Reseapch PReneurs

Researchpreneurs- Data Analytics Project

Final Presentation Group 5- 13.07.23

Van Anh Pham, Lance Pangan, Muhammed Ikbal Sevinc, Anna-Lena Rau

Research PRneurs

Our team

Anh Pham

Technical Lead

Ikbal Sevinc

Supervisor

Anna Rau

Documentation Lead

Lance Pangan

Project Manager

Table of contents

Research PRreneurs

Research PReneurs

01

MOTIVATION

Task & Generated Value

TASK

1. Analyze Researchpreneurs' LinkedIn data and advertising tool
   → target the appropriate users and boost registration rate of company's website

2. Analyze competitors text data on Twitter
   → identify most significant topics and provide recommendations for an enhanced content strategy.

Generated Value

- Understand most important topics, keywords etc. for the industry
- Increase search appearances
- Engage the target audience
- Provide a higher return on investment for the company‘s marketing effort

Research PReneurs

Research Preneurs

02

Data & Approach

Data Collection, Preprocessing & Topic Modeling

Used Data

LinkedIn Data

- Exported stats and analytics from LinkedIn
- Including company page analytics and follower demographics

Twitter Data

- Tweets from 6 most important competitors

Research PRneurs

TABLEAU-DASHBOARD

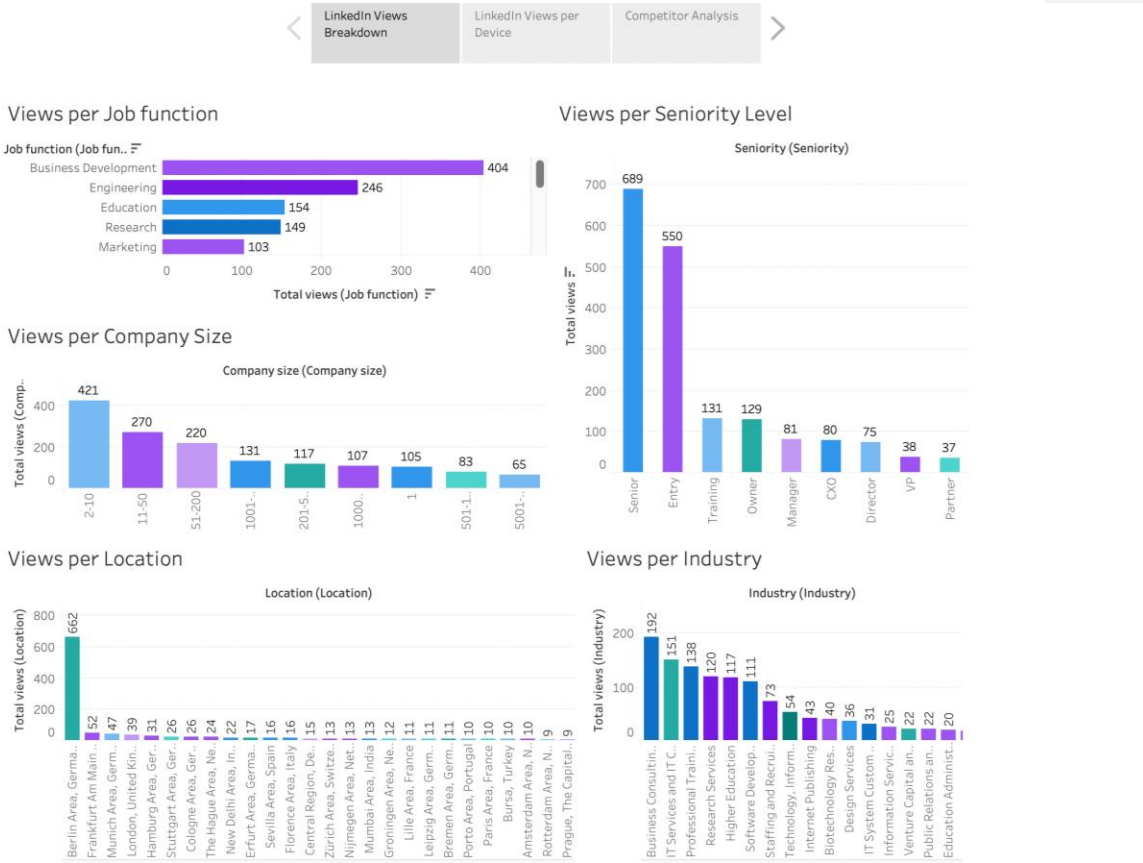
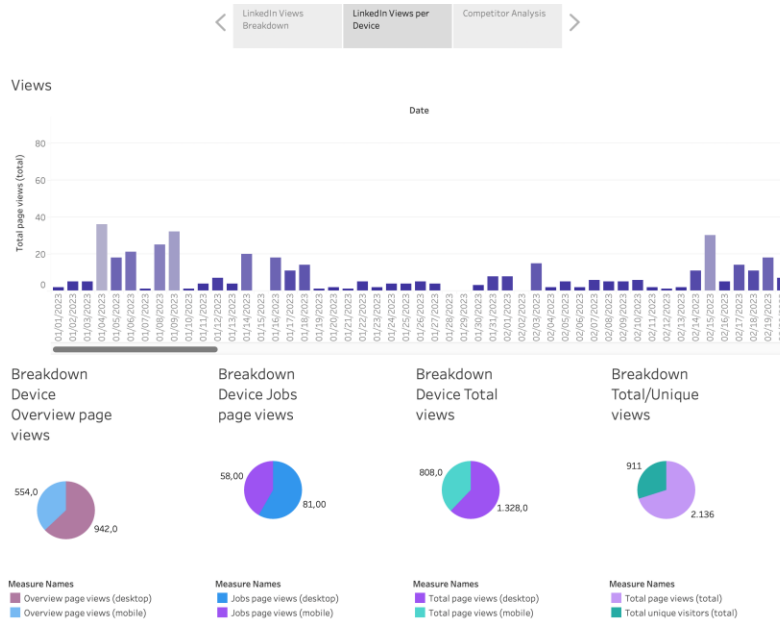


