Identifying Potential Customers using User Affinity towards products

Pradhyumna Rao
University of Rochester
pradhyumna.rao18@ur.rochester.edu

Ayush Singla
University of Rochester
asingla3@ur.rochester.edu

Abstract

As much as business owners would like to believe otherwise, a company's product or service cannot meet every customer's needs. Understanding your ideal customers, or the customers who are most likely to profit from and purchase your product or service, is the key to effective sales if you are marketing a product or service. This is where we try to create a model to help businesses launching new products to identify customers on social media and convert them to loyal customers.

Objective: This project provides a unique approach to identify ideal customers, or the customers who are most likely to profit from and purchase a product or service, and is the key to effective sales if you are marketing a product or service. The potential target market consists of the customers who fit into this category and are most likely to profit from and purchase what the company is offering.

Scope: We would be using Twitter social media platform to understand user tweets and gather information on their sentiments towards various products and map this information to product categories and industry type. Our aim is to determine the degree of the consumer's affinity towards a product and identify the right set of users to be targeted by the company.

Method: The targeted products have not yet been released, and the buzz has just recently been developed. To see this, we took a year's worth of tweets from January to August and retweets from September to December to validate the users we had taken out. Through further application of our model to analyze user sentiments, which will enable us to categorize people into potential customers, this validation offers us confidence in the users. We look at the various possibilities and categorize them into 4 groups after comparing user sentiments in the two periods. Positive and negative sentiments are examined. This provides us with our ultimate target market.

Keywords **: Sentiments, Data Analysis, Cluster, Business**

I. INTRODUCTION

For businesses of all sizes, social media is an effective platform to connect with prospects and clients.Effective social media analytics can help your company achieve extraordinary success by generating leads, acquisitions and eventually loyal brand supporters.

For the scope of our project, We will be using Twitter platform to understand user tweets and gather information on their sentiments towards various products and map this information to product categories and industry type. Our aim is to determine the degree of the consumer's affinity towards a product and identify the right set of users to be targeted by the company.

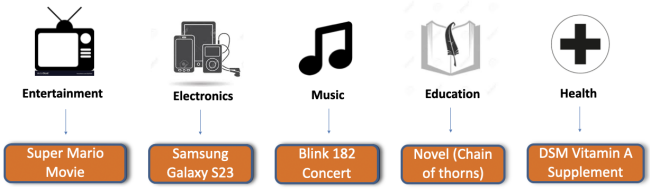

Fig 1: Different product of Industries

For each industry we have selected a product which is “Yet to be launched” in 2023. Then we looked at the Tweets of the previous year to get the sentiments of the products.

II. Literature Review

S.No	Domain	Paul B Chau et al	Sarvesh et al	S.-T. Yang et al	Longinos et al
1	Business Opportunities	✓	✗	✓	✓
2	Data Sample	✓	✓	✓	✓
3	Sentiments	✗	✓	✓	✗
4	Validation	✗	✗	✗	✓

Table 1: Comparative analysis of different research papers

Paul B. Chau et al. [1] focuses on whether the marketing researchers have demographic information on the current customers, or the general market population, or people with propensity to become customers He also presents a novel approach to the problem by exploiting the availability of a data sample from the general market population. Finally, he described an on-line lead management and delivery system that uses the mining approach described in this paper for insurance agents to obtain qualified customer leads.

Sarvesh Bhatnagarphan et al. [2] Discussed prior works on community detection and sentiment analysis. Worked majorly on Community based detection. Did analysis Of a closely related topic. provided insights into the graph structure of a community.

S.-T. Yang et al. [3] in their paper developed a personal business information management model with different modules to identify the potential customers and provide personalized knowledge. Based on performance of the previous projects, the existing customers are classified into critical and non-critical customers via the 80-20 principle. After classifying the existing customers into critical and non-critical customers, the module can filter out the key features and the corresponding confidence intervals to identify the potential customers.

Longinos Marin et al. [4] did further research into this subject by identifying the Role of these consumers in Brand Extension. They ran a regression to test the effect of additional information, identification, and their interaction on purchase intention.

III. Learning Methods

A. Data Extraction and Analysis

Extracting the data from twitter is the first step in our method. We extract data using product related keywords from different parts of The United States to get a comprehensive set of users and validate them by setting a cut-off date to further analyze the user sentiments and look at the polarity shift from one period to another.

