SENTIMENT ANALYSIS:

HOUSEHOLD CONSUMER PACKAGED GOODS (CPG)

ANALYSIS BY ROB BECOTTE

CONTENT

1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    1. Brand Sentiment
    2. 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 and promote environmental sustainability
3. **End State:** The startup will…
    1. Be armed with current consumer and market atmospherics
    2. Better target brand and marketing efforts
    3. Better align company vision with VCs or existing market leaders (M+A)

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

OBJECTIVES

1. Twitter Engagement Trends

A. **Magnitude - How much**

B. **Frequency – How Often**

C. **Timing - When**

D. **Clustering - Why**

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/branding

EXECUTIVE SUMMARY

Findings:

1. There is consistent high positive sentiment shared across brands
2. There is consistent evidence on the importance of environmental sustainability leading to high consumer sentiment
3. @SeventhGen is the consistent leader of the three on Twitter (Emulate their engagement style)
4. The SVM NLP Classification model scores best on Sensitivity (True Positive Rate)

KEY PERFORMANCE INDICATORS

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**

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**

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

BASELINE MODEL ACCURACY

VADER Sentiment Analyzer Performance

Data Set	Data Set	Data Set	Data Set
Amazon Reviews	Kaggle Twitter (India Tweets)	Reddit Twitter (India Tweet)	Kaggle Twitter
Long Reviews	Tweet	Tweet	Tweet
54% Accuracy	57% Accuracy	63% Accuracy	**64% Accuracy**

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

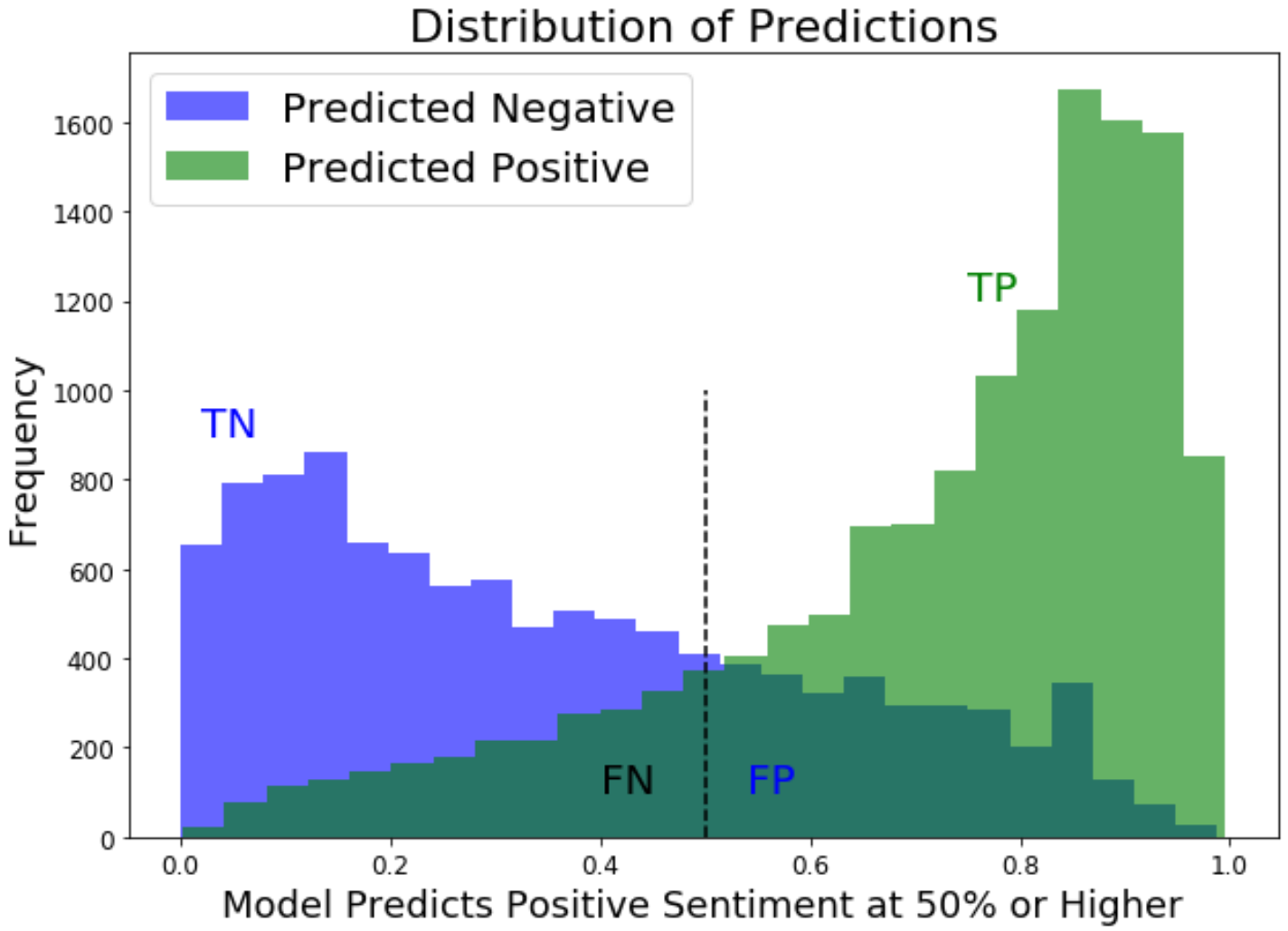
CUSTOM MODEL EVALUATION

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) 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)**

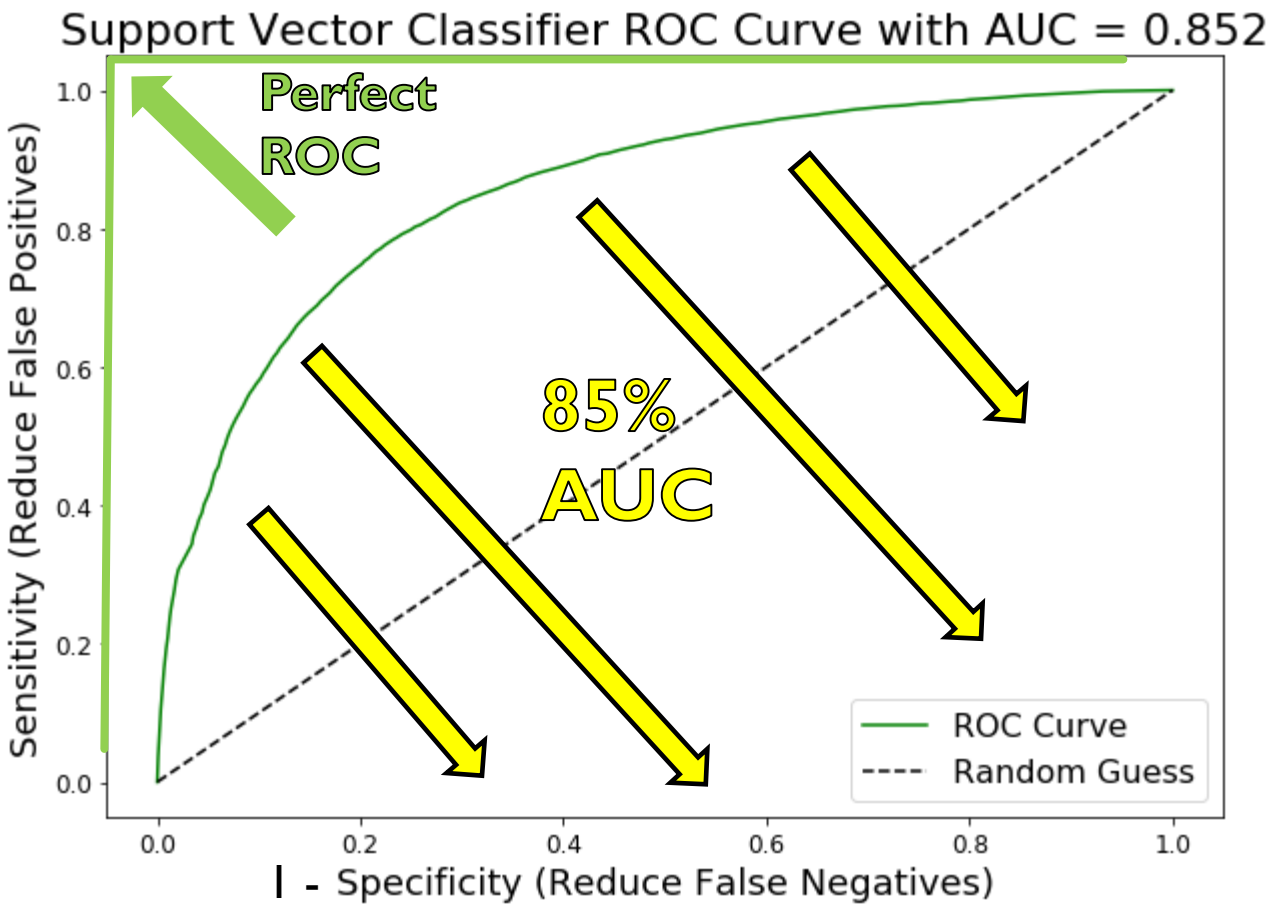
  Business Advice: *Avoid if looking for negative feedback*

