Springboard Data Science Course
Capstone Project 3

Natural Language Processing: Classify Amazon reviews based on the customer's ratings.

By: Joe Anson R. Aquino

August 11, 2023

Introduction

Many times, ratings are represented by a numerical value () or stars (★★★★★). However, the text feedback holds more value than the quantified ratings. Sometimes, the rating given may not accurately reflect the experience of the product. Given the text of the review of a product, we want to build a supervised, binary classifier model with the actual review text as the core predictor.

Approach

First, will gather data from Amazon Dataset contains the customer reviews for all listed Electronics, will do NLP Pre-Processing, Tokenization, Phrase Modeling, Vectorization before exploring the data before utilizing a machine-learning algorithm to predict if the review is negative or positive review. To justify this work, I conducted a comparative study and analysis using various classification algorithms, including Random Forest Classifier, XGBoost Classifier.

Here is the link for Amazon Dataset from the UCSD.EDU repository 142.8 million reviews (http://jmcauley.ucsd.edu/data/amazon)

Datasets

The **Amazon dataset** contains the customer reviews for all listed *Electronics* products spanning from May 1996 up to July 2014. There are a total of 1,689,188 reviews by a total of 192,403 customers on 63,001 unique products. The data dictionary is as follows:

- **asin** - Unique ID of the product being reviewed, *string*
- **helpful** - A list with two elements: the number of users that voted *helpful* , and the total number of users that voted on the review (including the *not helpful* votes), *list*
- **overall** - The reviewer's rating of the product, *int64*
- **reviewText** - The review text itself, *string*
- **reviewerID** - Unique ID of the reviewer, *string*
- **reviewerName** - Specified name of the reviewer, *string*
- **summary** - Headline summary of the review, *string*
- **unixReviewTime** - Unix Time of when the review was posted, *string*

Data Wrangling

The df is created from the Amazon dataset. If the file has been downloaded then the dataset is loaded from the local file. Otherwise the file is accessed and extracted directly from the repository.

	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
4	0528881469	[0, 0]	1	I've had mine for a year and here's what we go...	09 29, 2011	A24EV6RXELQZ63	Wayne Smith	Major issues, only excuses for support	1317254400
5	0594451647	[3, 3]	5	I am using this with a Nook HD+. It works as d...	01 3, 2014	A2JXAZZI9PHK9Z	Billy G. Noland "Bill Noland"	HDMI Nook adapter cable	1388707200
6	0594451647	[0, 0]	2	The cable is very wobbly and sometimes disconn...	04 27, 2014	A2P5U7BDKKT7FW	Christian	Cheap proprietary scam	1398556800
7	0594451647	[0, 0]	5	This adaptor is real easy to setup and use rig...	05 4, 2014	AAZ084UMH8VZ2	D. L. Brown "A Knower Of Good Things"	A Perfdect Nook HD+ hook up	1399161600
8	0594451647	[0, 0]	4	This adapter easily connects my Nook HD 7"...	07 11, 2014	AEZ3CR6BKIROJ	Mark Dietter	A nice easy to use accessory.	1405036800
9	0594451647	[3, 3]	5	This product really works great but I found th...	01 20, 2014	A3BY5KCNQZXV5U	Matenai	This works great but read the details...	1390176000

Only the `overall` and the `unixReviewTime` series are stored as integers. The rest are interpreted as strings (objects).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1689188 entries, 0 to 1689187
Data columns (total 9 columns):
asin              1689188 non-null object
helpful           1689188 non-null object
overall           1689188 non-null int64
reviewText        1689188 non-null object
reviewTime        1689188 non-null object
reviewerID        1689188 non-null object
reviewerName      1664458 non-null object
summary           1689188 non-null object
unixReviewTime    1689188 non-null int64
dtypes: int64(2), object(7)
memory usage: 116.0+ MB
```

The `unixReviewTime` is converted from Unix time to the more intuitive `datetime` datatype then `reviewTime` is dropped since the `unixReviewTime` series more accurately describes the time when each review was posted.

	asin	helpful	overall	reviewText	reviewerID	reviewerName	summary	unixReviewTime
0	0528881469	[0, 0]	5	We got this GPS for my husband who is an (OTR)...	AO94DHGC771SJ	amazdnu	Gotta have GPS!	06-01-2013
1	0528881469	[12, 15]	1	I'm a professional OTR truck driver, and I bou...	AMO214LNFCEI4	Amazon Customer	Very Disappointed	11-24-2010
2	0528881469	[43, 45]	3	Well, what can I say. I've had this unit in m...	A3N7T0DY83Y4IG	C. A. Freeman	1st impression	09-08-2010
3	0528881469	[9, 10]	2	Not going to write a long review, even thought...	A1H8PY3QHMQQA0	Dave M. Shaw "mack dave"	Great grafics, POOR GPS	11-23-2010
4	0528881469	[0, 0]	1	I've had mine for a year and here's what we go...	A24EV6RXELQZ63	Wayne Smith	Major issues, only excuses for support	09-28-2011

Each review is stored as string in the reviewText series.

```
I&#8217;m a big fan of the Brainwavz S1 (actually all of their headphones &#8211; have yet to be disappoin
ted with any of their products). The S1 has been my main set for active use (e.g., workouts, runs, etc.) s
ince the flat cable is very durable and resistant to tangles. The S5 keeps all the good features of the S1
and adds to it &#8211; the sound quality is richer and better defined.That&#8217;s not to say the S1 sound
s poor &#8211; they are quite good, in fact. But the S5 are better. The highs are better defined and the m
idrange 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 see
ms solid &#8211; as solid as the S1 or better. I love the flat cable! I know that&#8217;s something that i
s not appreciated by everyone, but for me it&#8217;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 aroun
d my mp3 player in my pocket. Easy to carry; very stressful on the cables and can lead to tangles with rou
nd wires. Flat wires, especially those with a thick jacket such as these, survive that abuse with zero pro
blems.The earbuds themselves are more sleekly shaped than the &#8220;can&#8221; style of the S1.Comfort is
in line with the customary Brainwavz style, which is to say it&#8217;s outstanding. It comes with a wide r
ange of tips to fit pretty much any ear, plus the Comply foam tips (which are my favorite). If fitted prop
erly you end up with zero ear irritation plus excellent sound isolation and bass response.These are an ove
r-the-ear design much like the S1. I never used that design prior to the S1 and it did take me a little ti
me to get accustomed to it. It became second nature quickly, and the design is a lot more stable when exer
cising than the conventional in-ear design.These are more expensive than the S1, but you can hear the diff
erence in price. Still, if you&#8217;re looking to keep the cost down a bit, the S1 is an excellent perfor
mer 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]
```

Each review is associated with a rating stored under the overall field. This serves as the quantified summary of a given review and will thus be used as the ground truth labels for the model.

NLP Pre-Processing

We'll work with reviewText to prepare our model's final dataframe. The goal is to produce tokens for every document (i.e. every review). These documents will make up our corpora where we'll draw our vocabulary from.

The following is a sample text in its original form. This is the same as what was inspected in the previous section.