In order to get polarity scores based on user tweets, we have used Vader polarity that assigns a score for each sentiment for every tweet and creates a compound score using the generated values.

Tweet Date	Tweet	UserName	Location	Scores	Compound
11/29/2022	I actually WANT THIS SEE THIS MOVIE! God damn, good job Nintendo and illumination!!! The Super Mario Movie is gonna be a HIT!	MoonNonaOG	LA	{'neg': 0.082, 'neu': 0.549, 'pos': 0.369, 'compound': 0.8556}	0.8556
10/7/2022	The new super Mario movie looks surprisingly good!	1776Productions	LA	{'neg': 0.0, 'neu': 0.392, 'pos': 0.608, 'compound': 0.8516}	0.8516
10/6/2022	I now have faith in the super mario movie!!!!	64Goblins	LA	{'neg': 0.0, 'neu': 0.433, 'pos': 0.567, 'compound': 0.8346}	0.8346
11/29/2022	I forgot that the Super Mario movie was actually real.	gallonegro_91	Chicago	{'neg': 0.0, 'neu': 0.672, 'pos': 0.328, 'compound': 0.5994}	0.5994
10/13/2022	Truly insane watching folks get worked up over the Super Mario movie. A movie that not only appears to made for children, but exclusively for the dumbest children.	SokolAdam	Bronx	{'neg': 0.187, 'neu': 0.682, 'pos': 0.131, 'compound': -0.4404}	-0.4404
10/19/2022	The Motion Picture Association has given The Super Mario Bros. Movie a PG (Paul Gale) rating.	PaulGaleNetwork	LA	{'neg': 0.173, 'neu': 0.783, 'pos': 0.045, 'compound': -0.8422}	-0.8422
11/25/2022	The Super Mario universe is insane like one minute Bowser and Mario are trying to murder each other with fireballs over a hostage princess	jpmark90	Austin	{'neg': 0.155, 'neu': 0.672, 'pos': 0.172, 'compound': -0.0516}	-0.0516

Fig 2: Different product of Industries

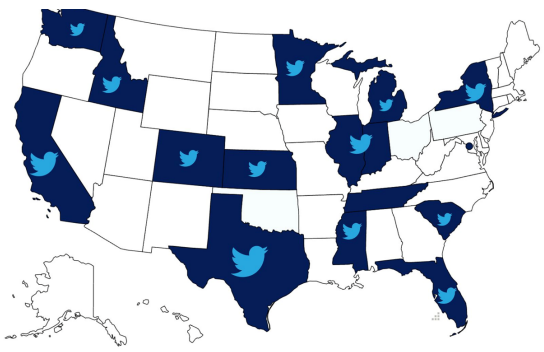

Fig 3: Volume of Data Extracted

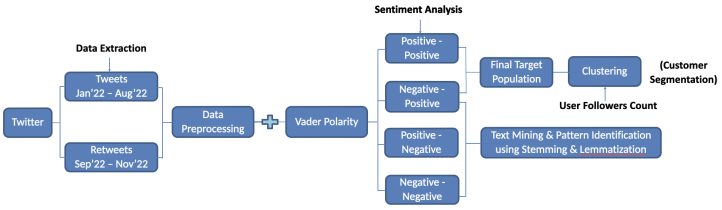

Fig 4: Project Flow

In order to see how user sentiments shift over time and to examine the elements that influence them, we devised a sentimental analysis and separated it into 4 categories. Having said that, the two key categories for our project are

users with negative sentiments that turn into positive sentiments and users with positive sentiments that remain positive. This helps us identify the target population for the

product that shows positive polarity as we approach the launch of the product. [5]

Polarity				
Jan'22 - Aug'22	Sep'22 - Nov'22	Final Category	Target Base Identified (#Users)	Tweets
Positive	Positive	PP	12,372	93,228
Positive	Negative	PN	7,040	45,201
Negative	Positive	NP	6,330	51,550
Negative	Negative	NN	4,595	27,172

Fig 5: Different Categories of Sentiments

Target base and no of tweets is identified for each category and 18k users are identified which can be a potential target base

We further looked at sentiment trends across different locations and months.

