predicted
1
2
3
4
5
80
60
40
20

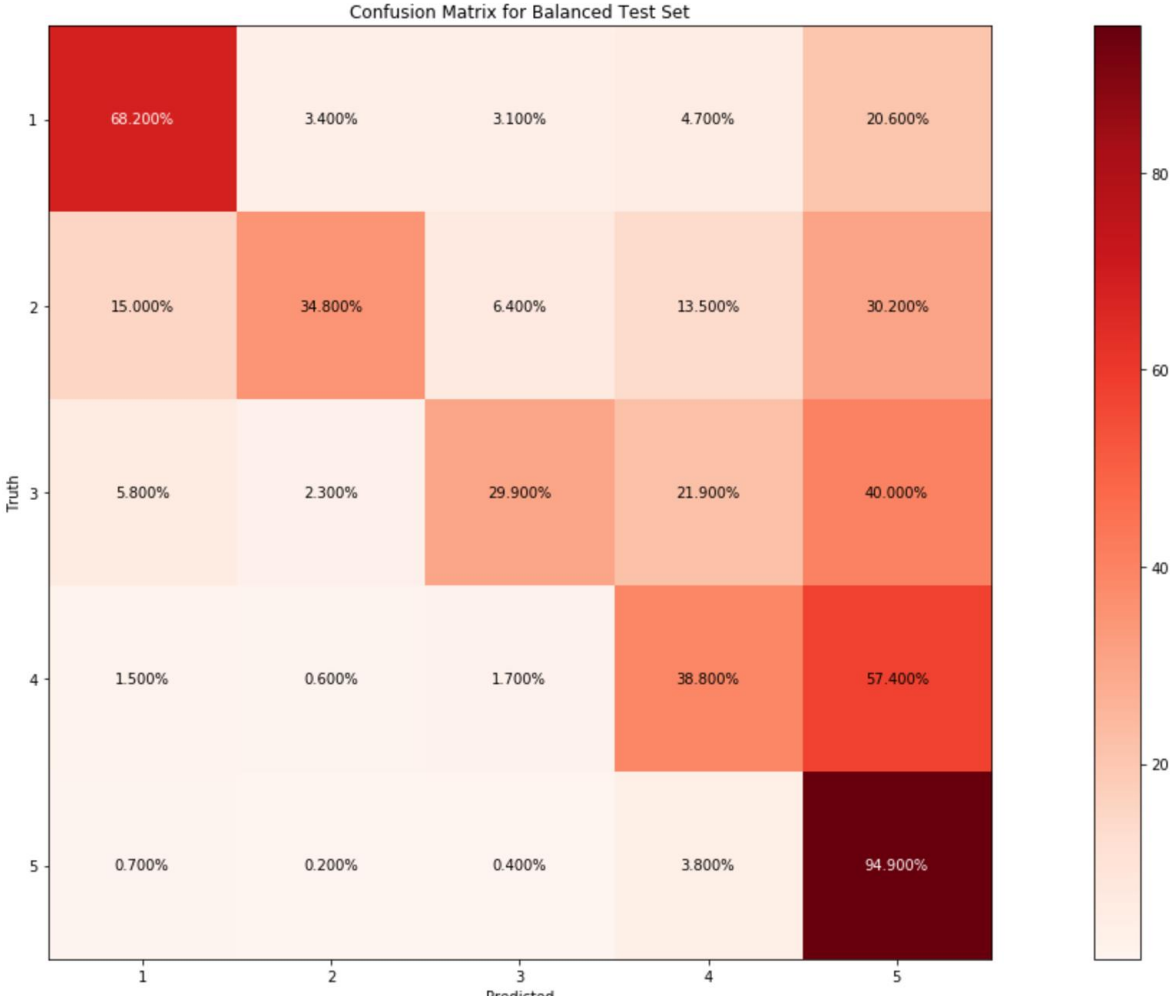
Machine Learning – Final Scores

- Applying the model to the reserved test set of our original `model_df`, we get comparably better results than the 20% accuracy of the baseline
- But the baseline was based on the balanced dataset and so it is appropriate that the score is taken on the trimmed dataset where under-representation was established:

Tuned XGBoost Scores:

```
Test Set Accuracy:        65.161%
 Test Set F1 Score:        0.652

Balanced Test Set Accuracy:        53.336%
 Balanced Test Set F1 Score:        0.533
```


Confusion Matrix for Balanced Test Set
Truth
1
2
3
4
5
68.200%
3.400%
3.100%
4.700%
20.600%
15.000%
34.800%
6.400%
13.500%
30.200%
5.800%
2.300%
29.900%
21.900%
40.000%
1.500%
0.600%
1.700%
38.800%
57.400%
0.700%
0.200%
0.400%
3.800%
94.900%
1
2
3
4
5
Predicted
80
60
40
20

Word Cloud

- Using the true labels of the reviews, the fifty most salient words in every rating can be taken to produce a word cloud
- Stop words derived from the NLTK library are excluded

WORD CLOUD FOR 1-STAR RATINGS

WORD CLOUD FOR 5-STAR RATINGS

Closing Thoughts – Conclusion

- Various NLP techniques and concepts were explored in the study
- Though word embedding was central to building the model, pre-processing steps were crucial
- The model actually extracts and quantifies *context* and therefore the essence of a review by its words make up the final dataframe
- The multi-class, discrete classifier approach makes our model reliant on the distinction of each star-rating – if a one-star review was misclassified as a five-star review, the model is agnostic to how far off 1 and 5 are
- It is more concerned in asking, "*What makes a 5-star review different from a 4-star review?*" than "*Is this review more approving than criticizing?*"

Closing Thoughts – Limits and Recommendations

- The model will not be able to handle words that it has not encountered during training, it will simply drop new, unrecognizable words
- No way of handling misspelled words – spell-check feature will only add to model complexity
- Misspelled words will be taken as they are during training
- As is usually the case in NLP, sarcasm or text that is intended to be ironic is interpreted by what is literally in the text and not by its underlying context