```
I&#8217;m a big fan of the Brainwavz S1 (actually all of their headphones &#8211; have yet to be disappoin
ted with any of their products). The S1 has been my main set for active use (e.g., workouts, runs, etc.) s
ince the flat cable is very durable and resistant to tangles. The S5 keeps all the good features of the S1
and adds to it &#8211; the sound quality is richer and better defined.That&#8217;s not to say the S1 sound
s poor &#8211; they are quite good, in fact. But the S5 are better. The highs are better defined and the m
idrange 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 see
ms solid &#8211; as solid as the S1 or better. I love the flat cable! I know that&#8217;s something that i
s not appreciated by everyone, but for me it&#8217;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 aroun
d my mp3 player in my pocket. Easy to carry; very stressful on the cables and can lead to tangles with rou
nd wires. Flat wires, especially those with a thick jacket such as these, survive that abuse with zero pro
blems.The earbuds themselves are more sleekly shaped than the &#8220;can&#8221; style of the S1.Comfort is
in line with the customary Brainwavz style, which is to say it&#8217;s outstanding. It comes with a wide r
ange of tips to fit pretty much any ear, plus the Comply foam tips (which are my favorite). If fitted prop
erly you end up with zero ear irritation plus excellent sound isolation and bass response.These are an ove
r-the-ear design much like the S1. I never used that design prior to the S1 and it did take me a little ti
me to get accustomed to it. It became second nature quickly, and the design is a lot more stable when exer
cising than the conventional in-ear design.These are more expensive than the S1, but you can hear the diff
erence in price. Still, if you&#8217;re looking to keep the cost down a bit, the S1 is an excellent perfor
mer 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]
```

HTML Entities

Some special characters like the apostrophe (’) and the en dash (–) are expressed as a set of numbers prefixed by `&#` and suffixed by `;` . This is because the dataset was scraped from an HTML parser, and the dataset itself includes data that predated the universal UTF-8 standard.

These *HTML Entities* can be decoded by importing the `html` library.

```
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 i
t – the sound quality is richer and better defined.That’s not to say the S1 sounds poor – they are quite g
ood, in fact. But the S5 are better. The highs are better defined and the midrange has more punch to it. T
he bass comes through clearly without moving into the harsh territory when the volume is pushed (as the S1
s can do). The overall sound quality is very pleasing.The build quality seems solid – as solid as the S1 o
r better. I love the flat cable! I know that’s something that is not appreciated by everyone, but for me i
t’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 thic
k jacket such as these, survive that abuse with zero problems.The earbuds themselves are more sleekly shap
ed than the “can” style of the S1.Comfort is in line with the customary Brainwavz style, which is to say i
t’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 isola
tion and bass response.These are an over-the-ear design much like the S1. I never used that design prior t
o the S1 and it did take me a little time to get accustomed to it. It became second nature quickly, and th
e 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 bi
t, the S1 is an excellent performer as well.Great sound, great comfort, wonderful cable design, and it com
es with a solidly made case and lots of eartips. Highly recommend.[Sample provided for review]
```

Since punctuation marks do not add value in the way we'll perform NLP, all the HTML entities in the review texts can be dropped. The output series `preprocessed` is our `reviewText` but without the special characters.

```
Im a big fan of the Brainwavz S1 (actually all of their headphones  have yet to be disappointed with any o
f 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 i
t  the sound quality is richer and better defined.Thats not to say the S1 sounds poor  they are quite goo
d, 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 c
an do). The overall sound quality is very pleasing.The build quality seems solid  as solid as the S1 or be
tter. I love the flat cable! I know thats something that is not appreciated by everyone, but for me its be
en 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 stres
sful on the cables and can lead to tangles with round wires. Flat wires, especially those with a thick jac
ket such as these, survive that abuse with zero problems.The earbuds themselves are more sleekly shaped th
an the can style of the S1.Comfort is in line with the customary Brainwavz style, which is to say its outs
tanding. It comes with a wide range of tips to fit pretty much any ear, plus the Comply foam tips (which a
re 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 youre 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]
```

The lemmatize_doc works as follows:

- Each review is broken down into a list of sentences
- Punctuations that only group words or separate sentences (hyphens therefore are excluded) are removed (replaced by whitespace) using RegEx

- Every sentence is further broken down into words (tokens)

Each of the sentences then becomes an ordered bag of words. Every word is then *tagged* to a part-of-speech. This word-tag tuple pair is then fed one at a time to the lemmatize_word function, which works as follows:

- Only modifiable words – nouns, verbs, adjectives, and adverbs – can be reduced to roots
- These words are lemmatized and appended to the root list
- Words that are not modifiable are added as they are to the root list

The output lists are linked together as a string using whitespace. In the end, each preprocessed review will retain its text form but with each word simplified as much as possible.

```
Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of the
ir product The S1 have be my main set for active use e g workouts run etc since the flat cable be very dur
able 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 n
ot appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz hea
dset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pock
et 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 mor
e sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which b
e 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 isolatio
n 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 bu
t you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excell
ent performer as well Great sound great comfort wonderful cable design and it come with a solidly make cas
e and lot of eartips Highly recommend [ Sample provide for review ]
```

Removing Accents

Each review is normalized from longform UTF-8 to ASCII encoding. This will remove accents in characters and ensure that words like "*naïve*" will simply be interpreted as (and therefore not differentiated from) "*naive*".

Removing Punctuations

The `preprocessed` reviews are further cleaned by dropping punctuations. Using regular expressions, only whitespaces and alphanumeric characters are kept.

```
Im a big fan of the Brainwavz S1 actually all of their headphone have yet to be disappoint with any of the
ir product The S1 have be my main set for active use e g workouts run etc since the flat cable be very dur
able 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 n
ot appreciate by everyone but for me its be work out wonderfully Although this as most other Brainwavz hea
dset come with an excellent hard shell case I usually tote my earbuds wrap around my mp3 player in my pock
et 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 mor
e sleekly shape than the can style of the S1 Comfort be in line with the customary Brainwavz style which b
e 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 isolatio
n 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 bu
t you can hear the difference in price Still if youre look to keep the cost down a bit the S1 be an excell
ent performer as well Great sound great comfort wonderful cable design and it come with a solidly make cas
e and lot of eartips Highly recommend   Sample provide for review
```

Converting to Lowercase

Every letter is also 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 the
ir product the s1 have be my main set for active use e g workouts run etc since the flat cable be very dur
able 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 n
ot appreciate by everyone but for me its be work out wonderfully although this as most other brainwavz hea
dset come with an excellent hard shell case i usually tote my earbuds wrap around my mp3 player in my pock
et 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 mor
e sleekly shape than the can style of the s1 comfort be in line with the customary brainwavz style which b
e 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 isolatio
n 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 bu
t you can hear the difference in price still if youre look to keep the cost down a bit the s1 be an excell
ent performer as well great sound great comfort wonderful cable design and it come with a solidly make cas
e and lot of eartips highly recommend   sample provide for review
```

Removing Stop Words

Stop words consist of the most commonly used words that include pronouns (e.g. *us*, *she*, *their*), articles (e.g. *the*), and prepositions (e.g. *under*, *from*, *off*). These words are not helpful in distinguishing a document from

another and are therefore dropped.

Note that the `stop_words` were stripped of punctuations just as what we have done to our dataset.

```
sample stop words: ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'youre', 'youve',
'youll', 'youd', 'your', 'yours']
```

```
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 t
hats say s1 sound poor quite good fact s5 well high well define midrange punch bass come clearly without m
ove harsh territory volume push s1s overall sound quality please build quality seem solid solid s1 good lo
ve flat cable know thats something appreciate everyone work wonderfully although brainwavz headset come ex
cellent hard shell case usually tote earbuds wrap around mp3 player pocket easy carry   stressful cable le
ad tangle round wire flat wire especially thick jacket survive abuse zero problem earbuds sleekly shape st
yle 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 quick
ly design lot stable exercise conventional ear design expensive s1 hear difference price still look keep c
ost bit s1 excellent performer well great sound great comfort wonderful cable design come solidly make cas
e lot eartips highly recommend   sample provide review
```

Removing Extra Spaces

Again, we make use of regular expressions to ensure we never get more than a single whitespace to separate words in our 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 t
hats say s1 sound poor quite good fact s5 well high well define midrange punch bass come clearly without m
ove harsh territory volume push s1s overall sound quality please build quality seem solid solid s1 good lo
ve flat cable know thats something appreciate everyone work wonderfully although brainwavz headset come ex
cellent 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 com
ply foam tip favorite fit properly end zero ear irritation plus excellent sound isolation bass response ea
r 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 lo
t eartips highly recommend sample provide review
```

Tokenization

The entries for the `processed col`umn are extracted to make up our *corpora*, which is simply a collection of all our documents. Each review is then transformed into an ordered list of words. This is the process of *tokenization* – the document is broken down into individual words or tokens.