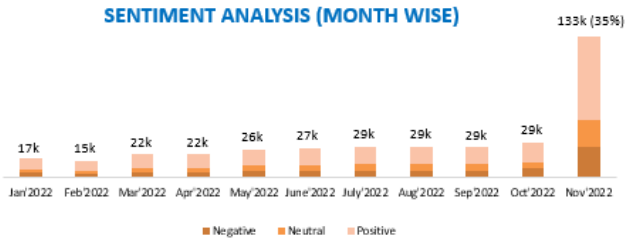

Fig 6: Month Wise Sentiment Analysis

Majority of the tweets were created during Nov '22 leading to a lower target base while the distribution of sentiments across different months has been steady with 60% positive sentiments.

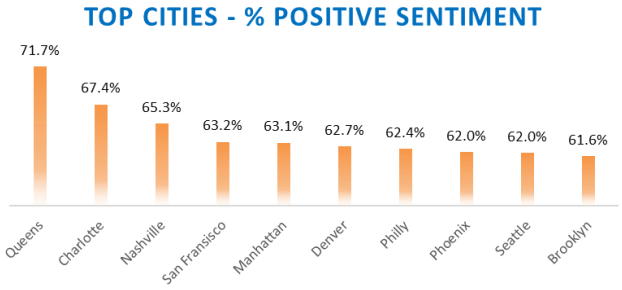

Fig 7: City Wise Positive Sentiment Analysis

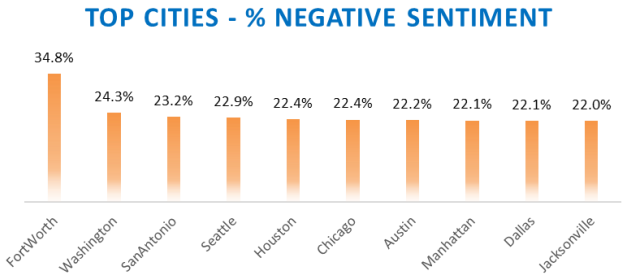

Fig 8: City Wise Negative Sentiment Analysis

While we observed the opposite trend at city level where we found top 10 cities with respect to positive and negative sentiment %. 72% Users in Queens show positive polarity and 35% users in FortWorth show negative polarity. Seattle and Manhattan are in both the lists

B. Cluster Segmentation

We have extracted the user sentiments towards various products from Twitter, But this gives us a very large pool of users which may or may not be completely relevant to a company. The important aspect is to divide into groups so it is easier for businesses to target a specific group of users and provide personalized offers to these customer segments based on their behavior. In order to accomplish this task, we have used K-means clustering as it is a very efficient clustering approach to segment data based on numerical attributes.

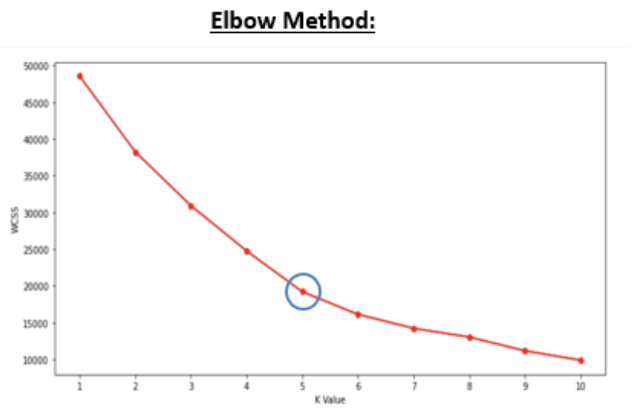

Fig 9: Number of Cluster - Elbow Method

Finding the ideal number of clusters to divide the data into is a critical stage in any unsupervised technique. We count the number of clusters needed to classify our data using the Elbow Method. We obtain k=5, which indicates that there will be 5 clusters into which the data will be divided.

