SENTIMENT ANALYSIS PRIMER:

HOUSEHOLD CONSUMER PACKAGED GOODS (CPG)

ANALYSIS BY ROB BECOTTE 16 MARCH 2020

EXECUTIVE SUMMARY

Findings:

1. There is consistent high positive sentiment shared across examined brands
2. There is evidence validating consumer interest in environmental sustainability
3. @SeventhGen is the consistent leader in sentiment of the three brands on Twitter (Emulate their engagement style)
4. The SVM NLP* Classification model performed best on the Sensitivity score, however, there seems to be higher positive sentiment when predicting on these brands than the numbers suggest

*Support Vector Machine Natural Language Processing

CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
   A. Brand Sentiment
   B. Engagement Time Series Analysis
5. Recommendations and Next Steps

PURPOSE

1. **Catalyst:** Provide an independent data point for a NYC based startup
2. **Hypothesis:** There is high positive consumer sentiment towards using household goods that reduce waste (plastic) and promote environmental sustainability
3. **End State:** The startup will…
   A. Be armed with current consumer and market atmospherics
   B. Better target branding and marketing efforts
   C. Better align company vision with VCs or existing market leaders (M+A)

OBJECTIVES

1. Twitter Engagement Trends

A. **Magnitude** - How much

B. **Frequency –** How Often

C. **Timing -** When

D. **Clustering –** Why (Hard Part)

2. Sentiment Analysis

A. **Top Brands** – How do consumers feel about them? (% Positive Tweets)

B. **Top Features** – Why do consumers like these brands? (Convenience, social reasons, environment, ingredients, price)

C. **Find Meaningful Words** – Align marketing and branding

YARD STICKS

1. Machine Learning NLP Sentiment Analysis Benchmark for social media (Twitter):

- 60-80% Accuracy Rate

2. Mention Count: A mention is when someone uses the @ sign immediately followed by your Twitter Handle.

- @DrBronner
- @MrsMeyersClean
- @SeventhGen

(This analysis included the '@' in the web scrape to reduce ambiguity of handles)

3. Tag Count: An act of endorsement, which can be very powerful coming from an influencer with an engaged audience made up of people similar to your target market.

- #plasticfree

(This analysis omitted the '#' in the web scrape to capture all data points)

Hypothetical Examples

- **Model accurately predicts positive and negative sentiment in 3 to 4 out of 5 Tweets**
- **"Hey @DrBronner, I love your products!"**
- **"We should live greener #plasticfree**

https://sproutsocial.com/insights/twitter-mentions/

http://nlpprogress.com/english/sentiment_analysis.html

COMPANY ENGAGEMENT ACTIVITY SINCE INCEPTION

@SeventhGen	@DrBronner	@MrsMeyersClean
13K Tweets	**30K Tweets**	3K Tweets
83K Followers	54K Followers	11K followers

CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    A. Brand Sentiment
    B. Engagement Time Series Analysis
5. Recommendations and Next Steps

DATA COLLECTION

Train/Test NLP Data Sets:

- Kaggle – Twitter and Reddit Tweets (Binary Pos/Neg Labels)
- AWS – 6M Amazon Product Reviews (1-5 Star Label)

Data Scrapes:

- Twitter –GOT3 Python API
  - ~400K Tweets scraped from January 1 2018 to March 1 2020
- Reddit – Pushshift Python API
  - ~100K Reddit 'r/SkincareAddiction' posts NOT analyzed

DATA COLLECTION

Scrape Output

	artifact	datetime	text	retweets	username
0	@SeventhGen	2018-01-01 03:34:56-05:00	Hey Everyone Get samples, test products and make a difference: join me @SeventhGen's #GenerationGood http://h3.sml360.com/-/27f3a	0	Shantele_Marie
1	@SeventhGen	2018-01-01 04:38:23-05:00	They've got fun products. Get samples, test products and make a difference: join me @SeventhGen's #GenerationGood http://h3.sml360.com/-/27f44	0	Shantele_Marie
2	@DrBronner	2018-01-01 11:00:05-05:00	Grateful for every person who believes in the All-One Mission, devoted to love, respect & equality for all. Every employee who mobilizes daily with a palpable passion. Every customer who feels called to be of service to the world, empowering us to do more, do better.	8	DrBronner
3	@DrBronner	2018-01-01 11:00:06-05:00	This year, we donated approximately $7 million to philanthropic causes—from animal advocacy & fair trade supply chains to drug policy reform & LGBTQIA equality. We did that together.	3	DrBronner
4	@DrBronner	2018-01-01 11:00:06-05:00	We have more work to do in 2018 to stand up for people-planet-animals, and one day achieve our mission of unifying the human race. Onwards!	4	DrBronner
...	...	...	...	...	...
375544	plasticfree	2020-02-28 16:21:07-05:00	@refill @cocacola maybe this is future! #plasticfree	0	EnvironmentPlym
375545	plasticfree	2020-02-28 16:30:12-05:00	Chessel Bay March Clean Up - Sat 14 March 2020 http://www.greenhampshire.co.uk/events/564/Chessel-Bay-March-Cleanup #Southampton #LitterPick #BeachClean #PlasticFree #NurdleHunters	0	GreenHampshire
375546	plasticfree	2020-02-28 16:31:17-05:00	Sé parte de la iniciativa para generar un cambio en nuestro planeta. #RegresandoAlOrigen #KiriPlanet #ECO #MedioAmbiente #EcoFriendly #ReduceWaste #ZeroWaste #PlasticFree	4	KiriPlanet
375547	plasticfree	2020-02-28 16:34:32-05:00	¡Empaque totalmente amigable con el medio ambiente!. #RegresandoAlOrigen #KiriPlanet #ECO #MedioAmbiente #EcoFriendly #ReduceWaste #ZeroWaste #PlasticFree	4	KiriPlanet
375548	plasticfree	2020-02-28 16:51:27-05:00	Be a planet saver with Tavos! #ecofriendly #paperstraws #plasticfree #planet #plasticfreeoceans #saveenvironment #sustainability #Biodegradable #Compostable	0	TavosCanada

375549 rows × 17 columns

Feature Engineering (My Additions)

year	month	day	month_year	hour	neg	neu	pos	compound	vader_pred	svm_pred	svm_proba
2018	1	Monday	2018-01	3	0.0	0.855	0.145	0.2960	1	1	0.625190
2018	1	Monday	2018-01	4	0.0	0.718	0.282	0.6705	1	1	0.676501
2018	1	Monday	2018-01	11	0.0	0.665	0.335	0.9594	1	1	0.714697
2018	1	Monday	2018-01	11	0.0	0.916	0.084	0.3182	1	1	0.722890
2018	1	Monday	2018-01	11	0.0	1.000	0.000	0.0000	0	1	0.685856
...	...	...	...	...	...	...	...	...	...	...	...
2020	2	Friday	2020-02	16	0.0	1.000	0.000	0.0000	0	1	0.710810
2020	2	Friday	2020-02	16	0.0	0.838	0.162	0.4019	1	1	0.702101
2020	2	Friday	2020-02	16	0.0	1.000	0.000	0.0000	0	1	0.778596
2020	2	Friday	2020-02	16	0.0	1.000	0.000	0.0000	0	1	0.736188
2020	2	Friday	2020-02	16	0.0	1.000	0.000	0.0000	0	1	0.739465

CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    A. Brand Sentiment
    B. Engagement Time Series Analysis
5. Recommendations and Next Steps

BASELINE MODEL ACCURACY (OFF THE SHELF ALGORITHM)

VADER Sentiment Analyzer Performance

Data Set	Data Set	Data Set	Data Set
Amazon Reviews	Kaggle Twitter #1 (Indian English Tweets)	Reddit	Kaggle Twitter #2
Long Varied Reviews	Tweet	Posts	Tweet
54% Accuracy	57% Accuracy	63% Accuracy	**64% Accuracy**

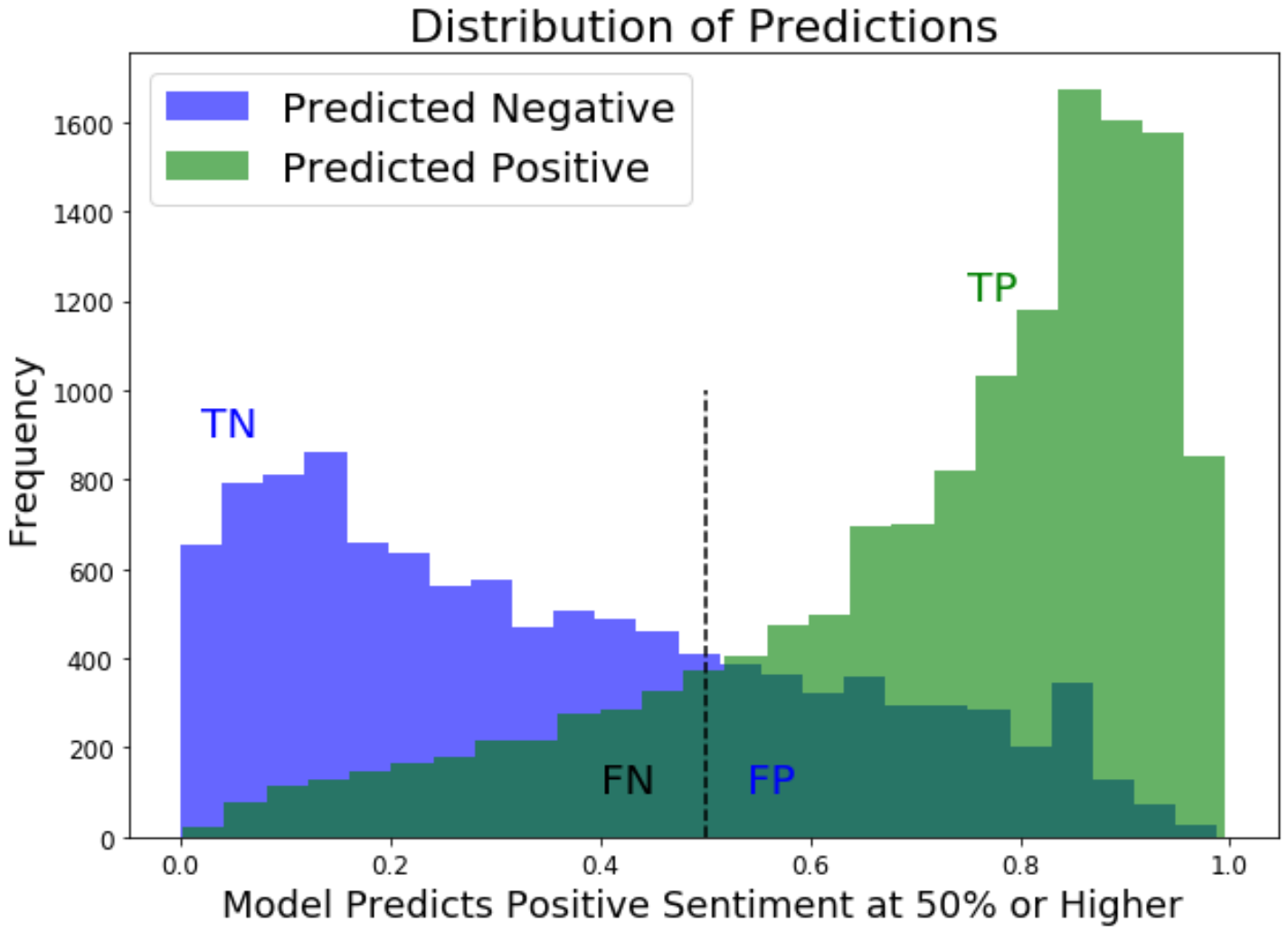
CUSTOM MODEL EVALUATION

Model	Compute Time	Best Parameters	Train Accuracy	Test Accuracy
VADER	5 Minutes		64%	64%
Random Forest	18 Minutes	TFIDF, 20K Tokens Grams: (1,3)	**99%**	75%
MNB	7 Minutes	Tandem Grid CV	89%	72%
RNN	27 Minutes	1 Hidden Layer, 600K Params	78%	76%
SVM	**360 Minutes**	20K Tokens C=1.0 Kernel='rbf'	95%	**78%**

SUPPORT VECTOR MACHINE (CLASSIFIER) RESULTS

Insight

- Predictions have an appropriate skew
- The high confidence predictions were generally accurate



SUPPORT VECTOR MACHINE (CLASSIFIER) TRAINING RESULTS

Insight

- Performs **best** at predicting positive sentiment **(Sensitivity)**

  Business Advice: *Use this model for identifying positive influencers and PR wins*