```
['im', 'big', 'fan', 'brainwavz', 's1', 'actually', 'headphone', 'yet', 'disappoint', 'product', 's1', 'ma
in', 'set', 'active', 'use', 'e', 'g', 'workouts', 'run', 'etc', 'since', 'flat', 'cable', 'durable', 'res
istant', 'tangle', 's5', 'keep', 'good', 'feature', 's1', 'add', 'sound', 'quality', 'rich', 'well', 'defi
ne', 'thats', 'say', 's1', 'sound', 'poor', 'quite', 'good', 'fact', 's5', 'well', 'high', 'well', 'defin
e', 'midrange', 'punch', 'bass', 'come', 'clearly', 'without', 'move', 'harsh', 'territory', 'volume', 'pu
sh', 's1s', 'overall', 'sound', 'quality', 'please', 'build', 'quality', 'seem', 'solid', 'solid', 's1',
'good', 'love', 'flat', 'cable', 'know', 'thats', 'something', 'appreciate', 'everyone', 'work', 'wonderfu
lly', 'although', 'brainwavz', 'headset', 'come', 'excellent', 'hard', 'shell', 'case', 'usually', 'tote',
'earbuds', 'wrap', 'around', 'mp3', 'player', 'pocket', 'easy', 'carry', 'stressful', 'cable', 'lead', 'ta
ngle', 'round', 'wire', 'flat', 'wire', 'especially', 'thick', 'jacket', 'survive', 'abuse', 'zero', 'prob
lem', 'earbuds', 'sleekly', 'shape', 'style', 's1', 'comfort', 'line', 'customary', 'brainwavz', 'style',
'say', 'outstanding', 'come', 'wide', 'range', 'tip', 'fit', 'pretty', 'much', 'ear', 'plus', 'comply', 'f
oam', 'tip', 'favorite', 'fit', 'properly', 'end', 'zero', 'ear', 'irritation', 'plus', 'excellent', 'soun
d', 'isolation', 'bass', 'response', 'ear', 'design', 'much', 'like', 's1', 'never', 'use', 'design', 'pri
or', 's1', 'take', 'little', 'time', 'get', 'accustom', 'become', 'second', 'nature', 'quickly', 'design',
'lot', 'stable', 'exercise', 'conventional', 'ear', 'design', 'expensive', 's1', 'hear', 'difference', 'pr
ice', 'still', 'look', 'keep', 'cost', 'bit', 's1', 'excellent', 'performer', 'well', 'great', 'sound', 'g
reat', 'comfort', 'wonderful', 'cable', 'design', 'come', 'solidly', 'make', 'case', 'lot', 'eartips', 'hi
ghly', 'recommend', 'sample', 'provide', 'review', '']
```

Phrase Modeling

Bigrams

Bigrams are generated from using the *gensim* phraser. Only those that pass the `bi_gram` criteria are considered. Sample bigrams below:

```
['2_00', 'make_convenient', 'matter_fact', 'actually_see', 'sure_problem', 'problem_design', 'work_everyth
ing', 'standard_camera', '1080p_120hz', 'set_ipad', 'control_cable', 'nikon_brand', 'tiny_size', 'tiny_cam
era', 'use_default', 'plug_network', 'light_fit', 'button_click', '4kb_qd', 'wheel_click', 'hold_device',
'ipod_phone', 'might_break', 'big_small', 'noise_ratio', 'less_200', 'design_camera', 'camera_function']
```

Trigrams

Trigrams are generated by applying another *gensim* phraser on top of a bigram phraser. Take for example the tokens *sd* and *card* . Because they appear often together enough, they become linked together as *sd_card*. In them together as well to tokenize turn, if sd_card appears adjacent to the token reader in enough instances, then the tri_gram model link *sd_card_reader* .

```
['play_blu_ray', 'samsung_galaxy_s4', 'old_macbook_pro', 'quality_top_notch', 'b_w_filter', 'one_living_ro
om', '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_re
p', '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_u
se', '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']
```

Single-character tokens are removed from every tokenized document. Our tokenized review, below is our final form

```
['im', 'big', 'fan', 'brainwavz', 's1', 'actually', 'headphone', 'yet', 'disappoint', 'product', 's1', 'ma
in', 'set', 'active', 'use', 'workouts', 'run', 'etc', 'since', 'flat', 'cable', 'durable', 'resistant',
'tangle', 's5', 'keep', 'good', 'feature', 's1', 'add', 'sound', 'quality', 'rich', 'well', 'define', 'tha
ts', 'say', 's1', 'sound', 'poor', 'quite', 'good', 'fact', 's5', 'well', 'high', 'well', 'define', 'midra
nge', 'punch', 'bass', 'come', 'clearly', 'without', 'move', 'harsh', 'territory', 'volume', 'push', 's1
s', 'overall', 'sound', 'quality', 'please', 'build', 'quality', 'seem', 'solid', 'solid', 's1', 'good',
'love', 'flat', 'cable', 'know', 'thats', 'something', 'appreciate', 'everyone', 'work', 'wonderfully', 'a
lthough', 'brainwavz', 'headset', 'come', 'excellent', 'hard', 'shell', 'case', 'usually', 'tote', 'earbud
s', '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', 'i
solation', 'bass', 'response', 'ear', 'design', 'much', 'like', 's1', 'never', 'use', 'design', 'prior',
's1', 'take', 'little', 'time', 'get', 'accustom', 'become', 'second', 'nature', 'quickly', 'design', 'lo
t', 'stable', 'exercise', 'conventional', 'ear', 'design', 'expensive', 's1', 'hear', 'difference', 'pric
e', 'still', 'look', 'keep', 'cost', 'bit', 's1', 'excellent', 'performer', 'well', 'great', 'sound', 'gre
at', 'comfort', 'wonderful', 'cable', 'design', 'come', 'solidly', 'make', 'case', 'lot', 'eartips', 'high
ly', 'recommend', 'sample', 'provide', 'review']
```

Creating the Vocabulary

```
ID: 0, Token: address
ID: 1, Token: around
ID: 2, Token: arrive
ID: 3, Token: back
ID: 4, Token: bad
ID: 5, Token: big
ID: 6, Token: come
ID: 7, Token: contact
ID: 8, Token: could
ID: 9, Token: day
```

Bag of Words Model

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

TF-IDF Model