	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

ROC: Receiver Operating Curve

AUC: Area Under (ROC) Curve



Insight

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

WHERE DID THE MODEL GUESS WRONG?

Twitter Training Data

- Models 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')
```

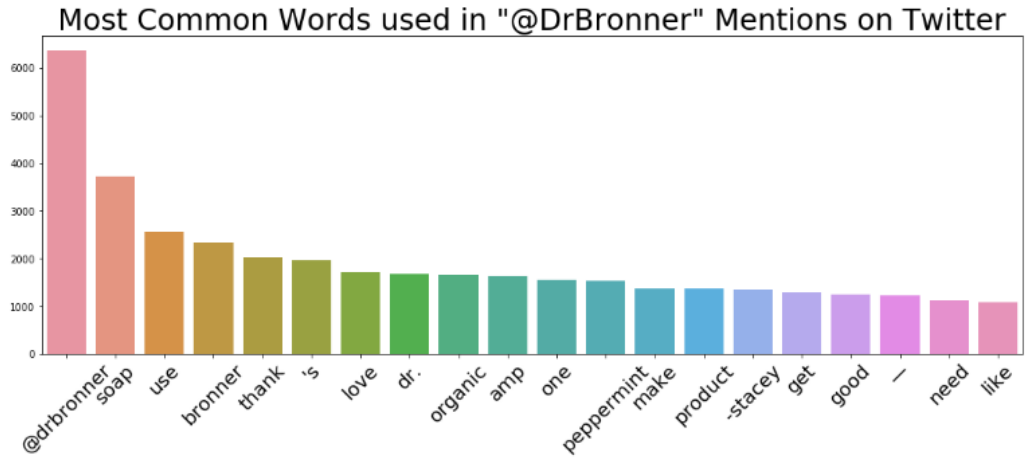
executed in 17ms, finished 16:36:15 2020-03-11

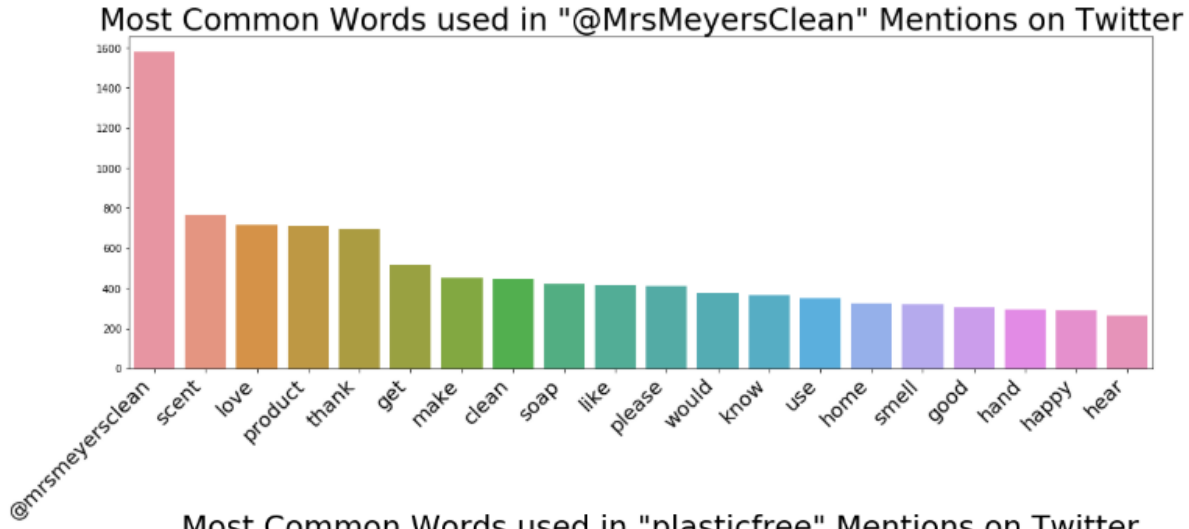
	predicted	actual	title
5371	0	1	@alesiaxx too bad about you bit becoming a red wings fan; you don't know what you are missing! LOL