- **Underperforms** when predicting negative sentiment **(Specificity)**

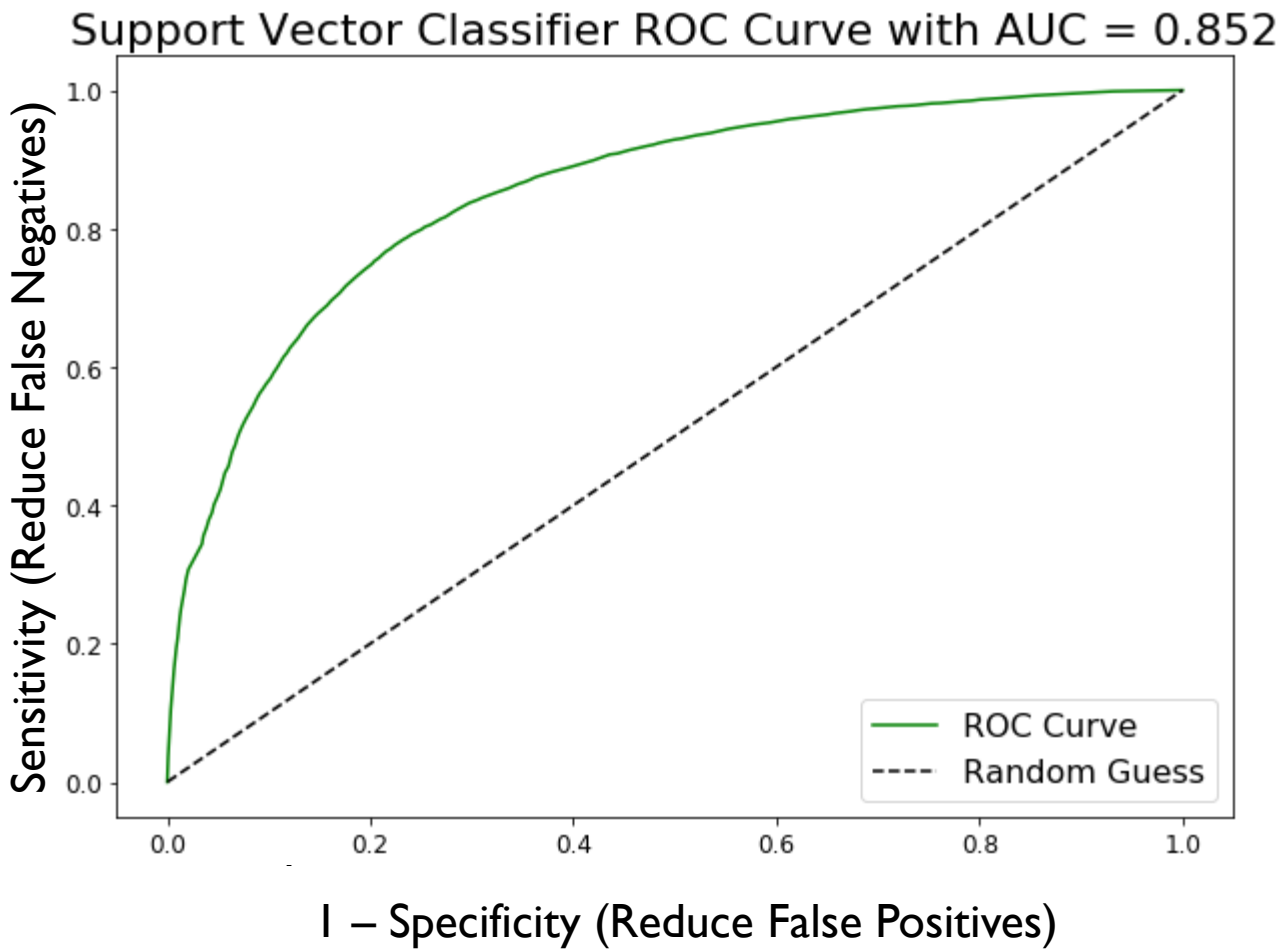
  Business Advice: *Avoid if looking for negative feedback*

	Predicted Negative Tweet	**Predicted Positive Tweet**	
Actual Negative Tweet	7668	Type I Error 3294	Specificity 70%
Actual Positive Tweet	Type II Error 2262	11774	**Sensitivity 84%**
		Precision 78%	Accuracy 78%

SUPPORT VECTOR MACHINE (CLASSIFIER) RESULTS

Insight

- 85% probability of rating a Positive tweet higher than a Negative Tweet



WHERE DID THE MODEL GUESS WRONG?

Twitter Training Data

- Model struggles to pick up sarcasm

- I randomly sampled 10 incorrect guesses… I agreed with the model on 6 of 10

```
1 wrong_guess_df.sort_values(by='predicted')
```

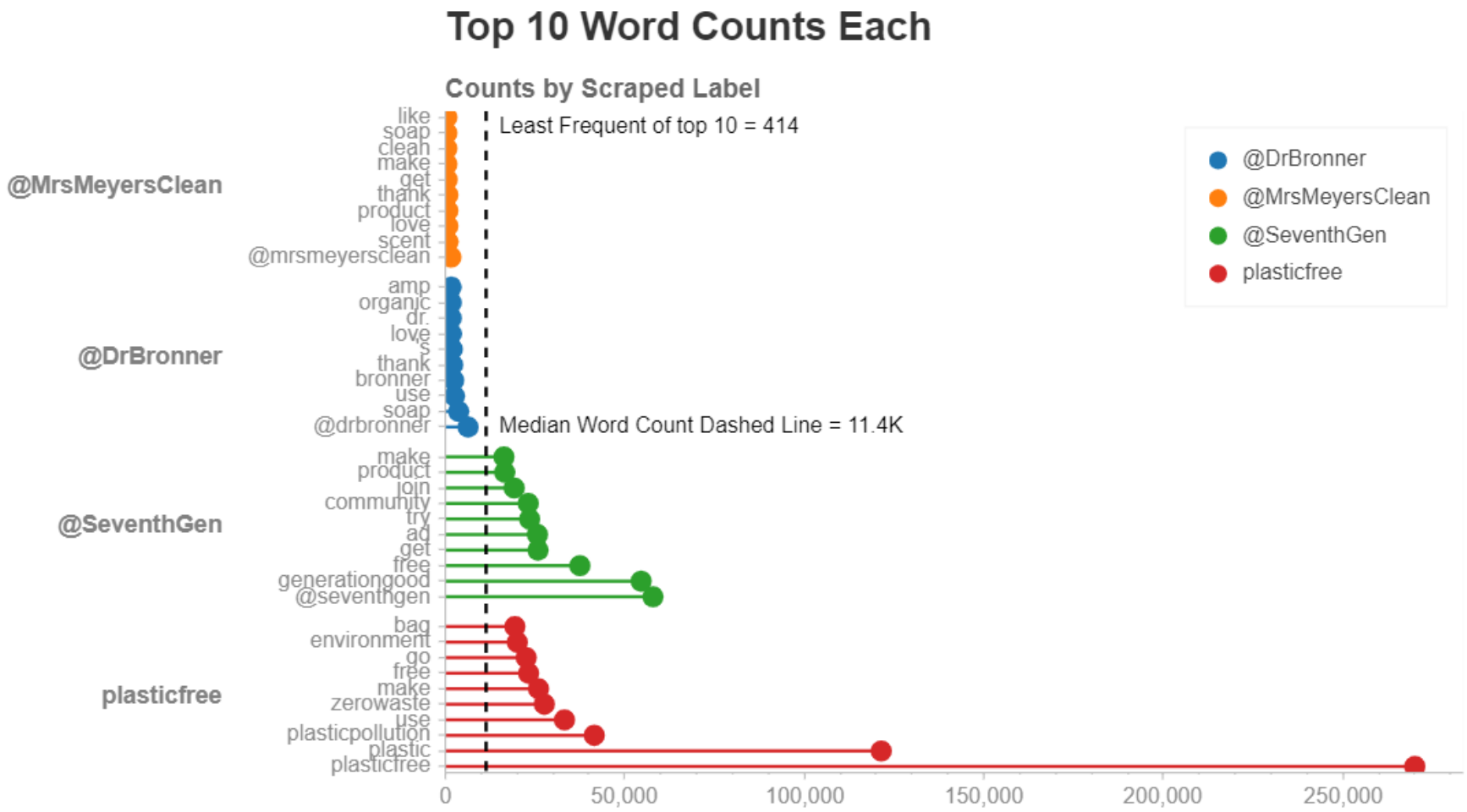
executed in 17ms, finished 16:36:15 2020-03-11