TABLEAU-DASHBOARD



Data Collection

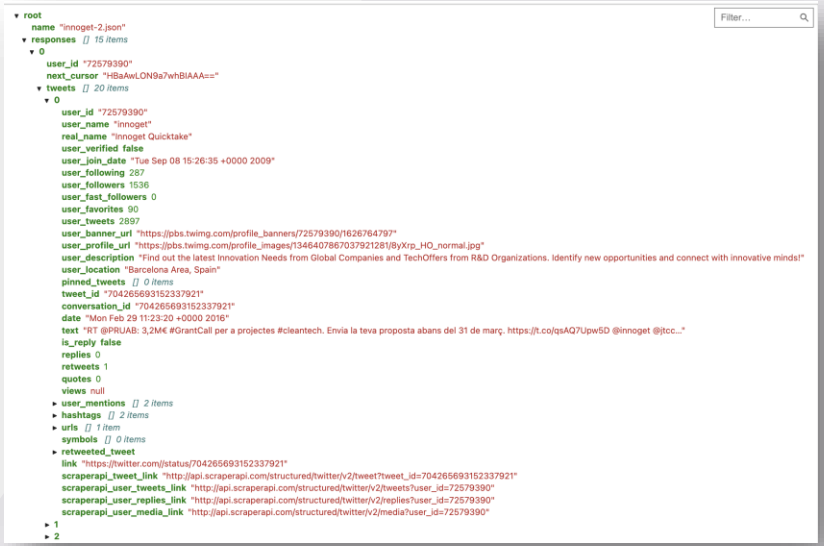
Data Scraping

Scrape all tweets of 6 competitors on Twitters

→ Total: 4984 tweets

Data Selection

Extract relevant data from the scraped data to use for analysis and modeling



Data Collection

	user_id	user_name	date	text	tweet_id	is_reply	replies	retweets	quotes
0	72579390	innoget	Mon Feb 29 11:23:20 +0000 2016	RT @PRUAB: 3,2M€ #GrantCall per a projectes #c...	704265693152337921	False	0	1	0
1	72579390	innoget	Fri Feb 26 11:35:21 +0000 2016	#Global #Biotech Reagents Market 2016 Industry...	703181551727570944	False	0	0	0
2	72579390	innoget	Thu Feb 25 10:15:16 +0000 2016	Tech Transfer Office in #Ohio #University help...	702799011187658752	False	0	0	0
3	72579390	innoget	Wed Feb 24 12:10:05 +0000 2016	What's your point regarding IP protection? Doe...	702465520109559808	False	0	0	0
4	72579390	innoget	Tue Feb 23 15:15:17 +0000 2016	New article about #Samsung and its investment ...	702149739492597761	False	0	0	0
...	...	...	...	...	...	...	...	...	...
4979	72579390	innoget	Tue May 17 20:25:37 +0000 2016	RT @harilaosv: A Hands-Off Approach to #OpenIn...	732668416016945153	False	0	2	0
4980	72579390	innoget	Wed Apr 27 08:04:02 +0000 2016	@innoget is attending the Open Innovation Summ...	725234035593744384	False	0	1	0
4981	72579390	innoget	Tue Mar 29 12:08:13 +0000 2016	RT @BIOFIT_EVENT: Welcome to #BioFIT2016 new s...	714786234791542785	False	0	2	0
