DSI Project 3

Web APIs & NLP

Done by: Group 2

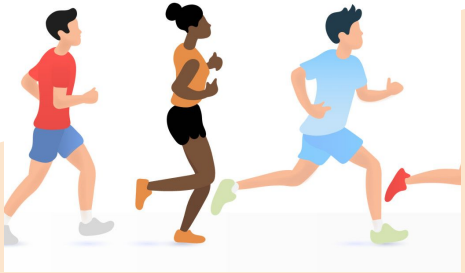
Chun Shan, Jason, Cindy, Shin

VS

Background

Retail brands such as Nike, Adidas etc have short seasons (up to 4 times a year).

These retail brands generate a lot of online discussion and comments daily and there are fast-moving conversations on Reddit.



+

=

Problem Statement

We are from Nike's commercial analytics team
Challenge: Hard to keep up with consumers' demand of multiple collections launching every season.
Stakeholder: Merchandise planning department has tasked us to find a solution to help with their planning for the upcoming season, by proxying popularity according to online conversations.

Understand **main differences** between Nike and its main competitor (Adidas) consumers' online conversation keywords

Understand whether consumer response towards Nike products are positive/ negative using **sentiment analysis**

Be able to **differentiate** Nike posts from competitor posts using a classification model

Find out the merchandise demand based on **top talked-about keywords and phrases**

Data Collection

For our data collection, we have decided to use Reddit, as it has proper segmentation of different communities in subreddits which help ease our web scraping process.

Data standardisation:

- Minimum 20 words in a post (title + selftext)
- 5000 posts from each brand's subreddit

Competitor selection: Adidas

Data Collection Process

- Create a function to filter post with minimum 20 words count while collecting the data
- Using https://files.pushshift.io to collect post and cut off at 5000
- Features that we collected:

Features	Description
author	The name of the author
subreddit	The subreddit community
selftext	The title of the post
title	The content of the post
created_utc	The epoch time

- Nike data was posted between 30 October 2022 and 10 November 2022
- Adidas data was posted between 22 October 2022 and 10 November 2022

Data Cleaning

- We notice that there are characters and hyperlink in certain post
- Removing:
  - Everything that comes with http
  - Newline: '\n'
  - Backslash: '\\'
  - Double space: ' '
  - Double quote', '\"'

Model: Air Force 1

Model: Air Max 95

Model: Dunk
Colour: Photon dust

Model: Adicolor Classics
(Primeblue collection)

Model: Ultraboost 22

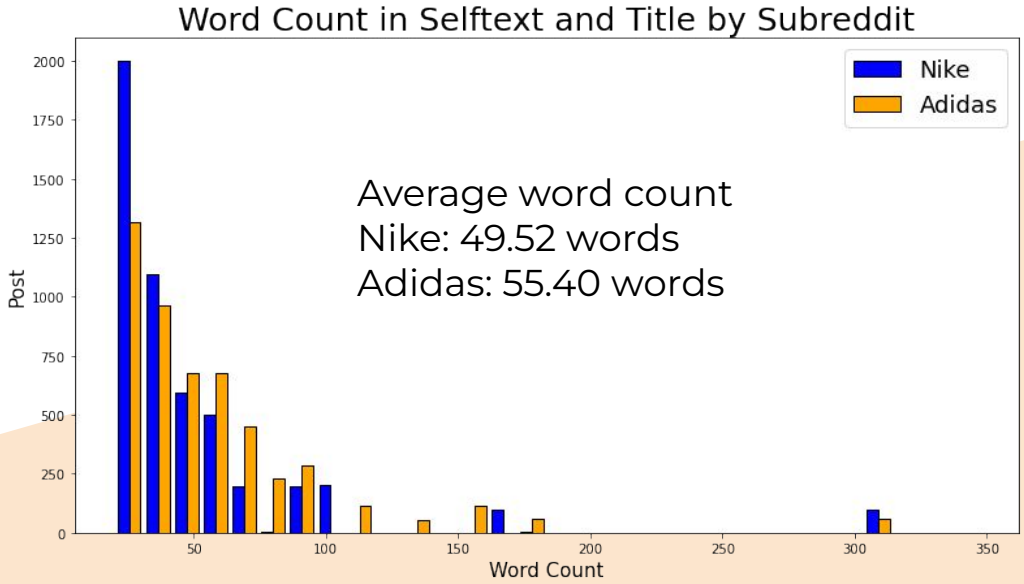
Model: Superstar (SST)
(Primeblue collection)