	predicted	actual	title
5371	0	1	@alesiaxx too bad about you bit becoming a red wings fan; you don't know what you are missing! LOL
7159	0	1	@brundlefly no, not those Zombies these zombies http://bit.ly/8VQY1
20876	0	1	i have an idea im going to get a gun go to taylor l's house and kill him!! IM A VERY BAD GIRL x]
20880	0	1	@anthonyjohnston Oh no...wasn't here when the mean Nurse arrived...hope you remembered to take my arm with you! Don't faint! x
7147	0	1	@A11woman Till he wakes as #bgt would not be on his watching agenda!
...	...	...	...
10857	1	0	@aineODM noo kindof wish it was now. how long you out there for? don't be lazy, write the novel! haha
10860	1	0	@bluntmag I just saw your Lyn-Z poster... I really love it but I live in the States and nowhere imports your mag here
10862	1	0	@apache_rose haha. I really love Jaylor! I wish they were a couple now
10224	1	0	... #andnav US server is back! seems that it was a provider problem
24994	1	0	@brianwelburn So happy to be going to work!! who wants to spend time in the sun eh!!

CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    A. Brand Sentiment
    B. Engagement Time Series Analysis
5. Recommendations and Next Steps

TOP 10 MOST OCCURRING WORDS BY LABEL



5.9M tokens in Corpus

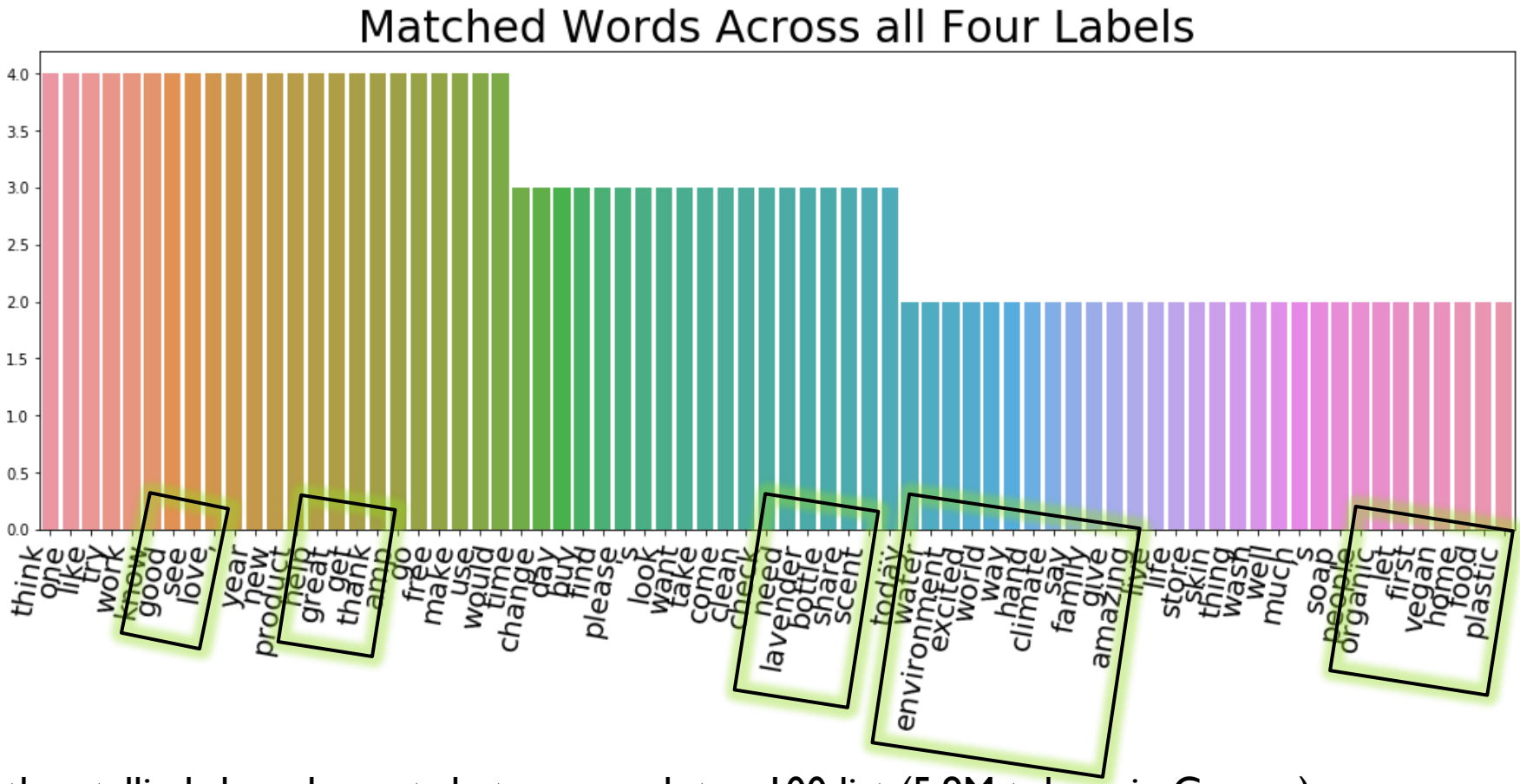
SHARED THEMES BETWEEN ACCOUNTS

Insights

- Highest shared interest in:

1. Positive words
2. Features: 'Lavender', 'scent', organic', 'vegan', 'plastic'
3. Environment

Business Advice: *Ensure business philosophy, branding, and actions align with these concepts*