```
Word: address, Weight: 0.113
Word: around, Weight: 0.060
Word: arrive, Weight: 0.093
Word: back, Weight: 0.051
Word: bad, Weight: 0.068
Word: big, Weight: 0.126
Word: come, Weight: 0.046
Word: contact, Weight: 0.103
Word: could, Weight: 0.054
Word: day, Weight: 0.061
Word: earlier, Weight: 0.141
Word: ease, Weight: 0.220
Word: ect, Weight: 0.181
Word: email, Weight: 0.213
Word: exception, Weight: 0.131
Word: exchange, Weight: 0.132
Word: expect, Weight: 0.067
Word: freeze, Weight: 0.259
```

Word Embedding for Feature Engineering

The downside of count-based techniques is that without regard to word sequence and sentence structure, the semantics get lost. 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.

Final Dataframe

The goal is to have a dataframe with observations corresponding to the product reviews. The `word_vec` model is used to gather all the unique tokens in the corpora. This enables us to generate the `word_vec_df` which makes use of the dimensions as the features of every word.

	0	1	2	3	4	5	...	95	96	97	98	99	96	97	98	99
get	2.027105	2.285539	0.325559	5.013640	-0.886071	-2.239007	...	-4.002328	5.016023	1.077161	-0.641248	1.685374	5.016023	1.077161	-0.641248	1.685374
gps	4.430076	-1.912336	9.017261	3.536343	-9.298578	-3.119642	...	-0.979287	6.612934	-4.329911	5.106474	7.303112	6.612934	-4.329911	5.106474	7.303112
husband	1.160196	4.993967	1.216444	2.476656	-3.268333	1.352225	...	-2.101081	5.262798	0.258067	-0.073419	-2.713029	5.262798	0.258067	-0.073419	-2.713029
otr	1.308931	-1.700580	2.168358	2.334967	0.032168	0.996840	...	0.947201	-0.885584	-0.979156	0.709682	1.676071	-0.885584	-0.979156	0.709682	1.676071
road	3.249146	2.961001	10.251733	3.529565	-7.407037	-1.892380	...	0.373690	10.218236	-3.727395	-3.895083	6.060423	10.218236	-3.727395	-3.895083	6.060423

5 rows × 100 columns

The word_vec_df is sliced by the words that appear in a given tokenized d review and the mean along every dimension is taken. The resulting model_array shape is therefore the word count on *axis 0* and the number of dimensions on *axis 1* . This singularizes multiple word embeddings into one observation for each review.

If multiple occurrences of a word occurs in a review, then this only emphasizes the token since the row is pulled towards the values of the vectors of that word.

```
tokenized_array = np.array(tokenized)

model_array = np.array([word_vec_df.loc[doc].mean(axis=0) for doc in tokenized_array])
```

Every document is provided the ground truth label by imposing its `overall` rating. This completes our finalized `model_df` dataframe.

	0	1	2	3	...	96	97	98	99	label
0	0.456036	-0.340525	-0.423433	1.381710	...	1.881948	-1.110850	-0.336875	0.206793	5
1	-0.127207	0.409159	0.727547	0.947930	...	1.551120	-0.726942	-0.235786	1.039625	1
2	-0.782381	-0.635467	0.310187	1.470468	...	1.439608	-1.896147	-0.506262	0.904398	3
3	0.079711	-0.142391	0.528292	1.845955	...	1.987752	-1.498181	-0.840509	0.573334	2
4	0.446526	-0.167763	0.217617	0.934713	...	2.020145	-0.689082	0.123893	1.976123	1

5 rows × 101 columns

Exploratory Data Analysis

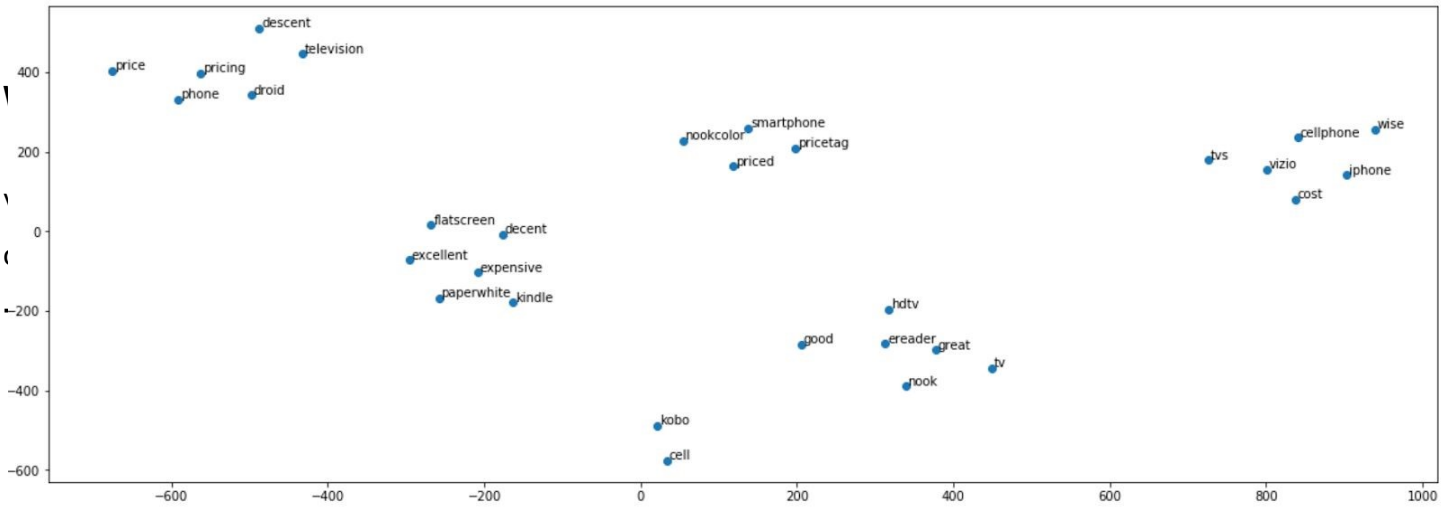
We'll implement several interesting Natural Language Processing techniques in order to explore our Amazon

dataset.

More on Word2Vec

```
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']
```

t-SNE



Named Entity Recognition

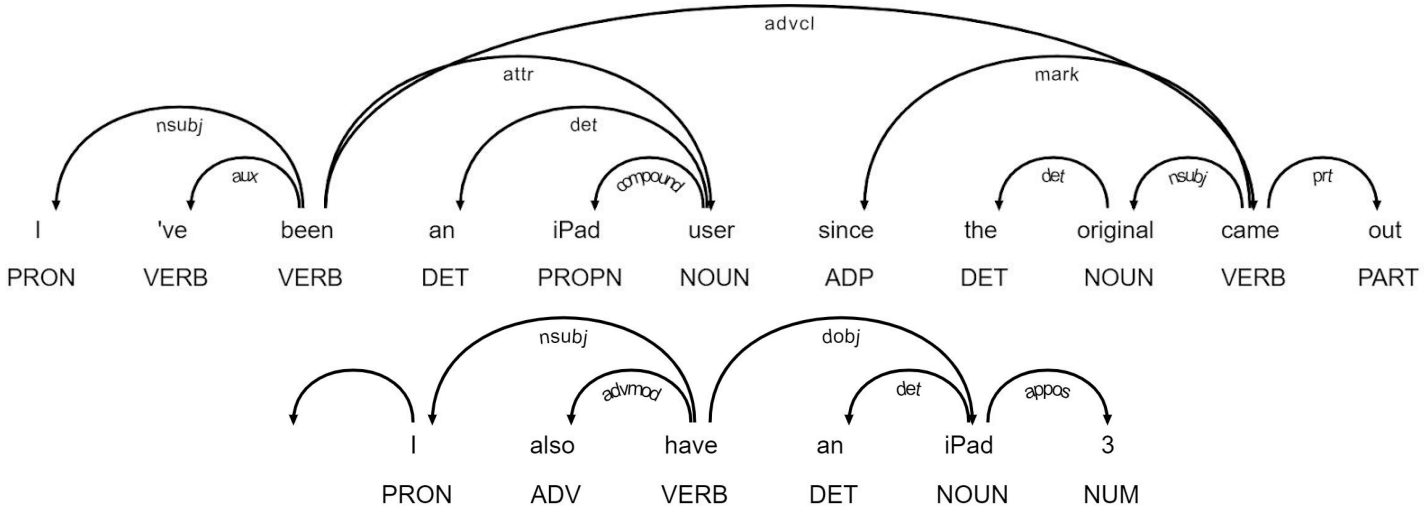
We've seen *gensim* perform word tagging to identify part-of-speech. Now we use *spaCy* to go further and identify what nouns in the documents refer to. Some Named-Entity Recognition (*NER*) classification tags include distinguishing persons, organizations, products, places, dates, etc.