Post/User Analysis

Top 3 active users

r/Nike:
1. Extroe (394 posts)
2. Nervous-Matter4576 (198 posts)
3. Syranial-Bean (198 posts)

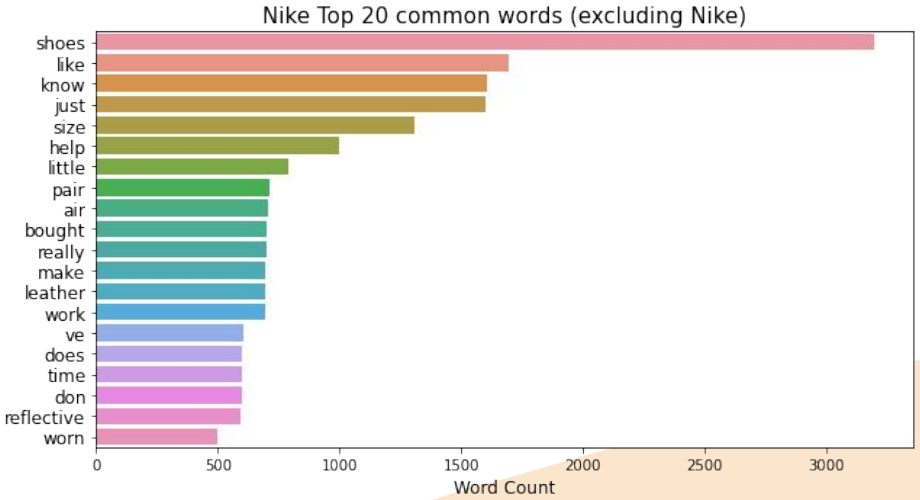
r/Adidas:
1. Neonnearvash (167 posts)
2. Alternative_Coconut6 (111 posts)
3. MCllooosey (110 posts)

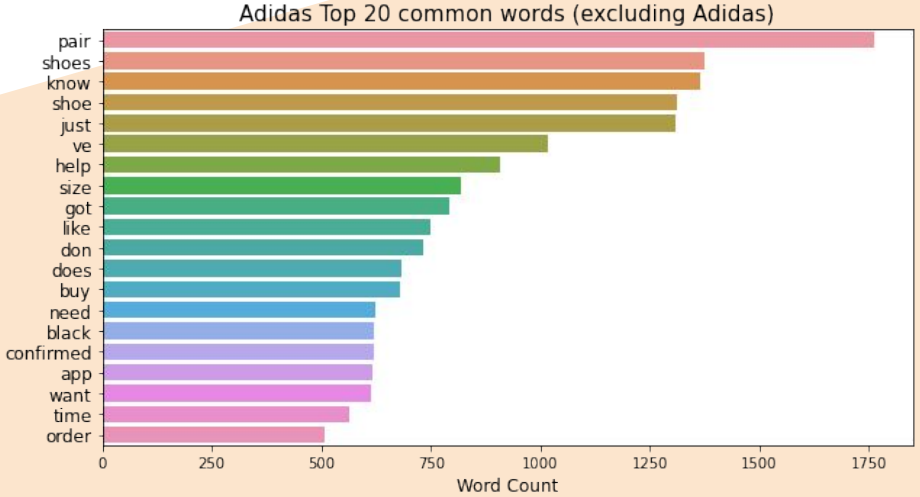


r/Nike has average lower word count (<50 words) than r/Adidas

Commonly Used Words

Both Subreddit	Nike	Adidas
shoes	little	shoe
like	air	got
know	bought	buy
just	really	need
size	make	black
help	leather	confirmed
pair	work	app
ve	reflective	want
does	worn	order
time		
don		





Commonly Used Words - Bigram

We can identify a few words that are related to each brand.