7159	0	1	@brundlefly no, not those Zombies these zombies http://bit.ly/8VQY1
20876	0	1	i have an idea im going to get a gun go to taylor l's house and kill him!! IM A VERY 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!!

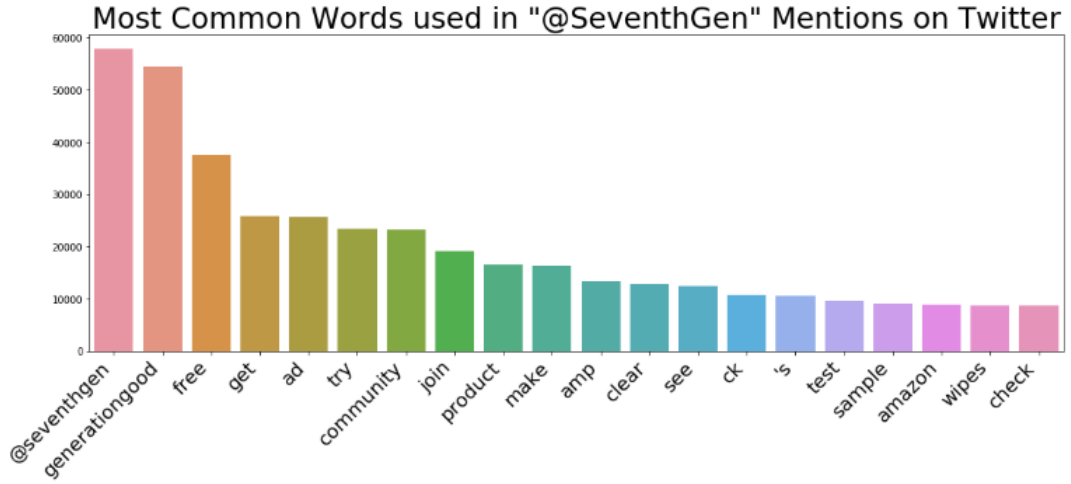
CONTENT

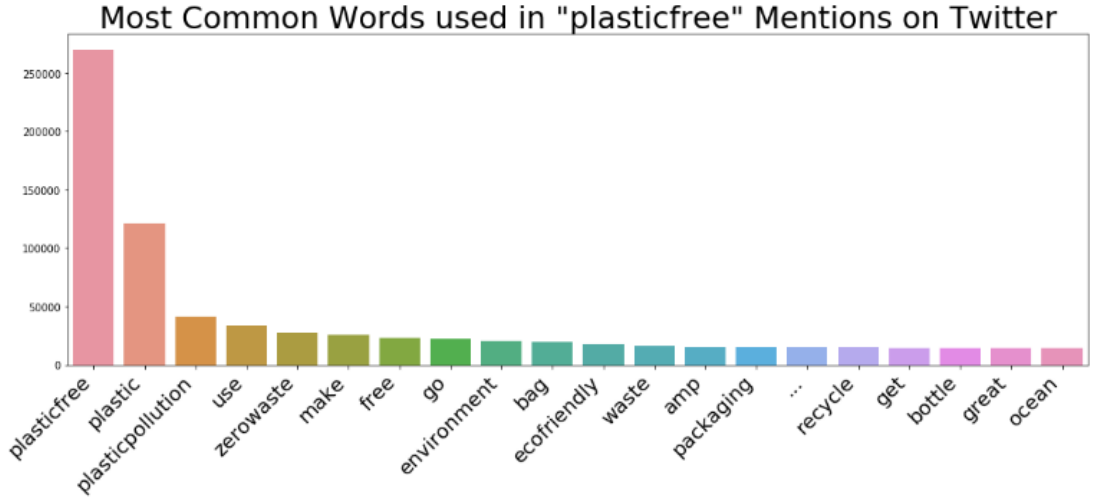
1. Orientation
2. Data Collection
3. Model Comparison and Performance
4. Findings
    A. Brand Sentiment
    B. Engagement Time Series Analysis