4982	72579390	innoget	Mon Mar 21 08:25:54 +0000 2016	IoT Development Enters the Open Innovation Cra...	711831185161375744	False	0	0	0
4983	72579390	innoget	Thu Mar 10 08:11:22 +0000 2016	New article about 5 Intellectual Property Mist...	707841261294764032	False	0	0	0

4984 rows × 9 columns

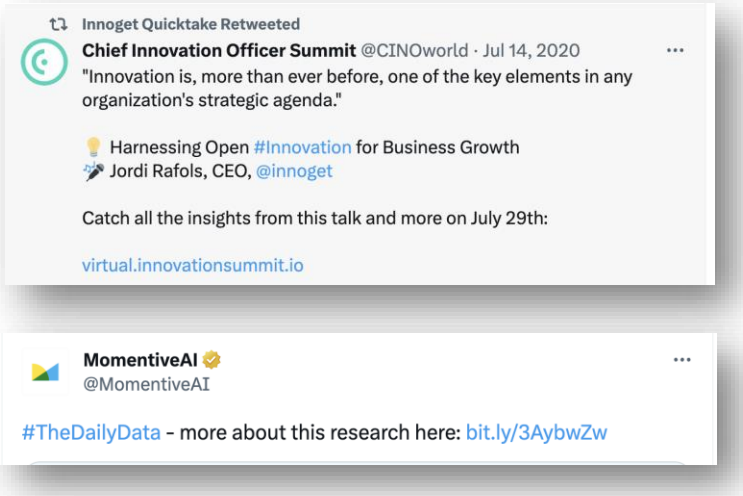
Data Cleaning

Retweet & Noise Tweet

- Retweets are duplicates of other users' tweets
- Noise tweets contain patterns that primarily serve as links to external sources

Technique:

Remove these tweets to prioritize content that directly contributes to topic modeling process



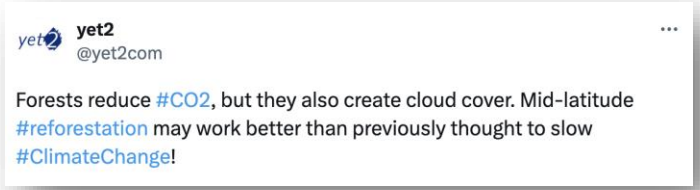
Data Cleaning

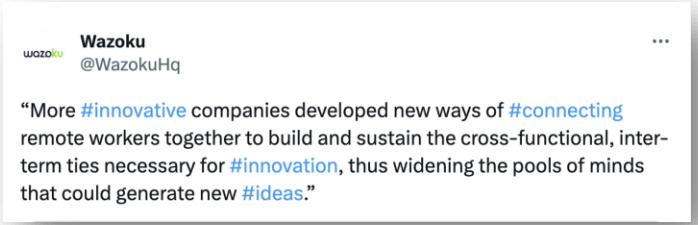
Hashtags:

- Hashtags are important and contribute to the topic analysis.

Technique:

- Remove "#" symbol
- Separate combined hashtags into individual words for better analysis.





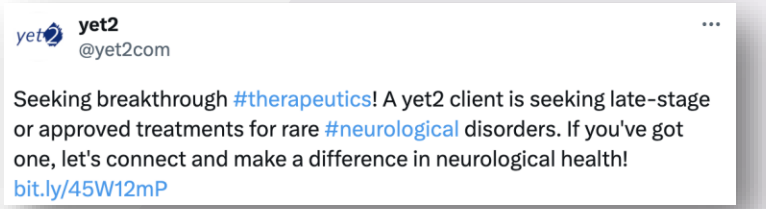
Data Cleaning

Uninformative Words

- Call-to-action words: Read more, more, here, sign up, …
- Company names
- Stop words
- Short words
- Emails, links, mentions
- Special characters: @, %, …
- Numbers

Technique:

Remove all of them



Research PRreneurs

LEMMATIZATION

Reduce words to their base or dictionary form to improve analysis accuracy and enhance information retrieval capabilities.