```
I've been an iPad user since the original came out. I also have an iPad 3. I have worked in IT for the pas
t few years 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 goe
s...BUILDThe device feels nice and solid. I'm a little surprised at how heavy it is, but that's not necess
arily a bad thing. The rubberized backing is always nice for added grip. It's not as nice as say unibody a
luminum, but it's not $500 either.SCREENThe screen is fantastic. But my problem is the same as when iPad g
ot Retina Display, other than the OS, 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 hig
her resolutions.LOCK SCREEN ADSYeah there are ads on my lock screen. I'm not sure why this is such a big d
eal. How much time do people really spend looking at the lock screen? The first 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 price tag. I
might pay the $15 to get rid of them so I can customize it, but I might not. I feel like this has been blo
wn out of proportion by other customers.SOUNDThe sound from the speakers is great. Much better than you wo
```

Dependency Tree

The capability of *spaCy*'s NER is based on deciphering the structure of the sentence by breaking down how tokens interact with and influence each other. Below is the dependency trees of the first two sentences

Dependency Tree



Topic Modeling

```
get 0.011734774
work 0.011198829
use 0.0095958505
router 0.009158912
device 0.008855804
```

The words that are the most characteristic of the topics are indeed thematic. And each word group do conjure a distinct topic.

```
Topic 1:
 get, 0.011734774336218834
 work, 0.011198828928172588
 use, 0.009595850482583046
 router, 0.009158912114799023
 device, 0.008855803869664669

Topic 2:
 sound, 0.034996841102838516
 speaker, 0.0209796279668808
 good, 0.015011309646070004
 headphone, 0.013791842386126518
 quality, 0.011160140857100487

Topic 3:
 lens, 0.03160819783806801
 bag, 0.018301470205187798
 camera, 0.014126963913440704
 use, 0.01146912481635809
 strap, 0.008649271912872791

Topic 4:
 cable, 0.031109070405364037
 work, 0.02897009812295437
 usb, 0.025115681812167168
 drive, 0.023603621870279312
 great, 0.017832685261964798

Topic 5:
 case, 0.03297600895166397
 cover, 0.012749520130455494
 screen, 0.011409485712647438
 like, 0.011152341961860657
 ipad, 0.010687867179512978

Topic 6:
 battery, 0.02778122015297413
 charge, 0.025451648980379105
 use, 0.017372693866491318
 phone, 0.016959380358457565
 one, 0.01309022307395935

Topic 7:
 camera, 0.0475146621465683
 use, 0.01570322923362255
 picture, 0.012167769484221935
 take, 0.011743015609681606
 video, 0.011700318194925785

Topic 8:
 card, 0.03473054617643356
 drive, 0.01358714234083891
 fan, 0.010703792795538902
 run, 0.0096812155097723
 memory, 0.009252899326384068

Topic 9:
 use, 0.023249300196766853
 keyboard, 0.02211669646203518
 tablet, 0.015327784232795238
 mouse, 0.011338042095303535
 like, 0.01100770290941

Topic 10:
 tv, 0.029768439009785652
 remote, 0.011976256966590881
 use, 0.010298002511262894
 get, 0.009141155518591404
 watch, 0.008147124201059341
```

Machine Learning

We'll further process our finalized model_df dataframe in order to make it compatible and easy to pipe into our Machine Learning model.

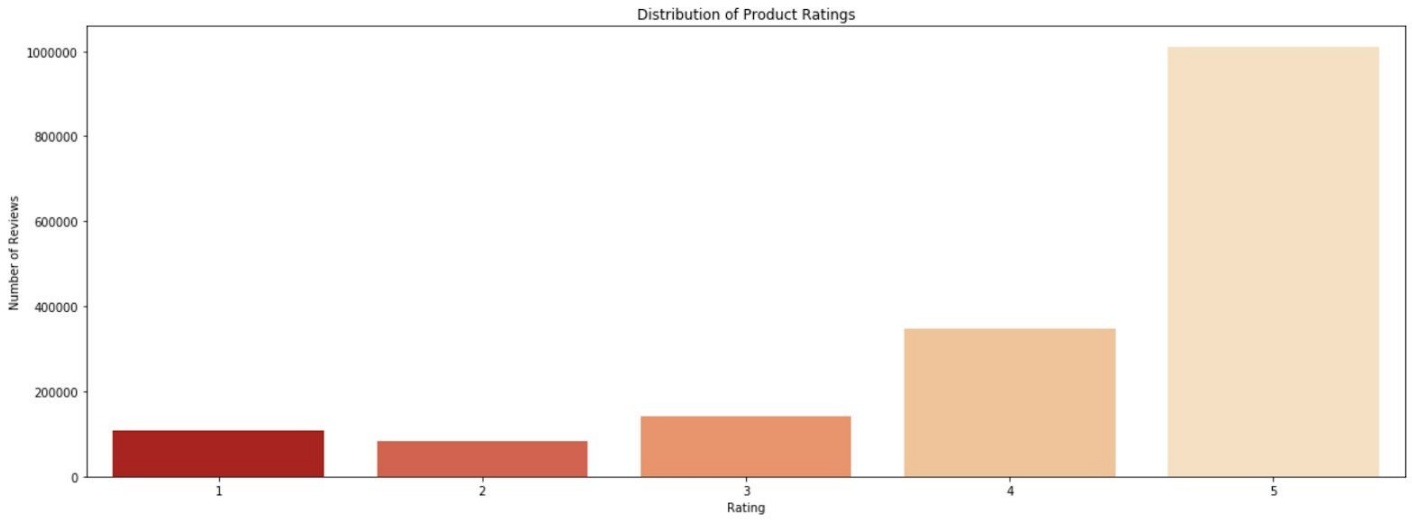
```
Original 'model_df' count: 1321
Final 'model_df' count: 1321
```

	0	1	2	3	4	5	6	7	8	9	...	91	92	93
0	-0.042715	0.449554	-0.058469	-0.210948	0.221794	-0.406341	0.152388	0.559175	-0.487808	-0.415364	...	0.166407	-0.031325	-0.102957
1	-0.058009	0.447886	-0.060280	-0.213946	0.239694	-0.400263	0.140340	0.551776	-0.484460	-0.416487	...	0.165706	-0.031013	-0.115166
2	-0.042507	0.462787	-0.075884	-0.228007	0.244245	-0.383637	0.138473	0.547498	-0.502998	-0.400627	...	0.175365	-0.023077	-0.118382
3	-0.051971	0.446820	-0.068522	-0.217371	0.231471	-0.385490	0.139727	0.540269	-0.464324	-0.403209	...	0.161128	-0.030951	-0.107705
4	-0.046612	0.502452	-0.093345	-0.246627	0.276240	-0.374039	0.146505	0.557952	-0.554885	-0.401337	...	0.166933	-0.031728	-0.136688

5 rows × 101 columns

Dealing with Unbalanced Data

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



If the model simply classified every review as `5` , then an accuracy of around 60% can be achieved given this exact dataset. Since this would outperform predictions made by chance, we should therefore ensure that we stratify the testing set where we base the final score of the model.

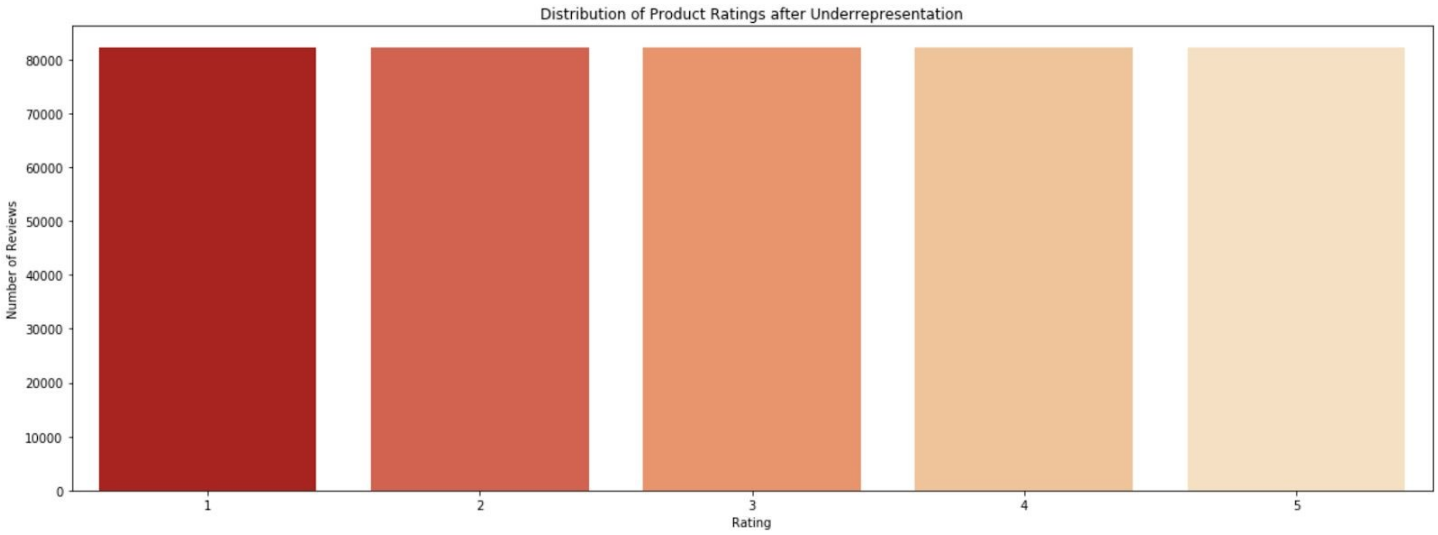
59.73%

Underrepresentation vs. Overrepresentation

In choosing this route to deal with imbalance, we create a trimmed version of our dataframe, `trimmed_df` . Each class is trimmed to have the same number of entries as the smallest class which is *Class 2*.