5. Recommendations and Next Steps

THE MOST FREQUENTLY USED WORDS









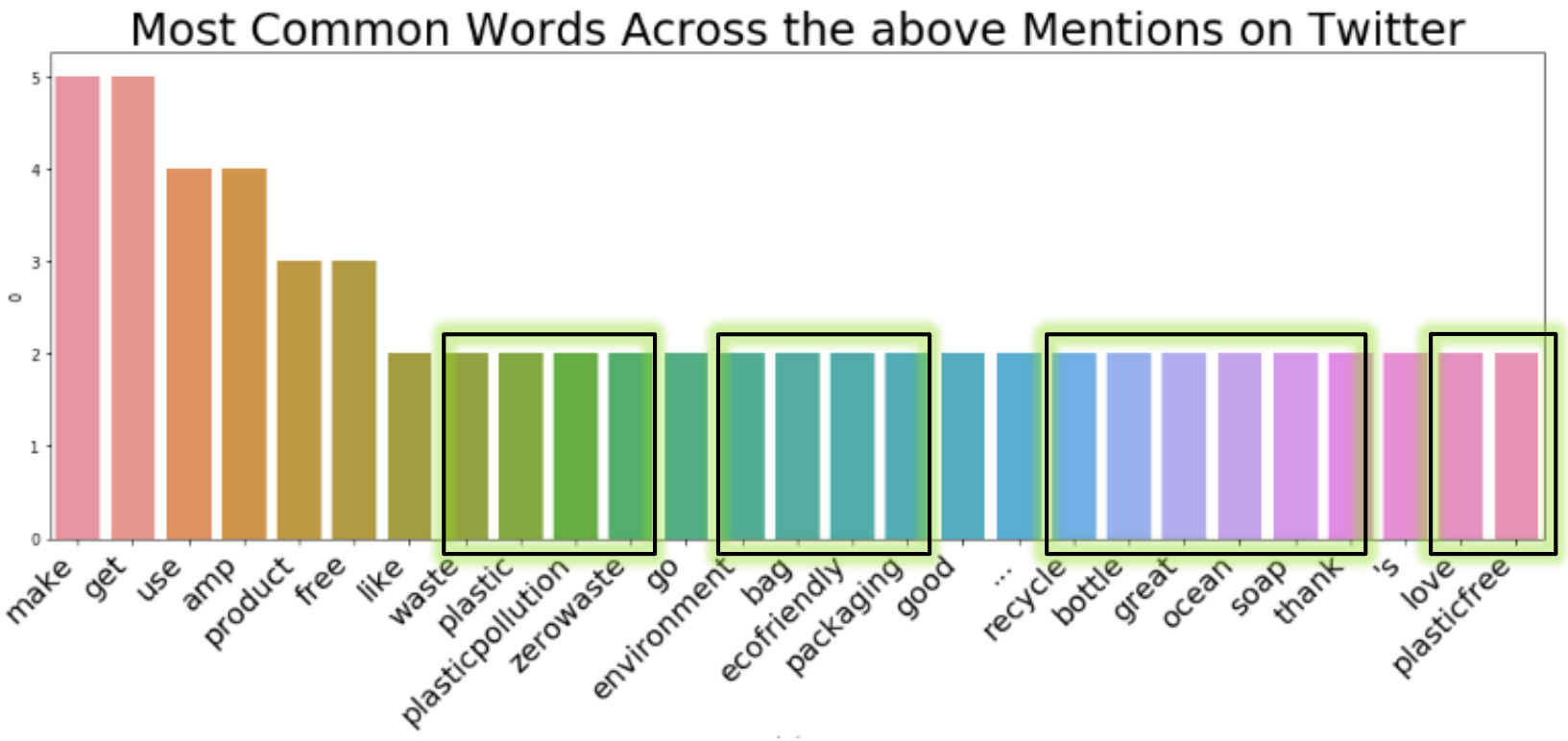
SHARED THEMES BETWEEN ACCOUNTS

Insights

- Highest interest in:

1. Reducing plastic
2. Reducing waste
3. Protecting the environment

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



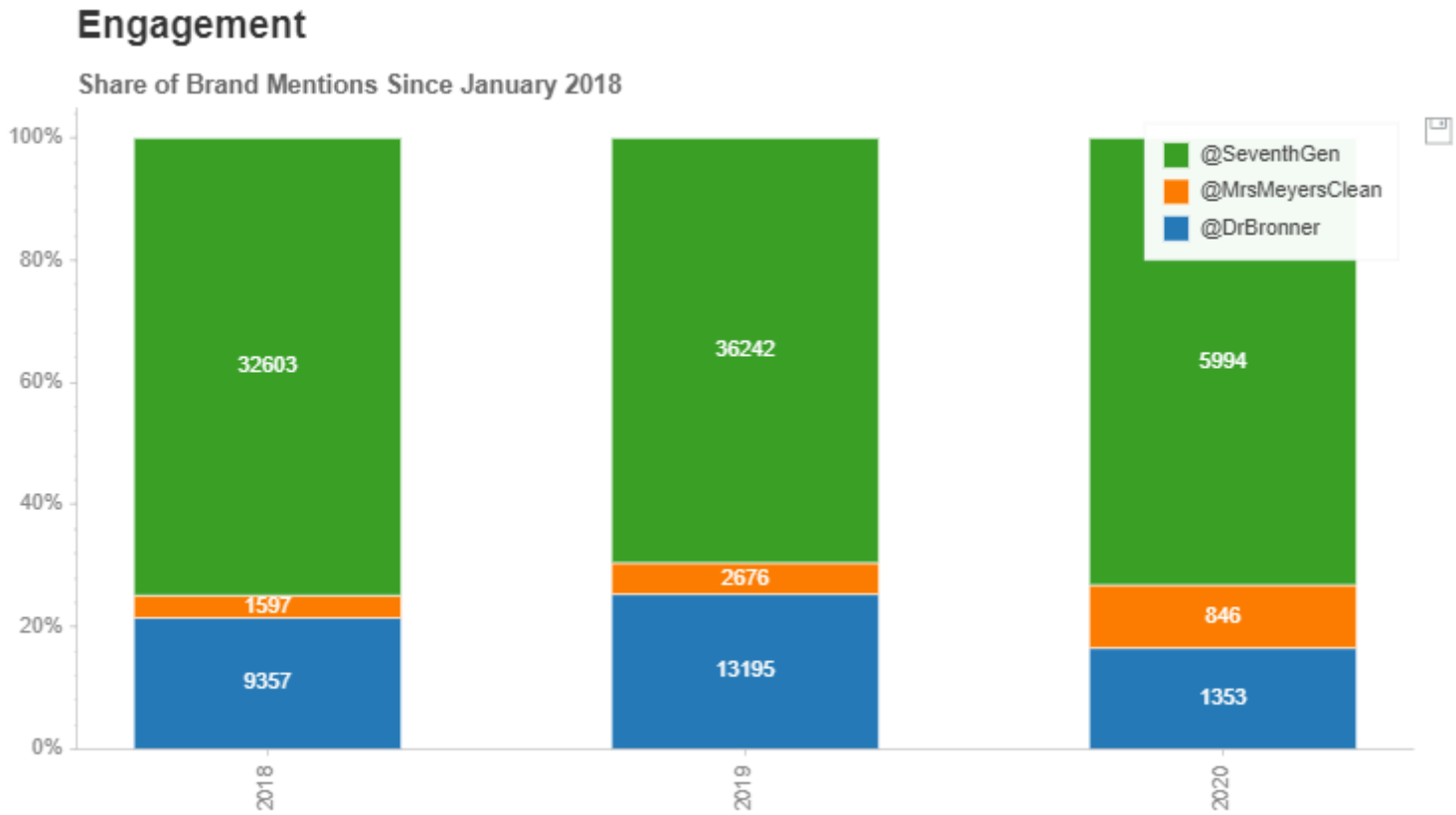
CONTENT

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

ENGAGEMENT BY BRAND BY YEAR

Insights

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