Final Preprocessed Data

Total remaining tweets = 4,131 tweets

Research PReneurs

Topic Modelling - Approach

Why TF-IDF & K-means?

- **Initial exploration**: a quick overview of potential topics or patterns in the data.
- **Small dataset and simple topics**: The dataset is not large or complex enough to significantly benefit from the more sophisticated methods such as LDA.
- **Interpretability**: easily interpretable clusters -> more helpful business people with limited data science understanding.

VECTORIZATION – TF-IDF

- Convert textual data into numerical vectors by assigning weights to each word based on its frequency in a document and rarity across the entire corpus.

To reduce the noise in data, rare and common words are removed:

- **Rare** words: appearing in **less than 5%** of the documents
- **Common** words: appearing in **more than 95%** of the documents

Result: 4,131 documents in the corpus and 25 words in the vocabulary

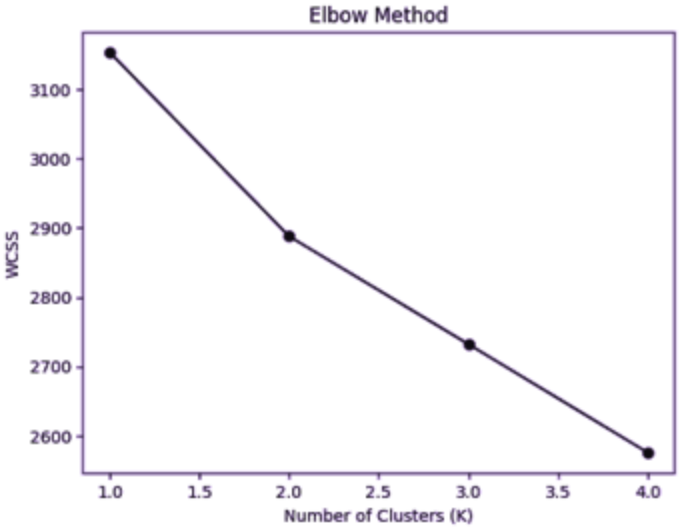
Topic Modelling

Elbow Methods

Main Idea: how the within-cluster sum of squares (WCSS) changes as the number of clusters increases.

- Add more clusters -> better separate the data points into distinct groups.
- However, after a certain point, we are no longer able to significantly improve the clustering by adding more clusters.
- The optimal number of clusters is the point at which the WCSS curve starts to bend or elbow.

We ran k-means clustering with k-values ranging from 1 to 5 -> elbow point at **k= 2**



Research PReneurs

Topic Modelling

K-means

Perform the K-means Clustering with k=2 to retrieve the cluster labels as well as common words within each cluster.

```
0 : innovation, open, webinar, technology, late, new, research
1 : challenge, new, technology, research, seek, solution, look
```

Topic: Innovation Exploration

Characterized by words related to exploring new ideas, developing new products or services, and finding new ways to solve problems

➔Cluster 0: Innovation Exploration

Research PReneurs

Topic: Challenges and Solutions

Characterized by words related to overcoming challenges, finding solutions to problems, and improving existing products or services

➔ Cluster 1: Challenges and Solutions

- Final Step: Assign the topic names to the original Tweets for evaluation

Check Cluster Names

	topic_name	text
4181	Innovation Exploration	"Because good ideas don't grow on trees" is the #weeklyquote. Don't wait any longer & make the leap to #OpenInnovation 3.0 #innovateSmarter https://t.co/YoosRtIMLW
990	Innovation Exploration	Open Innovation 2.0 #Conference 2015 announced in Espoo, Finland from July 6th to September 6th this year http://t.co/2NvNCxoDhv #OpenInnoEU
1314	Innovation Exploration	Want to learn how to apply our new Advanced Analytics feature to your innovation program? Watch our new webinar series 'Innovation Insights', learn more here: https://t.co/jF6WBVlNlm\n\n#Webinar #AdvancedAnalytics #Wazoku #IdeaManagement https://t.co/oMXfK2vTsd
2991	Innovation Exploration	Driving #innovation across your supply chain and across the entire value-chain should be paramount - Wazoku can help you to achieve your #sustainable innovation goals, faster and more cost effectively. #ThursdayThoughts https://t.co/fmZAk0fIp9 https://t.co/SdsFgXThOm
3902	Innovation Exploration	What defines the culture of an innovative organisation? Find out from our blog post: https://t.co/K4ZyBs8cqC #cultureofinnovation #cultureofchange #innovation https://t.co/4dDSj8ggN2