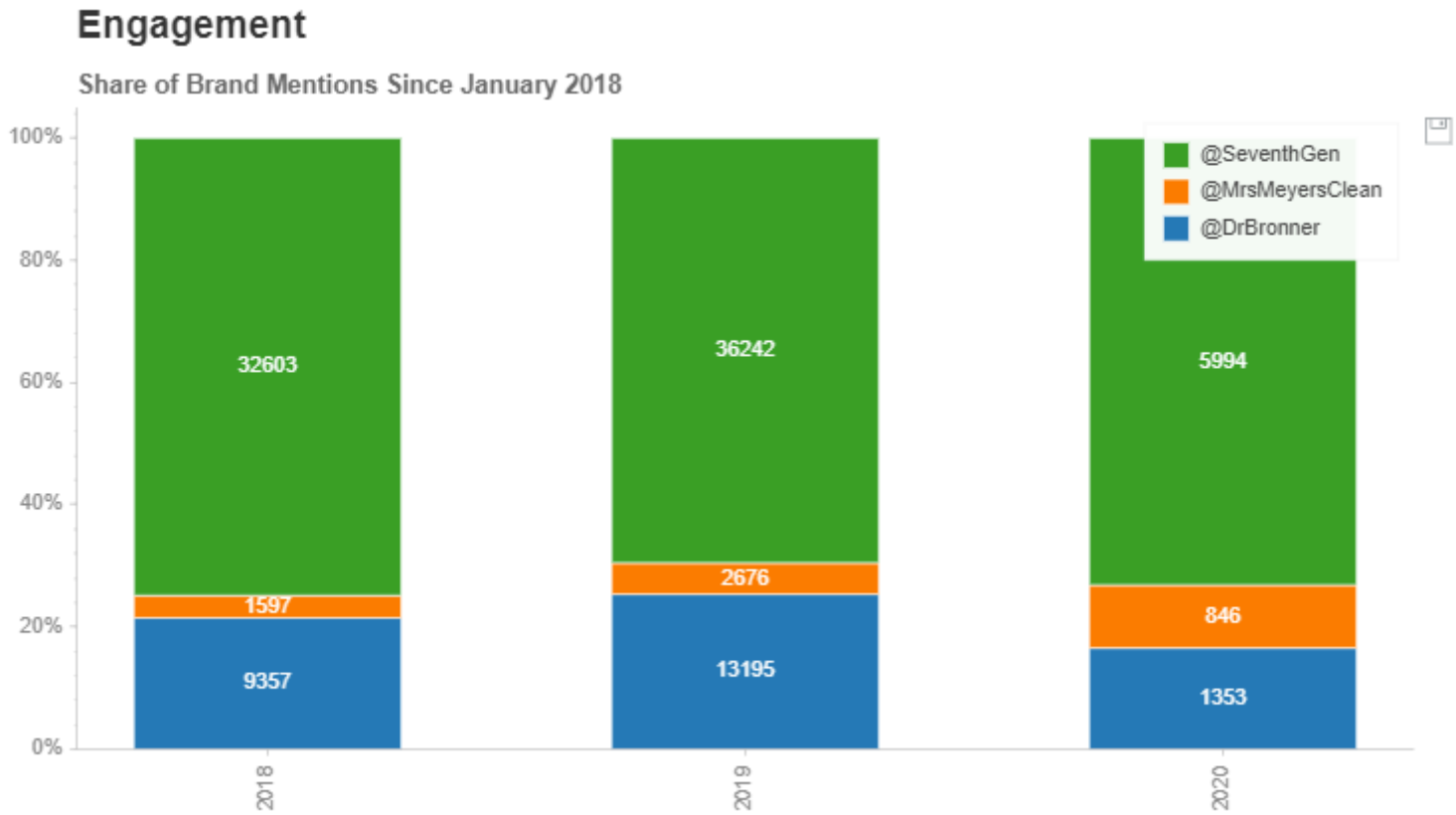


Collated top 100 words from each label, then tallied shared counts between each top 100 list (5.9M tokens in Corpus)

ENGAGEMENT BY BRAND BY YEAR

Insights

- @SeventhGen has the largest Twitter footprint
- @MrsMeyersClean is proportionately increasing their engagement year over year