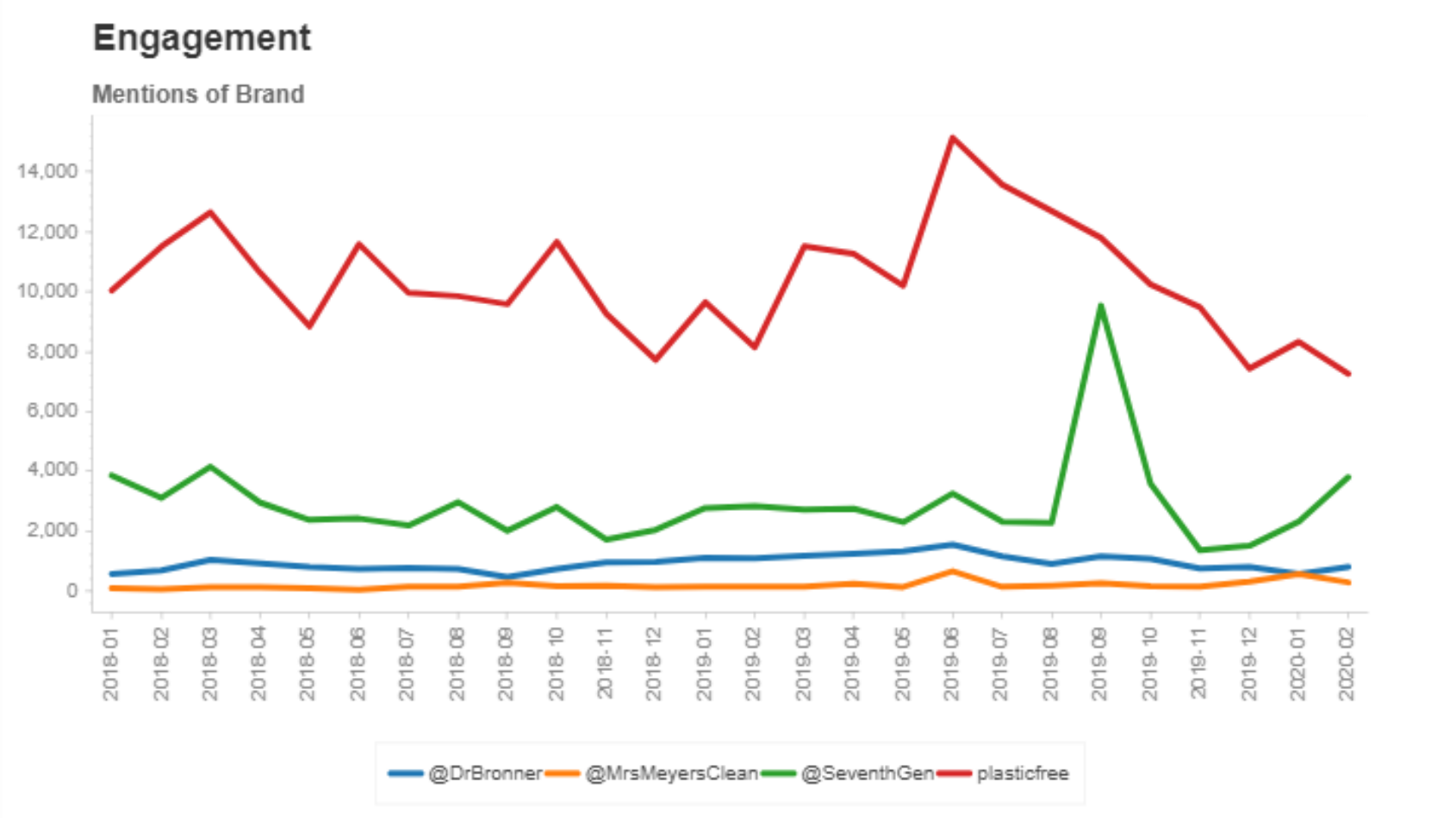
ENGAGEMENT BY MONTH

Insights

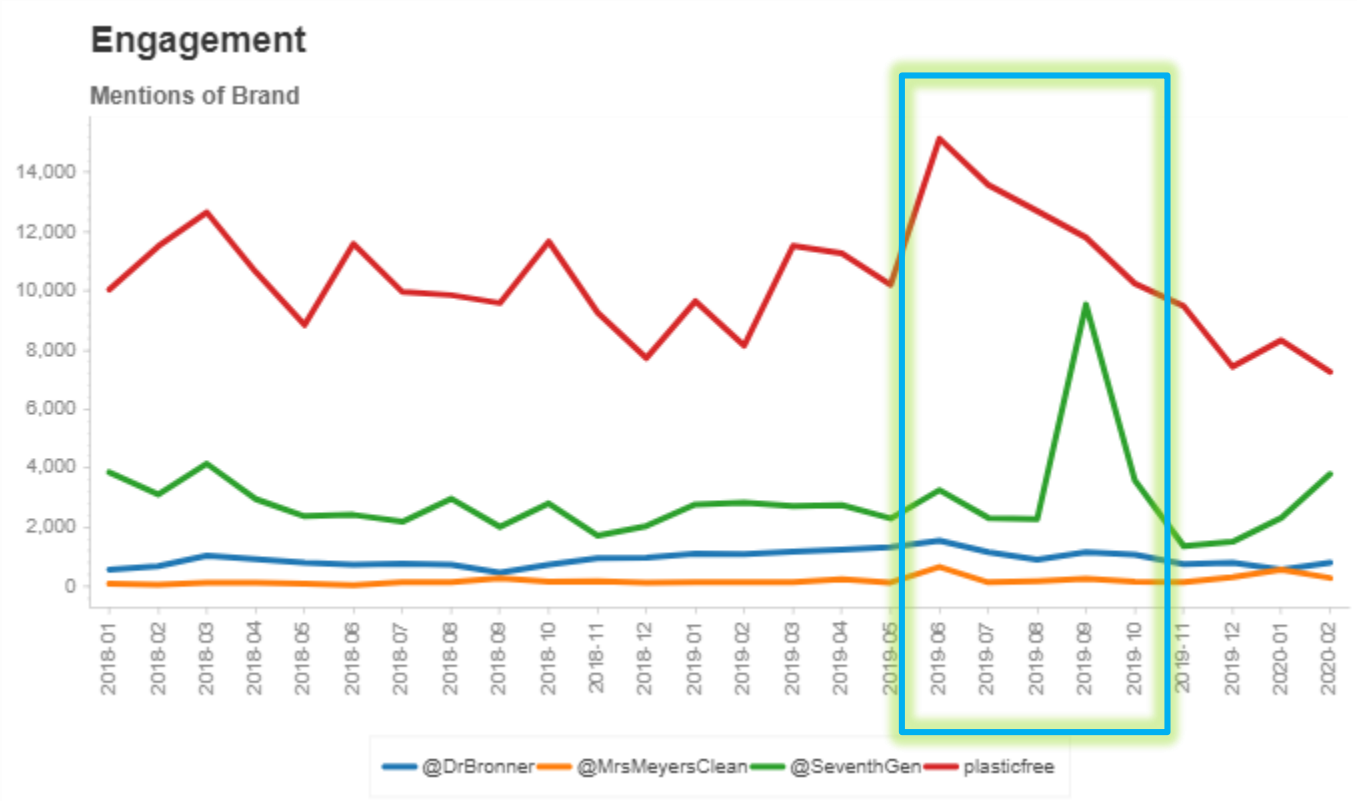
- #plasticfree averages 2.5X the engagement of @SeventhGen

- Generally plateaued mention counts across the board

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



#CLIMATESTRIKE

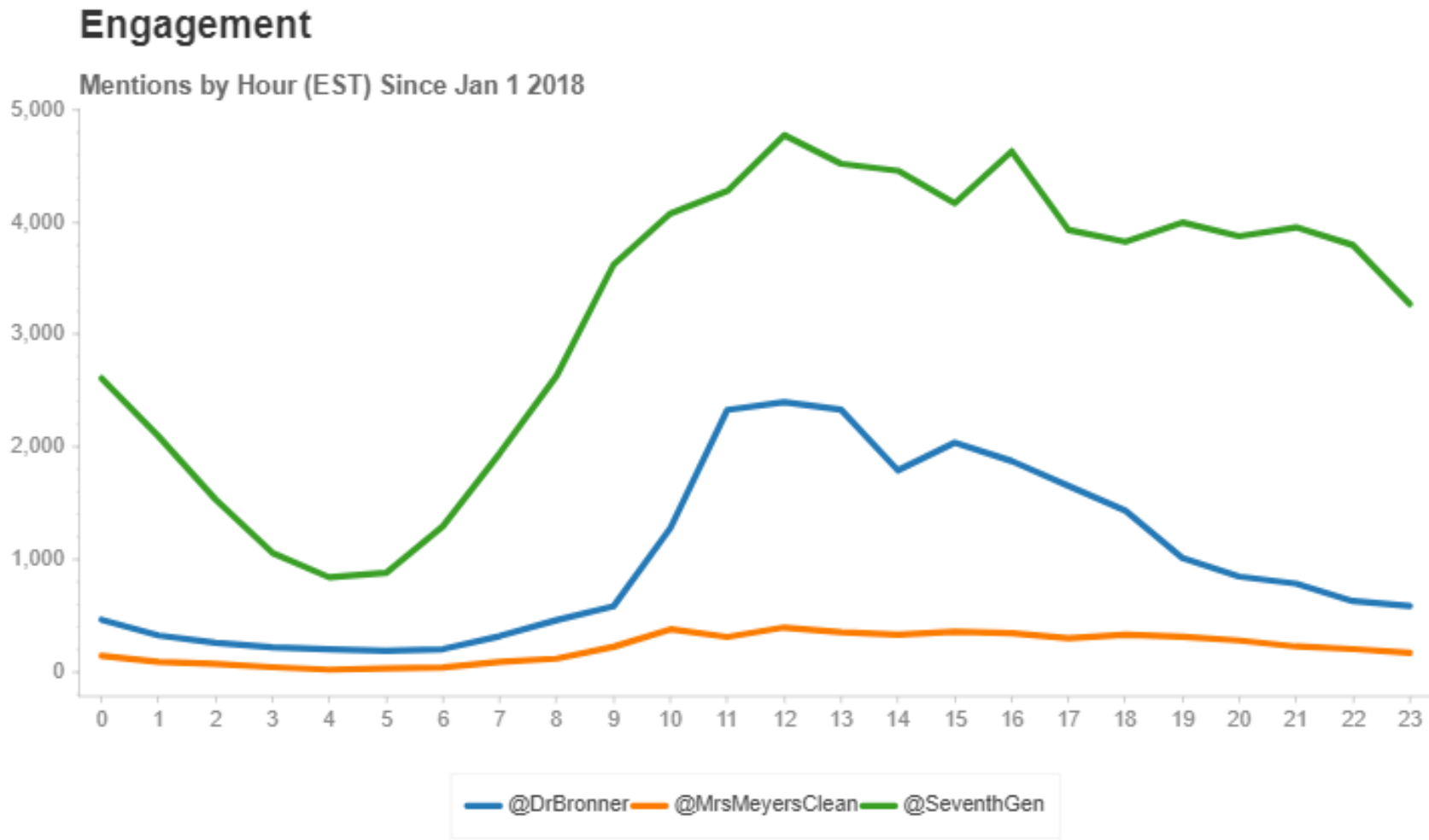


AGGREGATED 2 YEAR ENGAGEMENT BY HOUR

Insights

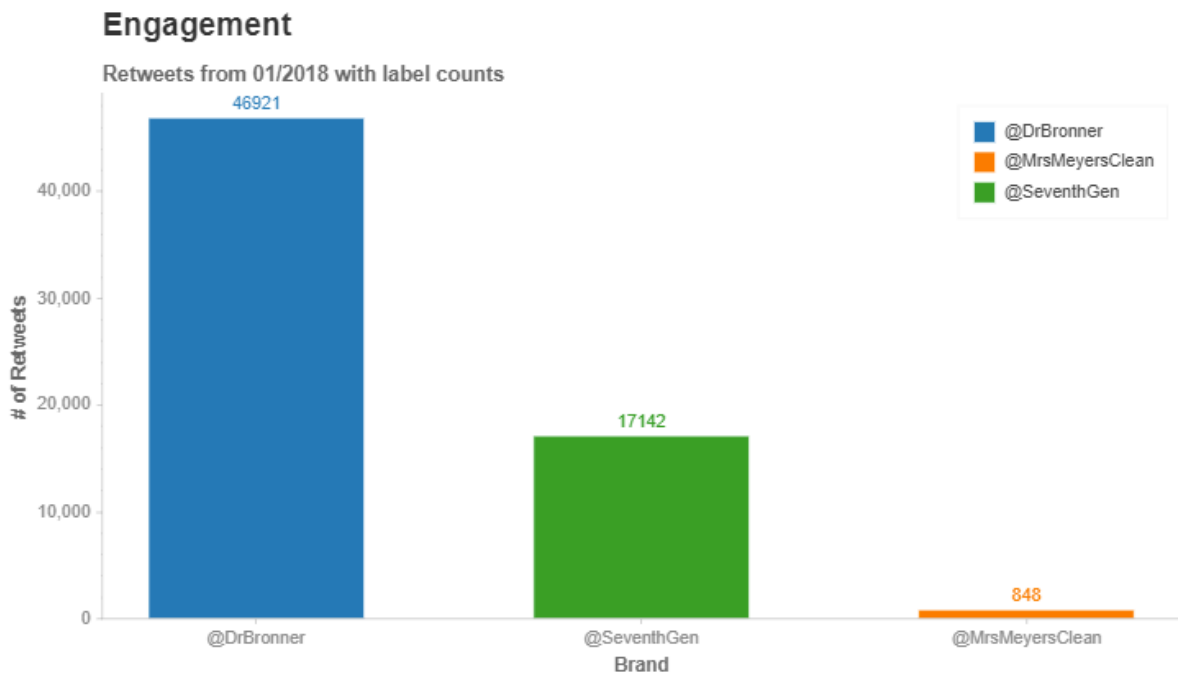
- Noticeable seasonality:
  - 6AM-12PM Ascent
  - 12AM-5AM Descent

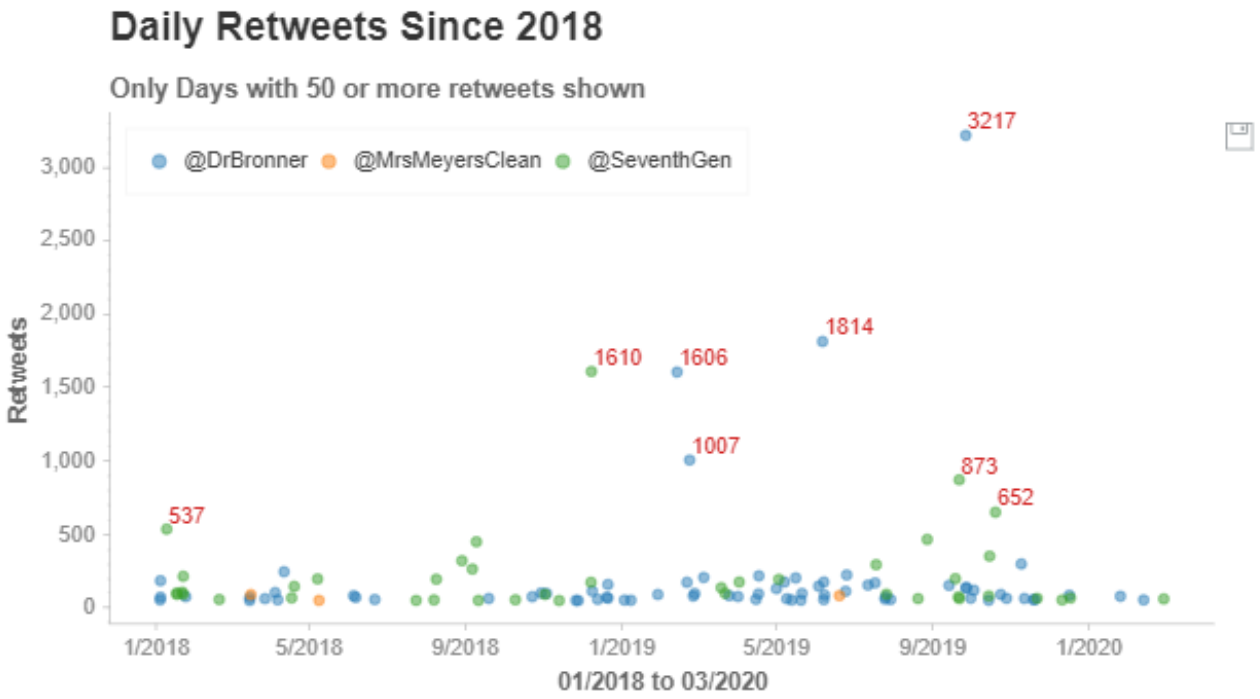