```
Size of Class 1: 82139
Size of Class 2: 82139
Size of Class 3: 82139
Size of Class 4: 82139
Size of Class 5: 82139
```

We see that we now have a perfectly balanced dataset after we performed underrepresentation.



Train-Test Split

The y is our target variable or the labels for the data. The X constitutes the features and are the predictor variables.

We evenly split the training and testing sets and *stratify* to ensure the ratio of classes in both sets are identical.

Scoring and Baseline

In our study, we will make use of two metrics to measure the model performance:

- Accuracy
- F1 Score

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 and so the probability that a correct prediction is made from pure guesswork is 20 %.

The F1 score is taking *precision* and *recall* into consideration. Taking into account false positives and false negatives for each class is especially important in inherently imbalanced datasets.

The baseline scores below are for when a model only randomly guesses the output labels – in this case, when every prediction is the same class. The scores are also based on an evenly distributed dataset.

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

Random Forest

Random Forest actually has a native way of supporting datasets that have class imbalance. We will therefore be able to use the original `model_df` instead of the sample `trimmed_df` :

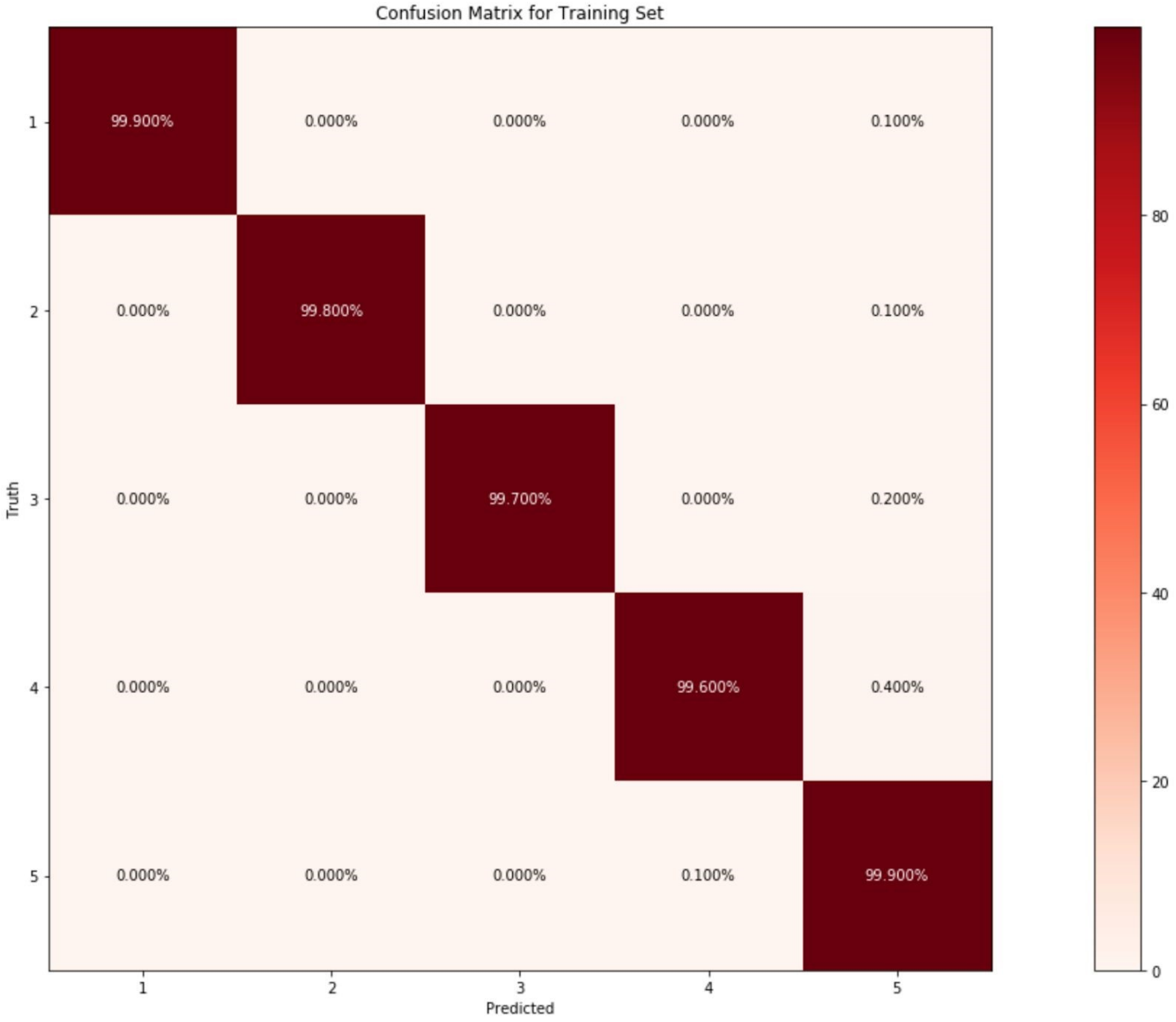
The `class_weight` attribute is provided with a dictionary that represents the associated weight of each class – the majority class is given a *1* and the rest are given the multiplying factor at which they would level with the largest class.

The criteria chosen is `entropy` which is similar to `gini` but instead of splitting nodes until there are pure classes, the nodes are split until the classes within have equal probability.

Our tuned Random Forest model got a very high score on the training data. The confusion matrix plotted below highlighted how the model almost perfectly classified each Amazon review accordingly.

However, these scores may be misleading since they are based on the data that the model were trained on. This is highly likely a result of *overfitting*. It is then important to rate our model more effectively without digging into our reserved test set.