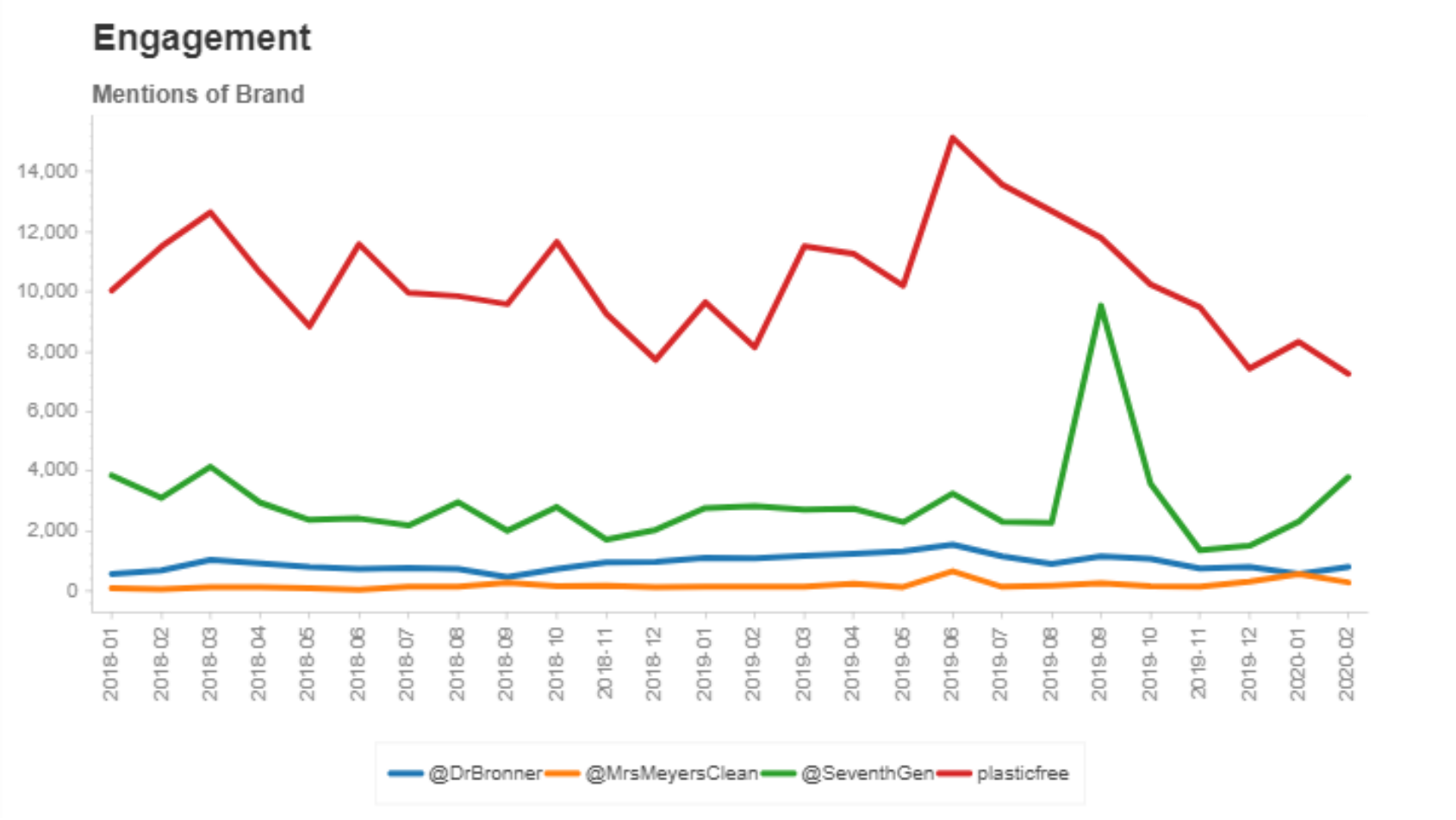
ENGAGEMENT BY MONTH

Insights

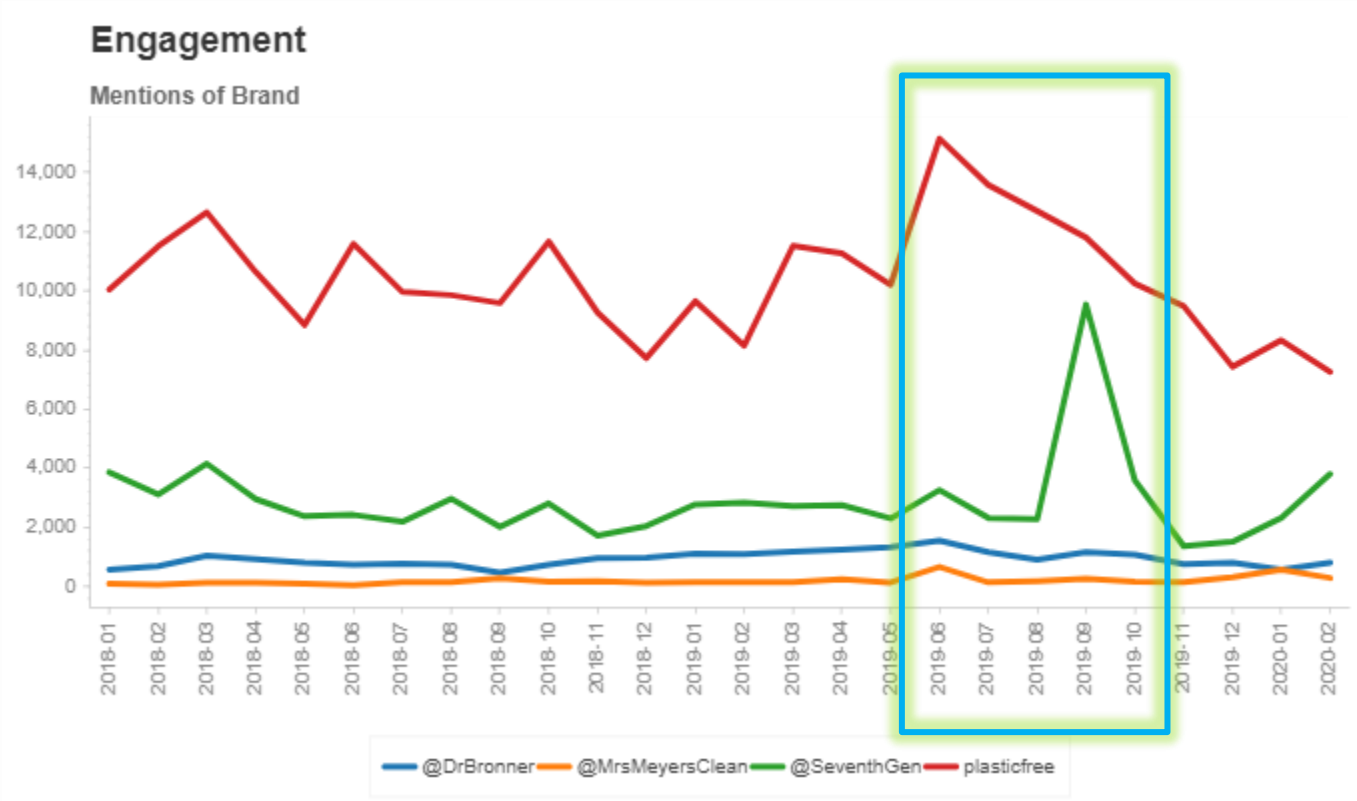
- #plasticfree averages 2.5X the engagement of @SeventhGen

- Generally plateaued mention counts across the board

  Business Advice: *Seek an opportunity to springboard product launch with a high-vis PR event*