Business Advice:
Engage your future consumers when they are active 7AM-7PM



ANOTHER METRIC…

@DrBronner is punching above their weight with retweets compared to @SeventhGen

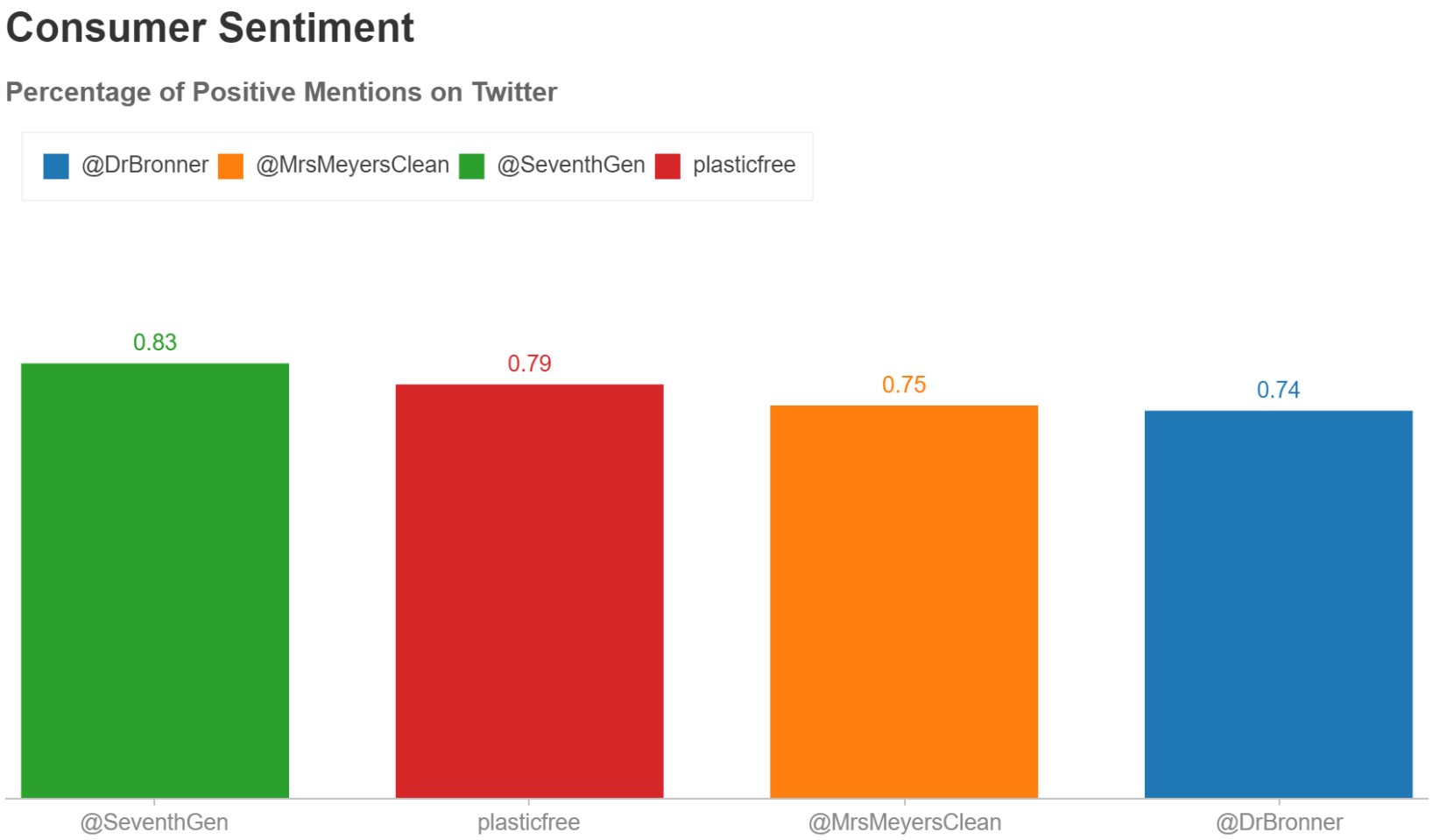




PERCENTAGE OF MENTIONS THAT ARE POSITIVE

Insights

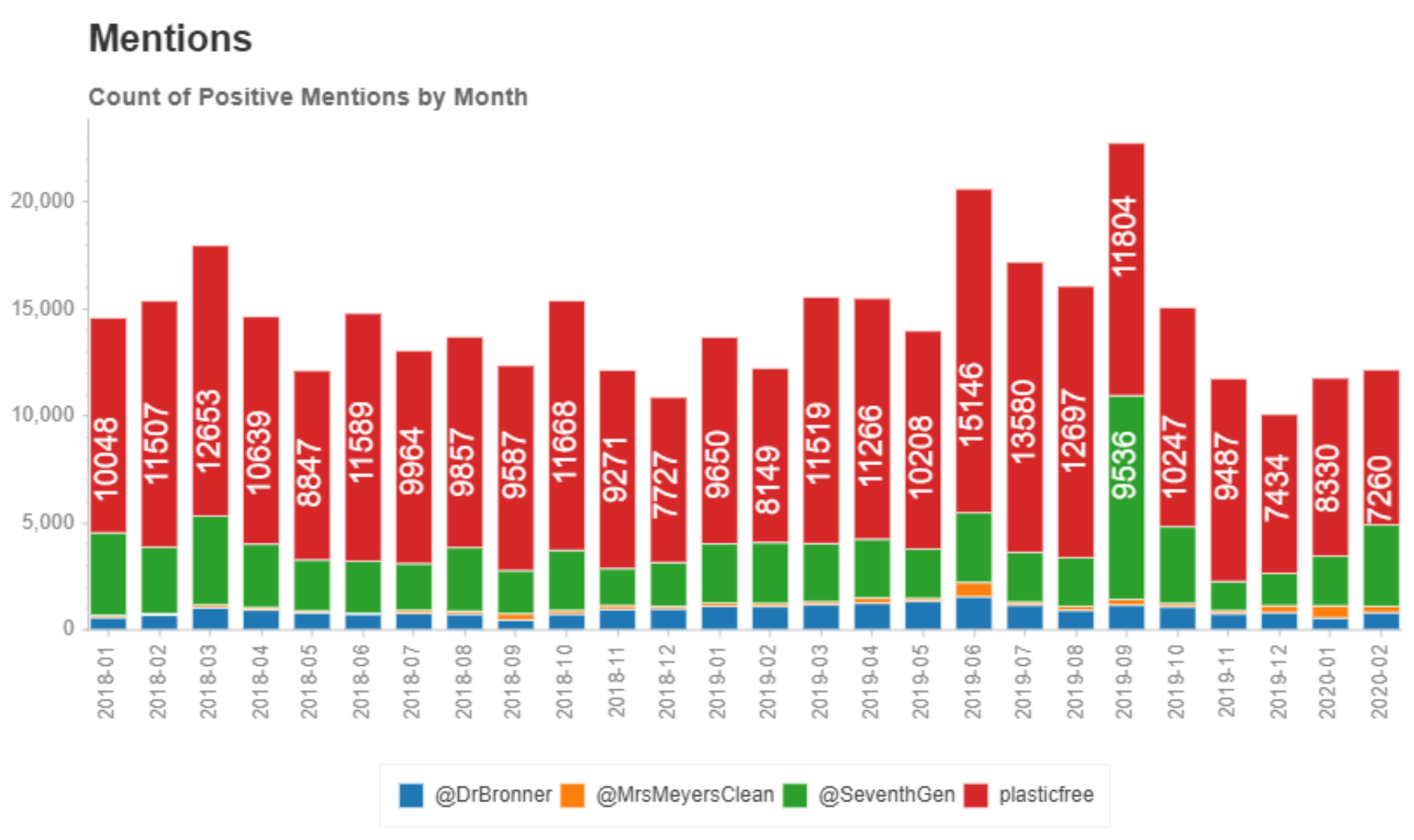
- Most mentions are classified as positive for all labels
- @SeventhGen receives the highest positive sentiment relative to mentions



LARGEST AMOUNT OF POSITIVE SENTIMENT

Insights