	topic_name	text
3074	Challenges and Solutions	Today is the day! At 4pm today we are hosting a panel discussion and live Q&A event with representatives from a range of European university technology transfer offices. \nhttps://t.co/dhM1A4c5SL https://t.co/fIcPiSXhwK
2620	Challenges and Solutions	Artificial neurons have the potential to help any number of illnesses from #dementia to heart disease and injury. Researchers @UniofBath in conjunction with @BristolUni, @UZH_en, and @AucklandUni have done just that. What else could it lead to?\nhttps://t.co/9qGiKyMYy0 https://t.co/DiEWgyL1TV
1812	Challenges and Solutions	A yet2 client is seeking new applications for #poultrymeat by-products\n\nhttps://t.co/XkkUCI04UQ https://t.co/m1wW3sHKYb
1142	Challenges and Solutions	In an exclusive interview with Simon Hill, CEO of Wazoku and Sarah Counts, COO of Wazoku they take us through the challenges and opportunities they faced in their mission to become a B Corp-certified company 🌍 https://t.co/WDCzLxWG8n \n\n#BCorpMonth #WeGoBeyond #Changetheworld
69	Challenges and Solutions	Planning to be in #NewYork on December 9th? Join the @FT Innovate 2015: Agility in an age of ambiguity - Innoget Blog http://t.co/B1cdCX1jd1