Nike

- **Style Names/Colourway:** 'air max'/'airmax 95', 'air force', 'jordan 1s', 'photon dust'
- **Product/Features:** 'basketball shoes', 'lace loops', 'flex experience', 'reflective lace'

Adidas

- **Style Names/Association:** 'nmd r1', 'shoes boost', 'kanye west'
- **Product/Features:** 'track jacket', 'boost sole'
- **UX/Service:** 'confirmed app', 'store pickup', 'adidas website'

Nike	Count
air max	398
nike shoes	299
does know	298
amp x200b	297
basketball shoes	297
photon dust	295
lace loops	295
reflective lace	295
loops 95s	295
dust reflective	295
flex experience	294
air force	205
ve worn	201
hey guys	200
jordan 1s	200
shoes really	198
13 wide	198
want know	198
tell legit	198
airmax 95	198

Adidas	Count
does know	285
need help	230
pair adidas	227
pair shoes	227
kanye west	226
adidas confirmed	226
adidas app	225
confirmed app	224
adidas shoes	171
store pickup	171
adidas website	169
track jacket	169
long term	169
don want	168
nmd r1	116
feels like	115
money dress	114
boost sole	114
shoes boost	114
ve received	114

Commonly Used Words – Trigram

Now we can see the combined word make more sense and more distinctive.

Nike

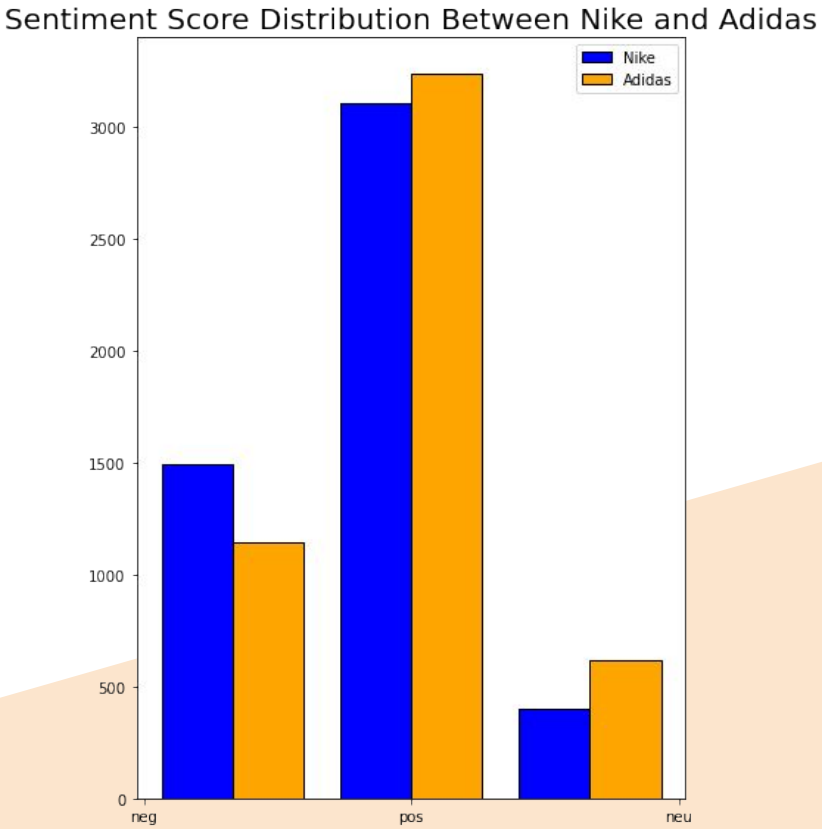
- **Style Names/Colourway:** 'photon dust reflective', 'air max 95', 'air max plus', 'colour photon dust', 'bought jordan 1s', 'air force 1s', 'mens dunks time'
- **Product/Features:** 'reflective lace loops', 'lace loops 95s'

Adidas

- **Style Names/Colourway:** 'bought adicolor classics', 'classics primeblue sst', 'adicolor classics primeblue', 'primeblue sst track', '22 boosts durable', 'zx 22 boosts'
- **Product/Features:** 'comfy sure age', 'like long term', 'boosts durable mean'
- **Association:** 'cut ties kanye', 'ties kanye west', 'adidas website'