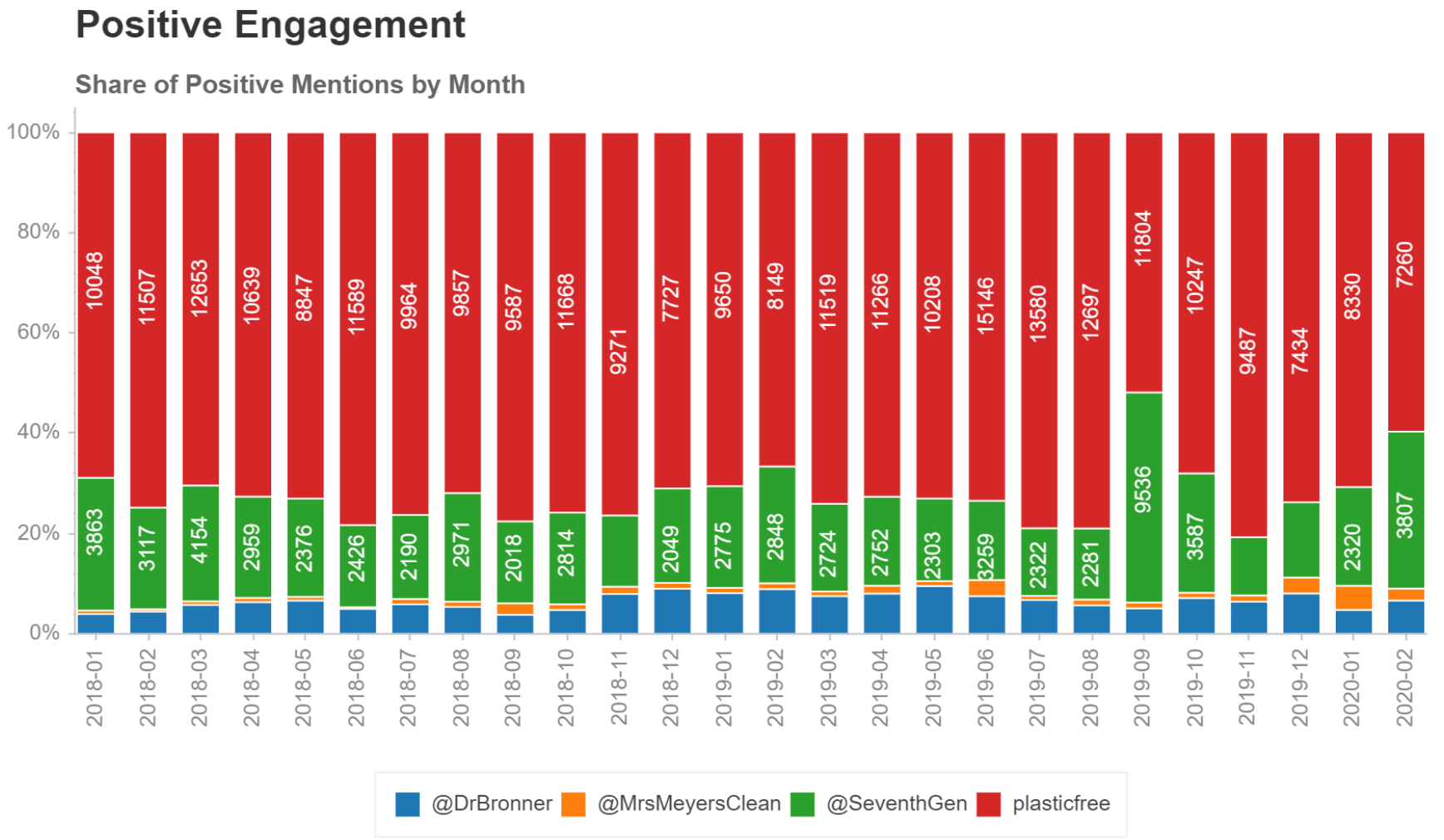
- #plasticfree generally outperforms all other labels combined



SHARE OF POSITIVE ENGAGEMENT

Insights

- #plasticfree leads positive engagement by both share and count



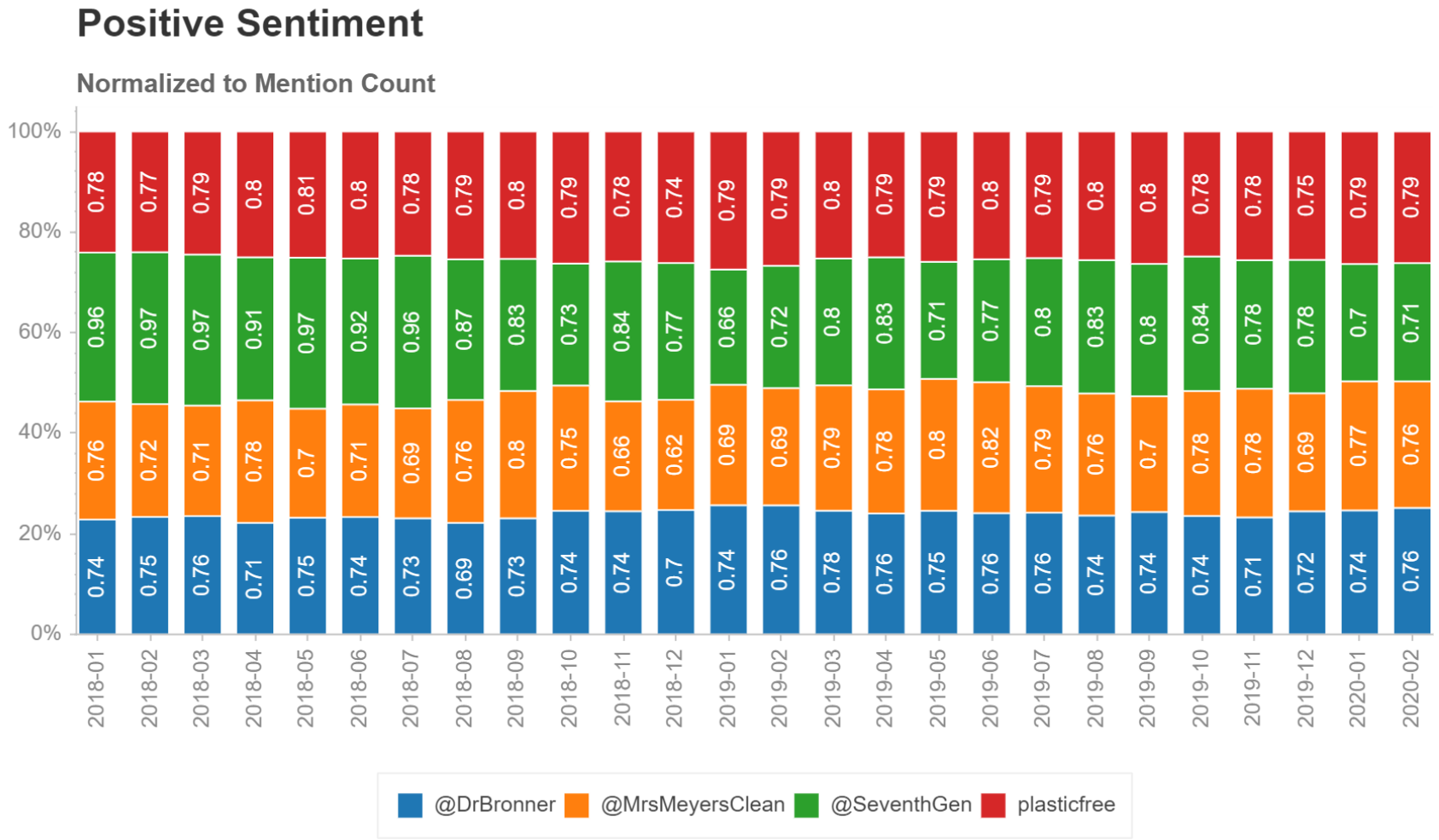
PROPORTIONAL RELATIVE SENTIMENT

Insights

- Each of the labels are proportionately positive given amount of engagement

Business Advice:

While high engagement counts are important, it alone does not supersede other factors such as price, convenience, quality, etc.



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 brands
2. There is consistent evidence on the importance of environmental sustainability leading to high consumer sentiment
3. @SeventhGen is the consistent leader of the three on Twitter (Emulate their engagement style)
4. The SVM NLP Classification model scores best on Sensitivity (True Positive Rate)

Recommendations:

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
4. Further analyze other metrics on additional platforms to validate or adjust these findings.