#CLIMATESTRIKE

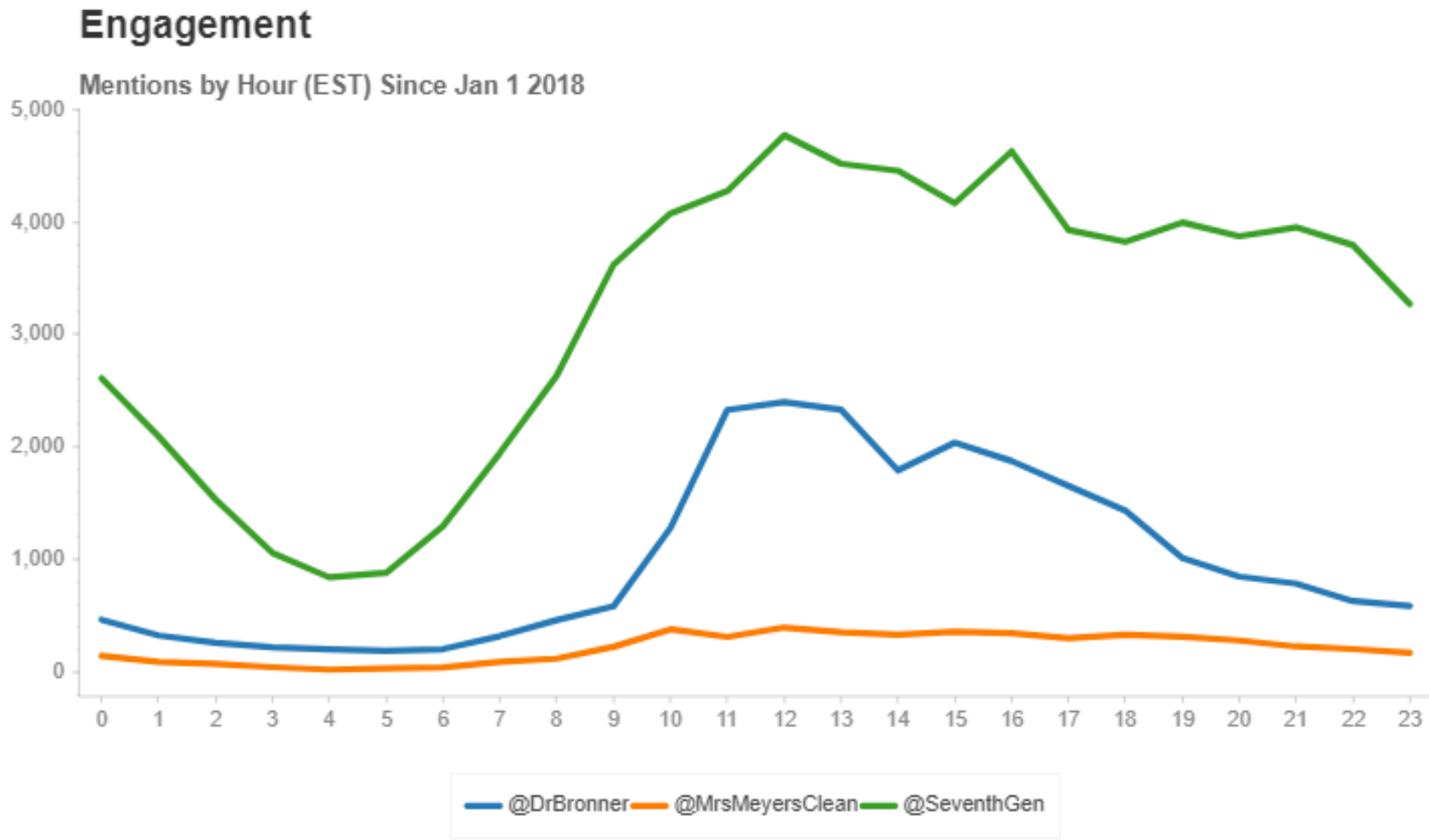


AGGREGATED 2 YEAR ENGAGEMENT BY HOUR

Insights

- Noticeable seasonality:
  - 6AM-12PM Ascent
  - 12AM-5AM Descent

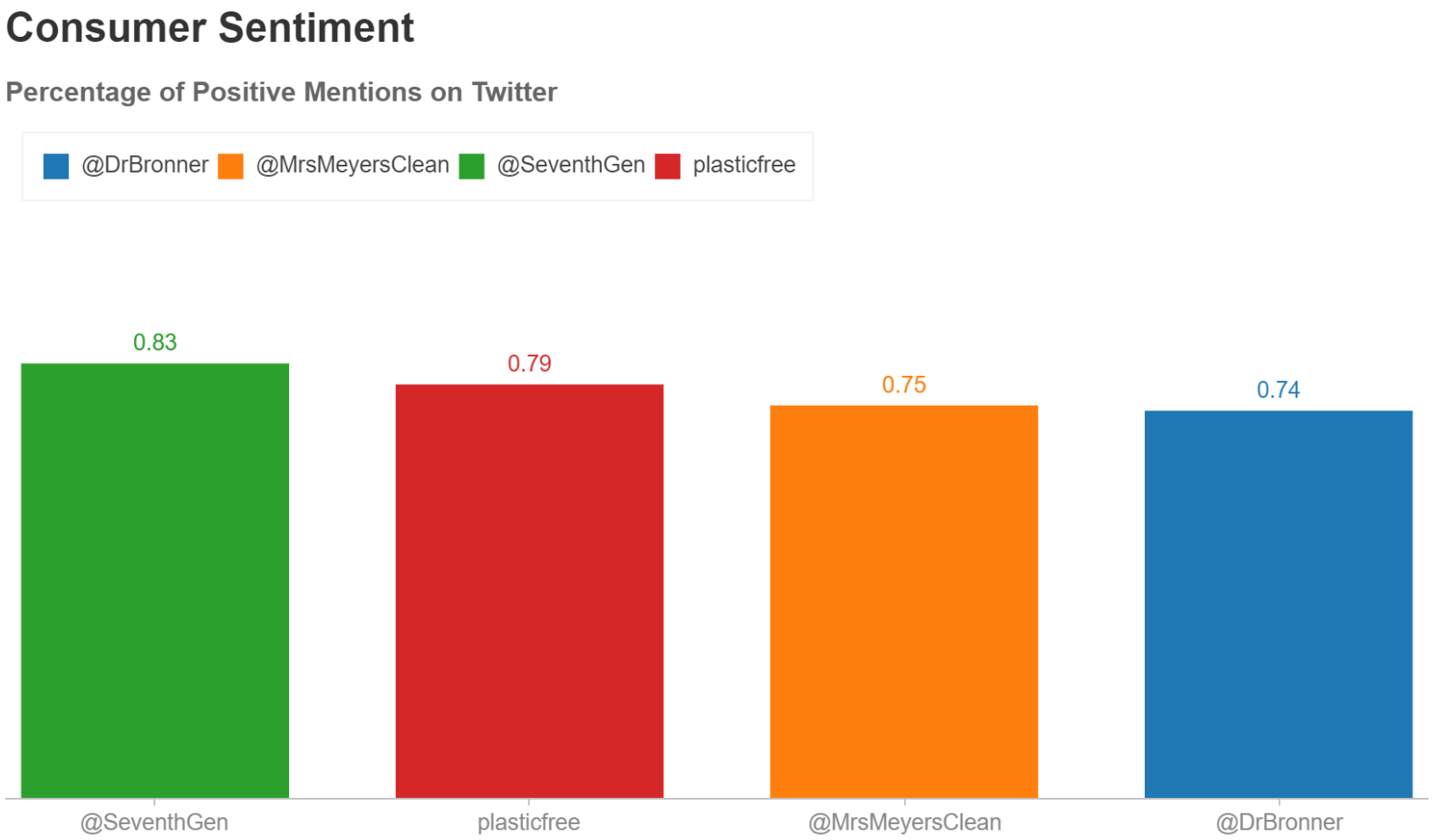
Business Advice:
Engage your future consumers when they are active 7AM-7PM



PERCENTAGE OF MENTIONS THAT ARE POSITIVE

Insights

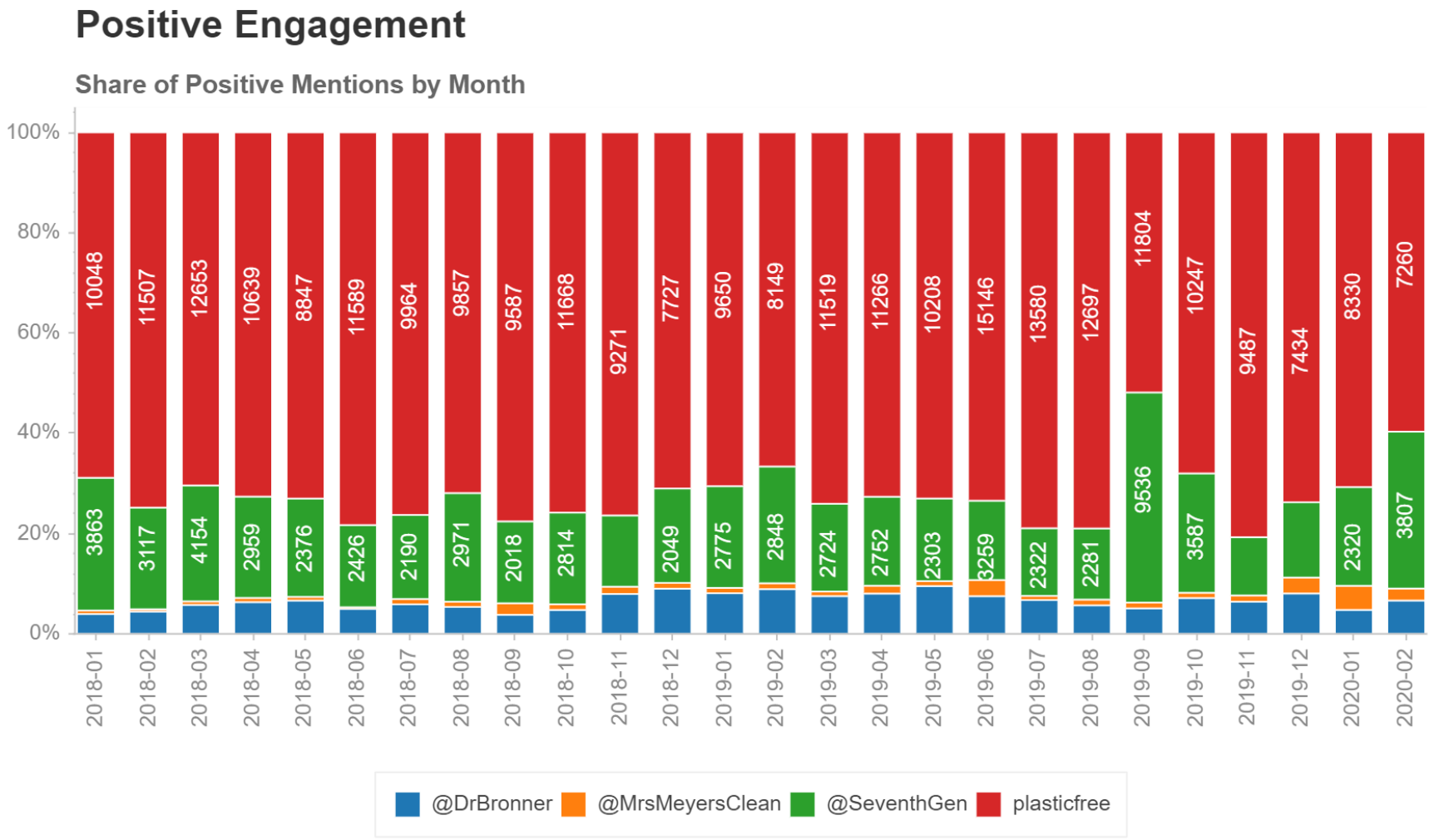
- Most mentions are classified as positive for all labels
- @SeventhGen receives the highest positive sentiment relative to mentions