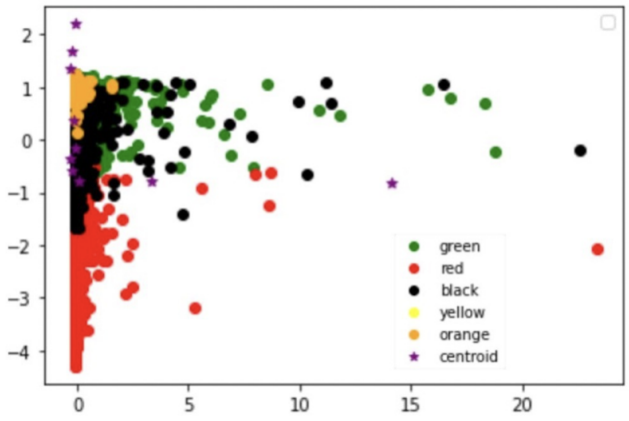

Fig 10: Cluster Analysis

We further segregate clusters into groups of labels based on similarity in behavior with each label having a different set of users that have different impact factors and can be targeted in a unique way.

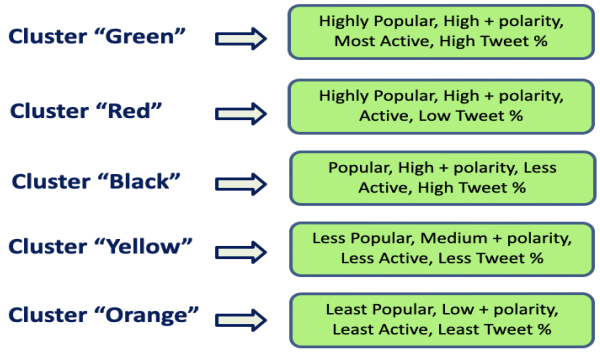

Fig 11: Different Labels of Cluster

IV. CODE SNIPPETS

When working on data science projects, an Exploratory Data Analysis (EDA) is essential. Knowing the data inside and out can help you make better decisions about which features, algorithms, and hyperparameters to use. The categorizing of data according to their type is an important part of the EDA process. That is where we extract the data and perform only those analyses which we require.

```
product='mario'
location='Indianapolis'
loc = '39.791000,-86.148003, 100km'
fifa_Ind = pd.DataFrame(itertools.islice(sntwitter.TwitterSearchScraper(
    'lang:en since:2022-01-01 until:2022-11-20 (fifa OR football) geocode:"{}"'.format(loc)).get_items(), 10000))
fifa_Ind['Tweet_Created_Date'] = pd.to_datetime(fifa_Ind['date']).dt.date
fifa_Ind['Location'] ='{}'.format(location)
fifa_Ind_v1=fifa_Ind[['date','Tweet_Created_Date','content','username','Location']] # picking the imp ones
fifa_Ind_v1.columns=['Date','Tweet_Created_Date','Tweet','UserName','Location'] # renaming columns
fifa_Ind_v1.tail(5)
```

Fig 12: Extract Tweets using Snscrape

```
import time
f7={}
d7={}
usernames=sms_final.UserName.values[0:1000]
for j in usernames:
  try:
    user = api.get_user(j)
    #usercreatedat[j]=user.created_at
    f7[j]=user.followers_count
    d7[j]=user.description
  except:
    pass
#print(usercreatedat)
print(len(f7))
print(len(d7))
```

Fig 13: Extract User Profile using Twitter API

```
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer

sid = SentimentIntensityAnalyzer()
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
```

```
fifa_full['scores'] = fifa_full['Tweet'].apply(lambda review: sid.polarity_scores(review))
fifa_full['compound']  = fifa_full['scores'].apply(lambda score_dict: score_dict['compound'])
fifa_full.head()
```

Fig 14: Vader Polarity

```
[ ] from sklearn.preprocessing import StandardScaler
    from sklearn import preprocessing
    import scipy.stats as stats

[ ] scaled_df = stats.zscore(data)

[ ] scaled_df

[ ] X = scaled_df[["Tweet_perc","days","Polarity","Followers"]]

[ ] #Importing KMeans from sklearn
    from sklearn.cluster import KMeans

    wcss=[]
    for i in range(1,11):
        km=KMeans(n_clusters=i)
        km.fit(X)
        wcss.append(km.inertia_)
```

Fig 15: Z-Score normalization for clustering

V. RESULTS

A. *Identified Target Base and Dollar Impact*

Product	Expected Product Price	Target Base	Conversion Rate	Customers Acquired	Dollar Impact
Samsung Galaxy S23	$800	5,419	10%	542	$434K
Blink 182 Concert	$200	12,162	10%	1216	$243K
Super Mario Movie	$20	18,702	10%	1870	$37K
DSM Vitamin A Supplement	$25	4,839	10%	484	$12K
Novel (Chain Of Thorns)	$20	5,294	10%	529	$11K

*Customer Acquired = Target Base*Conversion Rate*
*Revenue = Customers Acquired*Expected Product Price*

Table 2: Impact Analysis

Based on the Target Base identified we can convert this into dollar impact in terms of revenue generated. Taking into consideration a reasonable conversion rate (say 10%), we can calculate the revenue a company can generate as they acquire these customers.