```
Training Set Accuracy: 99.839%
Training Set F1 Score: 0.998
```



Cross-Validation

Cross-validation makes the most of the training data by splitting the training set into *folds* and further subjecting each fold to train-test splits. Cross-validation can thus test against overfitting and the resulting scores can better reflect how the model performs on data it has not seen before.

```
Training Set Accuracy: 61.722%
Training Set F1 Score: 0.617
```

XGBoost

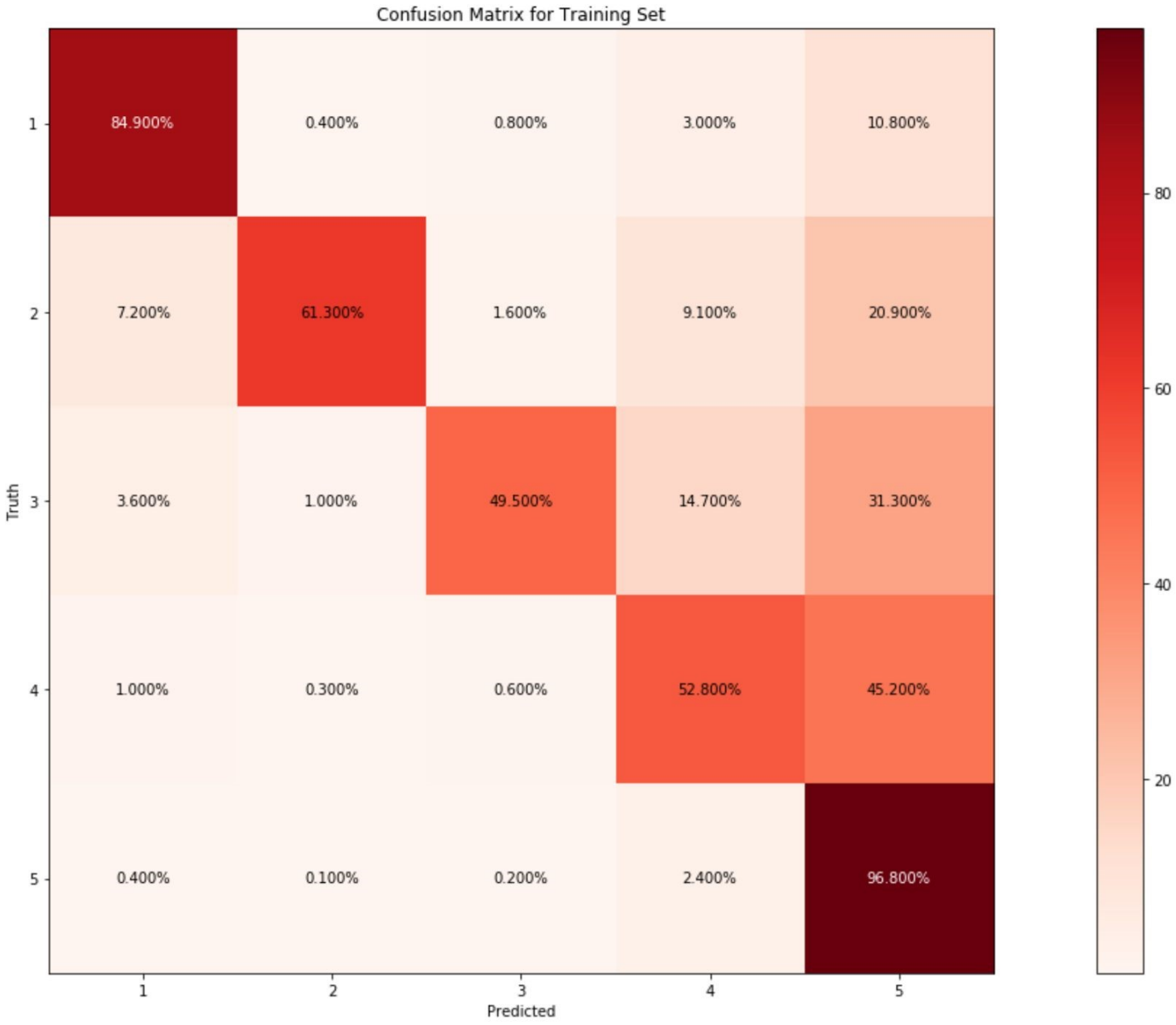
Let's now try to create a model based on a popular boosting technique and see how it compares with our Random Forest model (which is a tree-based bagging approach). XGBoost has become a staple in Kaggle competitions because of its high rate of success and its ease-of-use.

The class notation for our *XGBoost* object `boost` begins from 0, and so we perform an element-wise shift of our labels from *1* to *0*, from *2* to *1*, from *3* to *2*, etc. We tune our model using the maximum number of depths, the learning rate (*eta*), the number of classes, etc. We expect our outputs to be multi-class and so we select `softprob` as our *objective*.

The array of predicted labels `y_pred` contains lists of probabilities for each class per product review. The class that is deemed most likely is chosen by the *argmax* and the labels are shifted back to their original state.

The `micro` approach in averaging the F1 score means that the false positives, true positives, and false negatives are taken into account across all classes. This is in contrast with the `macro` approach that instead averages the F1 scores of each class independently.

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



To fairly compare our boosting results with our Random Forest outcome, we perform cross-validation on three folds of the training data set as well.

However, since the XGBoost implementation we used is not supported by *scikit-learn* 's `.fit` method, the cross-validation must be done using `xgboost` 's own API. The output `boost_cv` is actually a *pandas* dataframe that tabulates the results of the cross-validation.

	train-merror-mean	train-merror-std	test-merror-mean	test-merror-std
0	0.366273	0.000547	0.385373	0.000749
1	0.360975	0.000926	0.379174	0.000721
2	0.358772	0.000521	0.377226	0.000836
3	0.356486	0.000489	0.375637	0.000913
4	0.354330	0.000464	0.374665	0.000778

We get the training set cross-validation score by getting the *merror* mean on the 50th `num_boost_round`, which is the final boosting phase. The *merror* is an accuracy error rate metric meant for multi-class labels.

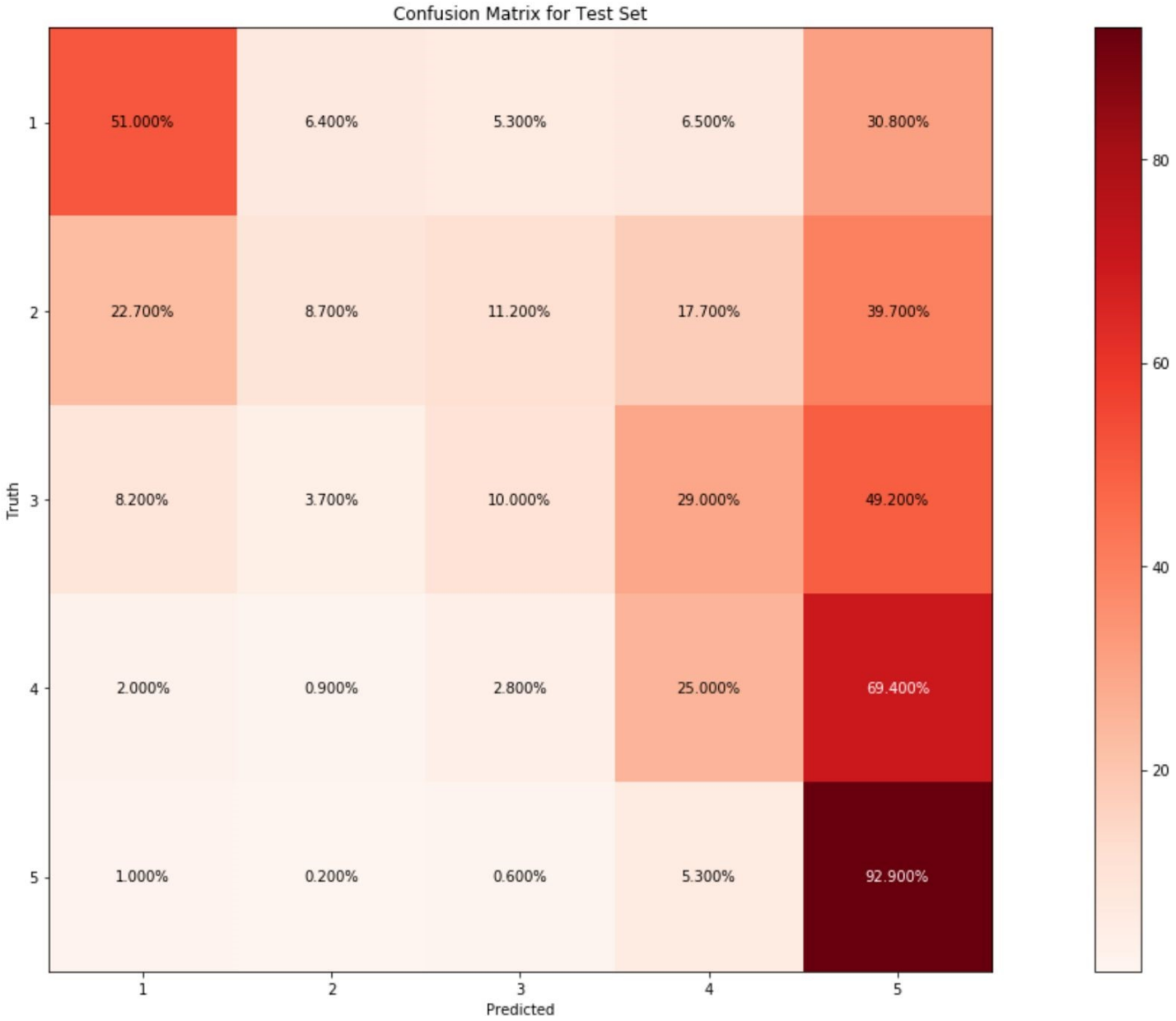
We can get a sense of how accurate the model is by subtracting the *merror* value from a perfect score of 100%.