Research PReneurs

Research PReneurs

03

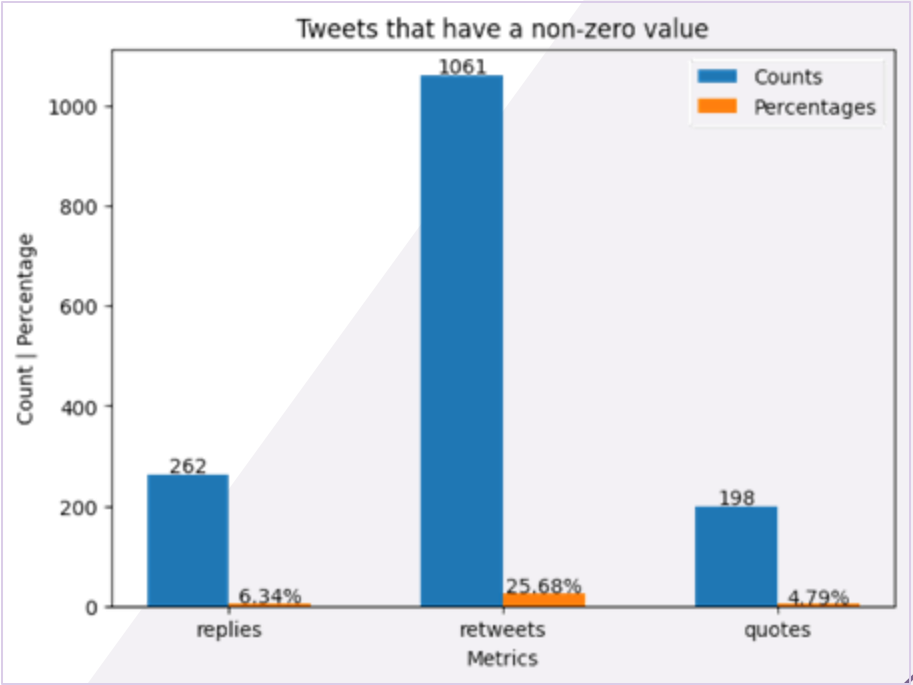
RESULTS

Evaluation& Conclusion

Engagement Rate

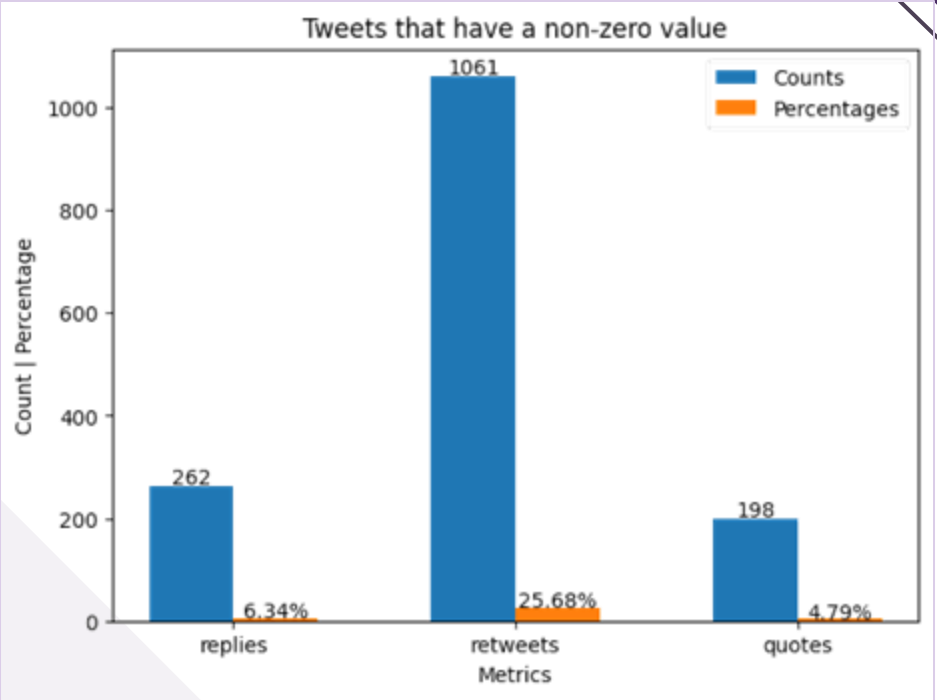
- Retweets are more prevalent compared to replies and quotes (25.68%)
  → Significant number of tweets have resonated with users and have been shared with a wider audience

➔ General engagement rate is low



Engagement Rate

- Majority of tweets in the dataset have not received replies (6.34% Reply-Rate)
  →Engagement level in terms of direct responses to tweets is relatively low
- Quotes are the least common interaction (4.79% Quote-Rate)
  →Users are less inclined to quote tweets in their own posts compared to retweeting or replying