Therefore, we have a validated method to help business and marketing teams to create an optimal strategy to look for potential customers.

B. Reasons for change in user sentiments

We further look into the reasons on why the user sentiments changed over a period of time by using NLP (Stemming and Lemmatization in NLTK) [6]. We successfully observe trends in which any event for the product drastically changes the mood of the user. Example : A movie trailer changes the sentiment from negative to positive or a leak, buzz creates change in mood.

Fig 16: Positive Word Cloud

Keywords like ‘trailer’, ‘incredible’ and ‘leak’ clearly show the sentiment of users changed post trailer launch of movie that in turn led to an overall positive impact.

Fig 17: Negative Word Cloud

While there are some users who changed their sentiments from positive to negative after learning Chris Pratt is giving voice for Mario character.

These insights can be extremely fruitful for the business as this gives deep dive into user thoughts and expectations. Accordingly, business can include these recommendations and insights and make better and informed decisions for their upcoming products.

In addition to this, it is very interesting to understand what kind of users are we targeting for a product and that is where we looked at user profiles and extracted common keywords to identify the most common user profiles

Fig 18: Identified User Profiles

As expected, most of the positive sentiments are shown by father, husband and people from the entertainment industry. Interestingly, retired military men and health content people are excited about the movie release next year

C. Product Level Insights

Now that we have performed sentiment analysis for each product, clustered the users into various groups and further looked at deep dive at user profiles and shift in sentiment, we will now focus on product level insights and look at top cities with respect to target base identified and average positive polarity shown by these users. New York , Los Angeles and Chicago are the top 3 cities when it comes to final target base which makes sense also since they have the higher population as compared to other cities

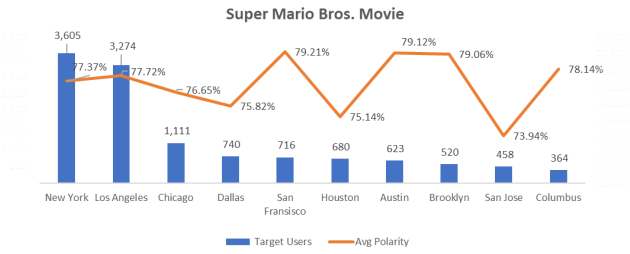

Fig 19: Super Mario Bros. Movie

San Francisco, Austin & Brooklyn show 79% positive polarity which is the highest among these 10 cities.

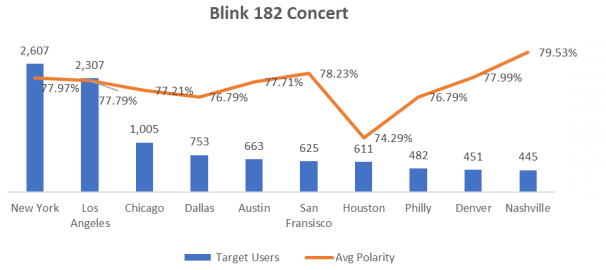

Fig 20: Blink 182 Concert

Users at Nashville seems to be most excited about the Blink 182 Concert to be held next year with average positive polarity of 80%

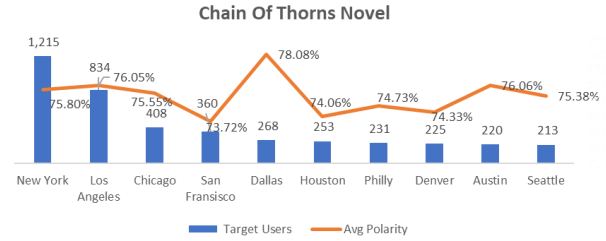

Fig 21: Chain of Thorns

Among the top 10 cities, Dallas inspite of having a low target base shows highest average positive polarity close to 78%. Most probably, users in Dallas have read and liked the earlier two novels by the same author and awaiting the launch of the third part