```
Training Set Accuracy: 64.617%
```

Final Scores

Seeing that the boosting model outperformed the Random Forest approach in the three-fold cross validation, we can now apply our model on the testing set that we have put aside early on.

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



Our results above were actually based on the original `model_df` dataset that had the massive class imbalance. Let's now reassign our `X` and `y` variables to the balanced `trimmed_df` sample dataset we've created.

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

At 53.3% on a perfectly balanced training data set, we have achieved a better result compared to the 20% accuracy of our baseline.

Word Cloud

Using the true labels of the reviews, we can take the fifty most salient words in every rating and produce a word cloud. The same `stop_words` we derived from the NLTK library are excluded.

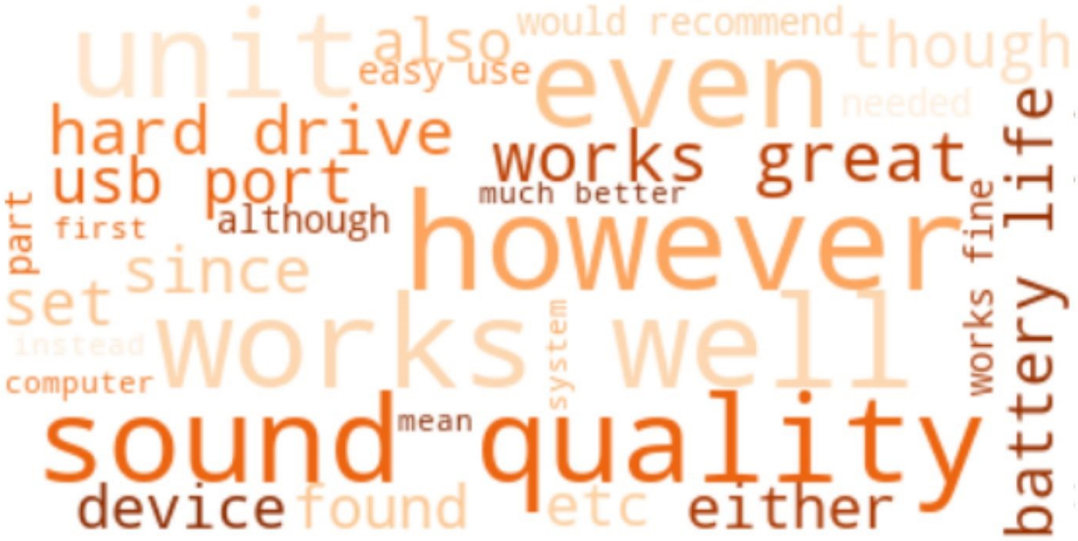
We see that some of the words are quite descriptive of the rating, with "problem" and "issue" frequently appearing in one-star reviews, and "quality" and "highly recommend" in top reviews.

Word Cloud for 1-Star Ratings

Word Cloud for 2-Star Ratings

Word Cloud for 3-Star Ratings

Word Cloud for 4-Star Ratings



Word Cloud for 5-Star Ratings

Conclusion

A lot of Natural Language Processing techniques were covered in the study. Just some of the concepts explored include topic modeling – where similar texts were clustered together according to a topic, named entity recognition (NER) – where nouns were given identifying labels like *place* or *time*, and dependency trees – where parts-of-speech tags and sentence structure were discerned. Though the *Word2Vec* phase was central to our final model, the pre-processing steps were perhaps just as crucial. Before tokenization, each document had to be decoded from UTF and encoded to ASCII, and converted to lowercase. The texts were stripped of accents, stop words, and punctuation, and multiple whitespaces were dropped. Words were simplified to their root words to compact the vocabulary as much as possible. Tokens that were often used together were also singularized through phrase modeling.

Beyond word use and word frequency, our model extracts and quantifies *context* . Every token in all the reviews is understood by its neighboring words and embedded in a given number of dimensions. All the interactions of a word with all the other words it has been associated with are expressed in vectors. And all the words in a given review are averaged according to each of the dimensions to create its 100 features. So the essence of a review by its words makes up the final dataframe.

What we have is a multi-class model where each of the five classes corresponds to a review's star rating. This is then a discrete approach where each class is independent of each other. In a situation where a 5-star rating is misinterpreted by the model as a 1-star review, then the model has simply misclassified – it is agnostic to how far off *1* and *5* are. This is in contrast with a *continuous* approach whereas a misclassification of a 5-star review as a 1-star review would be more penalizing. Our model then is reliant on the distinction of each kind of review. It is more concerned with asking "What makes a 5-star review different from a 4-star review?" than asking "Is this review *more* approving than criticizing?"

Limitations and Recommendations

Though we have observed satisfactory results in our model compared to the baseline, there are several limitations in the way the model handles data. These could serve as areas of improvement. First, despite a rich vocabulary, the model will not be able to handle words that it has not encountered during training. If an unknown word appears in a review, the word is dropped from the dimension-averaging step since has not been referenced in our word_vec_df .

Because each word is simplified by lemmatization during pre-processing, then alternate forms of a token shouldn't necessarily be a concern. However, the model cannot identify if a word is misspelled and will identify one simply as a new word. Incorporating a spellchecker would add to the computational cost and will certainly add to the model's complexity.

Finally, 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. Because sarcasm is usually detected by readers through the mood and sentiment of the document, it takes adding another layer of NLP just to approximate whether the review is sarcastic or not to properly work with such text. This supplement layer will not only utilize tagged sarcastic text as supervised labels but must also consider the review's given product rating in its judgment to detect sarcasm.