SHARE OF POSITIVE ENGAGEMENT

Insights

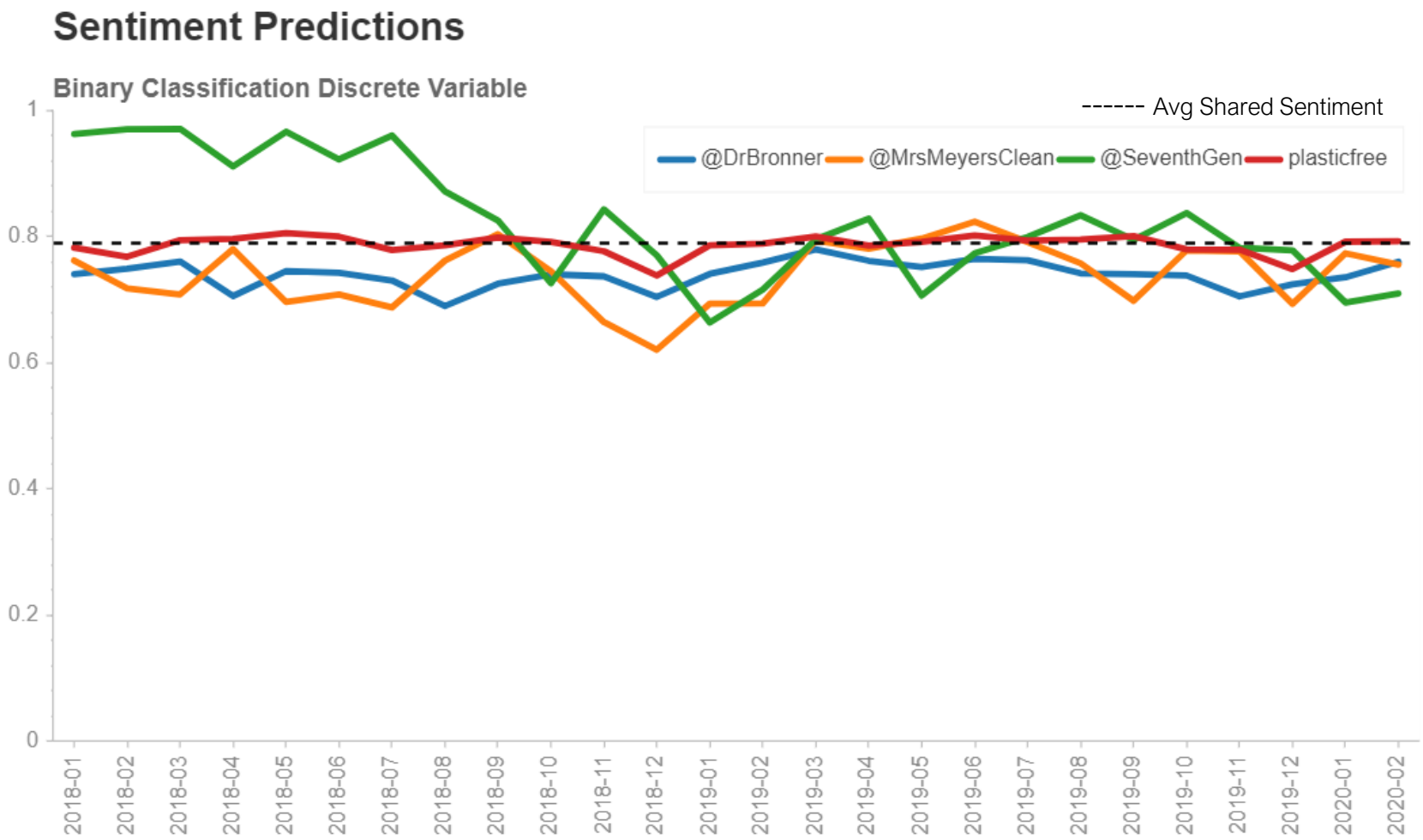
- #plasticfree leads positive engagement by both share and count



MONTHLY AVERAGED POSITIVE SENTIMENT PERCENTAGES

Insights

- #plasticfree is the current leader of the four in positive sentiment
- @SeventhGen maintains a higher averaged percentage, though it continues to incrementally decrease
- Avg Sentiment during period was 79%



CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    A. Brand Sentiment
    B. Engagement Time Series Analysis
5. Recommendations and Next Steps

SUMMARY AND NEXT STEPS

Hypothesis:

There is high positive consumer sentiment towards using household goods that reduce waste and promote environmental sustainability

Findings:

1. There is consistent high positive sentiment shared across examined brands
2. There is evidence validating consumer interest in environmental sustainability
3. @SeventhGen is the consistent leader in sentiment of the three brands on Twitter (Emulate their engagement style)
4. The SVM NLP* Classification model performed best on the Sensitivity score, however, there seems to be higher positive sentiment when predicting on these brands than the numbers suggest

Recommendations without additional analysis:

1. Align business philosophy, model, branding, and actions consistent with these companies (sustainability)
2. Seek out large scale events and influencers to promote brand and launch
3. Engage consumers while they are active: 7AM to 7PM