Nike	Count	Adidas	Count
reflective lace loops	295	bought adicolor classics	114
lace loops 95s	295	classics primeblue sst	114
photon dust reflective	295	adicolor classics primeblue	114
guys date tag	198	primeblue sst track	114
air max 95	198	sure age like	113
air max plus	198	mean know look	113
date tag tell	198	comfy sure age	113
hey guys date	198	need sizing help	113
tag tell legit	198	term usage reviews	113
tell legit thank	198	boosts durable mean	113
colour photon dust	197	hella comfy sure	113
reflective help pls	196	age like long	113
dust reflective help	196	like long term	113
bought jordan 1s	100	know look good	113
know nike shoes	100	22 boosts durable	113
air force 1s	100	cut ties kanye	113
know right subreddit	100	zx 22 boosts	113
drop follow instagram	99	good hella comfy	113
mens dunks time	99	durable mean know	113
mens drop friday	99	ties kanye west	113

Sentiment Analysis



r/Nike	
pos	62.18
neg	29.80
neu	8.02

r/Adidas	
pos	64.76
neg	22.84
neu	12.40

r/Adidas has higher positive sentiment as compared to r/Nike

Baseline Model

Preprocessing:
- Tokenize
- Lemmatize
- Stem

Model	Vectorizer	Words	Train Score	Test Score	False Positive	False Negative
Naive Bayes	Count	Tokenize	0.9972	0.9970	5	5
Naive Bayes	Count	Lemmatized	0.9984	0.9970	6	4
Naive Bayes	Count	Stemmed	0.9978	0.9973	4	5

Tokenize	Stem
has	ha
anybody	anybodi
tried	tri
these	these
boots	boot
before?	before?
how	how
are	are
they	they
in	in
terms	term
of	of
quality	qualiti
and	and
fit?	fit?
do	do
they	they
look	look
better	better
or	or
worse	wors
in	in
hand	hand
compared	compar
to	to
pics?	pics?

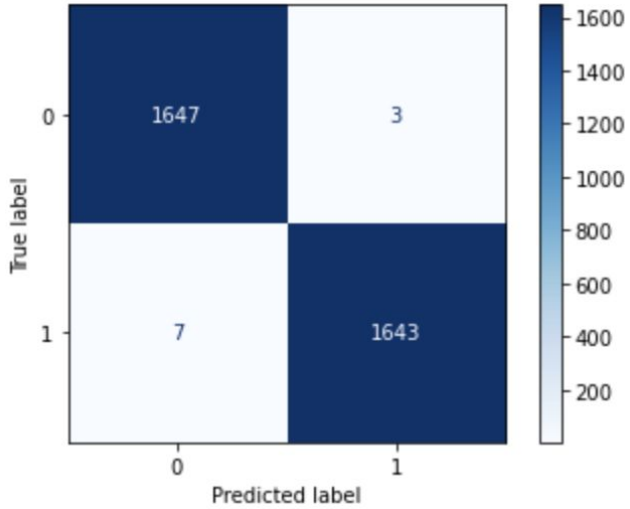
Model Tuning

Vectorizer:

- Countvectorizer
- TF-IDF Vectorizer

Model	Vectorizer	Train Score	Test Score	Precision	False Positive	False Negative
Naive Bayes	Count	0.9973	0.9967	0.9976	4	7
Naive Bayes	TF-IDF	0.9969	0.9970	0.9982	3	7

Model Tuning



Classifier:
- Naive Bayes (MultinomialNB)
- Random Forest Classifier

Model	Vectorizer	Train Score	Test Score	Precision	False Positive	False Negative
Naive Bayes	Count	0.9973	0.9967	0.9976	4	7
Naive Bayes	TF-IDF	0.9969	0.9970	0.9982	3	7
Random Forest	TF-IDF	1.0	0.9964	0.9964	6	6

Final Model

Minimise falsely identifying posts that are related to Nike (Identifier: 1) → Minimise false positives (Precision)

	Model	Vectorizer	Train Score	Test Score	Precision	False Positive	False Negative
	Naive Bayes	Count	0.9973	0.9967	0.9976	4	7
→	Naive Bayes	TF-IDF	0.9969	0.9970	0.9982	3	7
	Random Forest	TF-IDF	1.0	0.9964	0.9964	6	6

Final Model

```
gs_tvec.best_params_
```

```
0.9956716417910447

{'tvec__max_features': 5000, 'tvec__ngram_range': (1, 2)}
```

Recommendations

Sentiment Analysis
Should focus on improving sentiment before launching brand campaigns on r/Nike

Utilise popular bigram/trigrams
Can be used to determine which product lines are in-demand for campaigns

Improve generalizability of classifier
Train classifier on other online forums for more generalizable results across online demographics