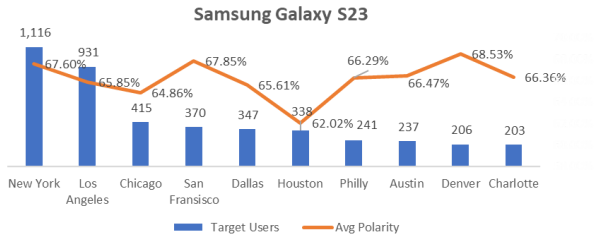

Fig 22: Samsung Galaxy S23

Interestingly, none of the top 10 cities has average positive polarity greater than 70%. This can be attributed to the fierce competition in the market when it comes to smartphones.

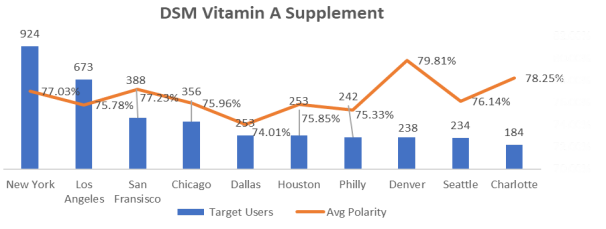

Fig 23: DSM Vitamin A Supplement

Users at Denver are eagerly awaiting the launch of the health supplement as average positive polarity is close to 80%.

VI. Conclusion

We can draw the conclusion that the data we extract based on location and product search provides us with a target base sufficient for analysis and insight extraction. The goal is to turn these users into potential clients by using clustering to divide them into Labels. We can also successfully track changes in user sentiment across many categories and examine the underlying causes.

We can then look at the Time analysis of how the interest changes of consumers with reference to each attribute and successfully say that we have gained adequate insight For Identifying Potential Customers For Any Business once we derive product level insights.

VII. Further Research

With all this in mind, there are many other ways we can look at our problem and have meaningful additions in order to further refine our target population and help businesses make better data driven decisions.

1. A company's actual sales and user demographics data can be extremely useful in further validating our method and refine the target population identified. For a company, it makes more sense to invest in potential users that show similar behavior as their customers.
2. Predict user's willingness to buy a product which can be an important factor in deciding whether or not to invest in that user
3. Extract data from other social media platforms like Amazon marketplace to increase the volume of data and thus more target customers for a product
4. Predict whether a product is an emerging product by analyzing user sentiments and topic modeling and observe the market trends.

References

1. Chou, Paul & Grossman, Edna & Gunopulos, Dimitrios & Kamesam, Pasumarti. (2000). Identifying prospective customers. 447-456. 10.1145/347090.347183.

2. Bhatnagar, S., Choubey, N. Making sense of tweets using sentiment analysis on closely related topics. *Soc. Netw. Anal. Min.* 11, 44 (2021). https://doi.org/10.1007/s13278-021-00752-0

3. Identification and Personalized Knowledge Provision of TSPs, IFAC Proceedings Volumes, Volume 42, Issue 4, 2009

4. Marin L, Ruiz De Maya S, Rubio A. The Role of Identification in Consumers' Evaluations of Brand

Extensions. Front Psychol. 2018 Dec 17;9:2582. doi: 10.3389/fpsyg.2018.02582. PMID: 30618996.

5. Namugera, F., Wesonga, R. & Jehopio, P. Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. *Comput Soc Netw* 6, 3 (2019).

6. Chou, Paul & Grossman, Edna & Gunopulos, Dimitrios & Kamesam, Pasumarti. (2000). Identifying prospective customers. 447-456. 10.1145/347090.347183.

7. Musumali, Benjamin. (2019). An Analysis of why customers are so important and how marketers go about understanding their decisions.

Bibliography

1. https://sloanreview.mit.edu/article/how-to-identify-the-best-customers-for-your-business/

2. https://edwardlowe.org/how-to-identify-a-target-market-and-prepare-a-customer-profile/

3. https://www.business.qld.gov.au/running-business/marketing-sales/market-customer-research/plan-conduct