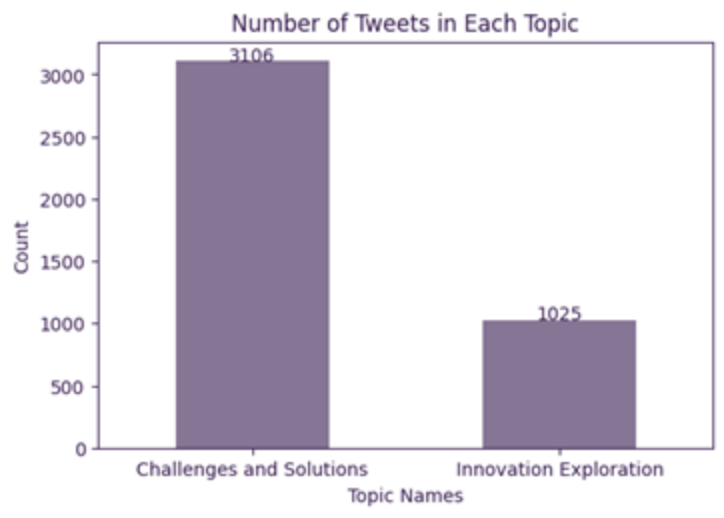
Research PReneurs

Engagement Rate per competitor

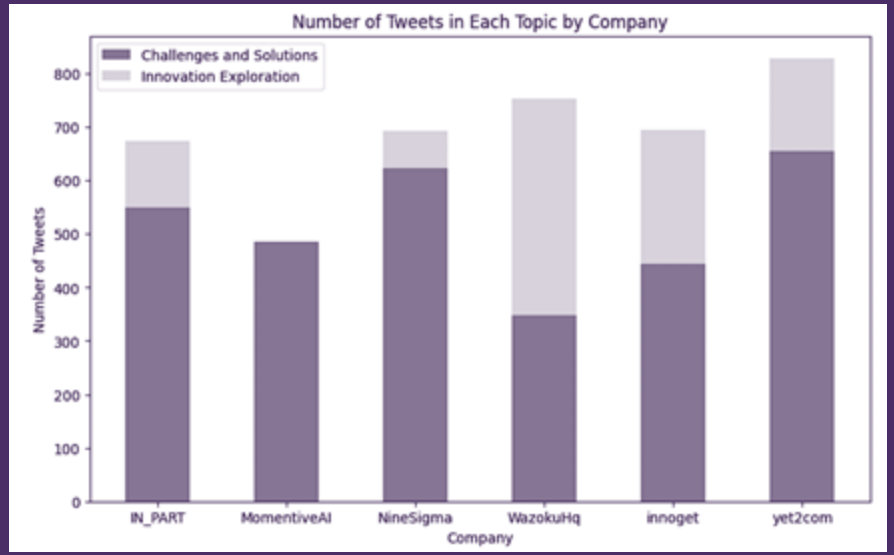
	Company	replies_count	replies_percentage	retweets_count	retweets_percentage	quotes_count	quotes_percentage
0	innoget	15	2.158273	139	20.000000	4	0.575540
1	IN_PART	33	4.896142	144	21.364985	41	6.083086
2	NineSigma	18	2.597403	276	39.826840	42	6.060606
3	WazokuHq	31	4.111406	176	23.342175	19	2.519894
4	MomentiveAI	150	30.864198	248	51.028807	82	16.872428
5	yet2com	15	1.809409	78	9.408926	10	1.206273

- MomentiveAI has the highest overall engagement rate, followed by WazokuHq and NineSigma
- IN_PART has the highest reply rate, but retweet and quote rates are lower than the average
- Innoget has the lowest overall engagement rate

Tweets per Topic



- Analyze how the topics relate to the overall engagement rate and identify any patterns or trends that emerge
- Challenge and Solutions Topic is dominating



Tweets Per Topic and Company

- Most of the companies focus on Challenges and Solution topic
- Only innoget and WazokuHq (>50%) have higher amount of tweets on Innovation Exploration

Research PReneurs

Engagement Rate per topic

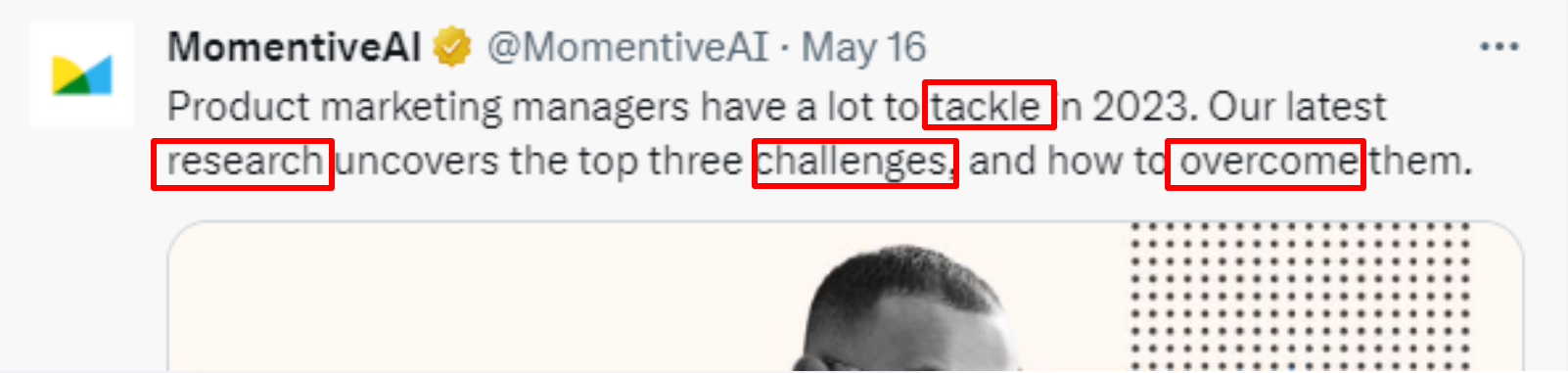
topic_name	Replies Count	Replies %	Retweets Count	Retweets %
Challenges and Solutions	234	7.53	818	26.34
Innovation Exploration	28	2.73	243	23.71

- Challenges and Solutions topic generates more engagement in terms of replies and retweets→ active discussions and sharing of content
- Innovation Exploration topic exhibits lower levels of direct interaction through replies, still garners considerable engagement through retweets.

Recommendation 1: Incorporate "Challenges and Solutions" Topic

Topic of "Challenges and Solutions" has shown significant engagement, with a higher count of tweets and retweets:

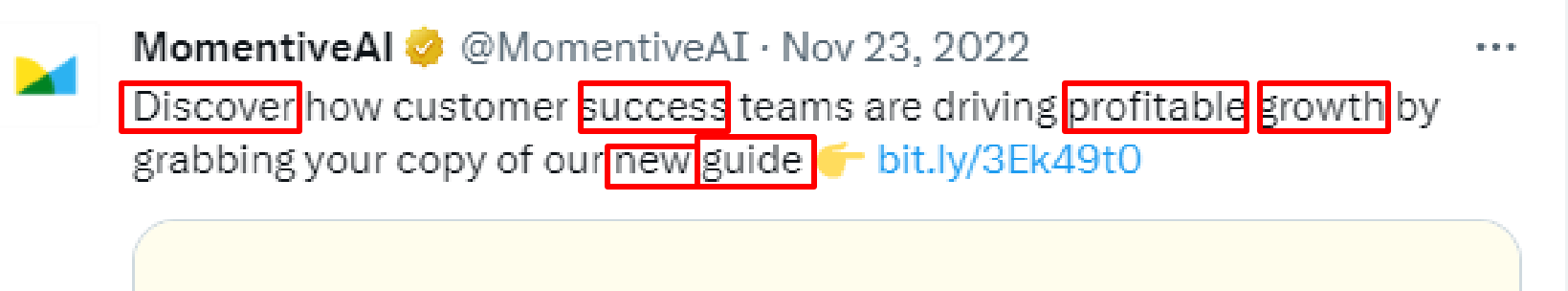
- Share practical solutions and insights to industry challenges



Research PReneurs

Recommendation 1: Incorporate "Challenges and Solutions" Topic

- Boost interaction by encouraging audience participation.
- Share success stories demonstrating effective problem-solving:

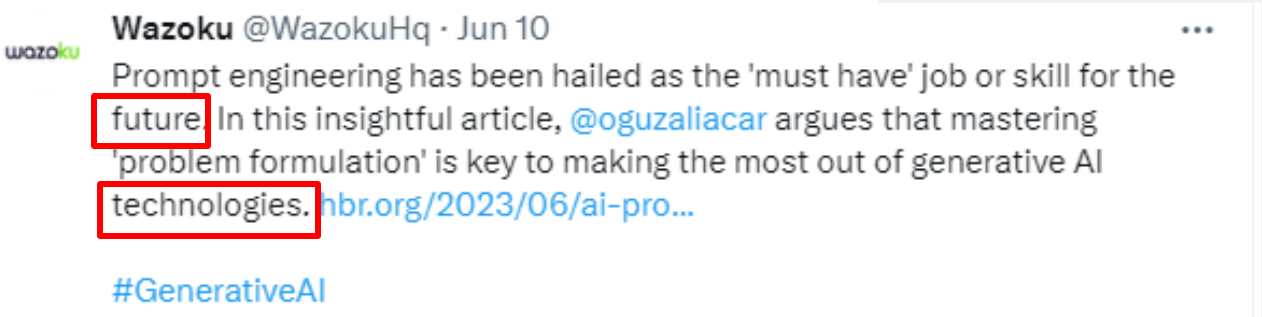


Research PReneurs

Recommendation 2: Emphasize "Innovation Exploration" Topic

While the count of tweets with replies in the "Innovation Exploration" topic was relatively lower, it exhibited a higher count of retweets:

- Spotlight emerging trends and innovative industry ideas
- Share thought leadership articles and future-oriented insights:



Recommendation 2: Emphasize "Innovation Exploration" Topic

- Highlight company's innovative projects to boost engagement
- Foster retweets with unique and inspiring content:

Research PRneurs

04

Next Steps

Limitations & Areas for further research

Limitations and Areas for Further Research

Limited Data

Final dataset is around 4k tweets → quite small to draw conclusions about engagement

Non diverse topic

Subject is not diverse in nature → makes it difficult to come up with topics that are significantly different

Low engagement performance

Current engagement performance is low → suggests that the tweets are not resonating with the target audience

- Increase the size of the dataset → looking at tweets from a wider range of companies or industries

→ Better topic modelling → identify the most engaging topics + draw more accurate conclusions about engagement

THANK YOU FOR YOUR ATTENTION!

Do you have any questions?

Research PReneurs