ANALYTICS STRATEGY CHALLENGE 2024

TEAM SAS
IIT KHARAGPUR

SHAMIK BHATTACHARJEE
ARUNAVA BHATTACHARYA
SWARALIPI DATTA

DATA-DRIVEN PHARMA MARKETING: LEVERAGING PHYSICIANS' DIGITAL FOOTPRINTS

ISB
ADVAITA 2024
ProcDNA

OVERVIEW

ISB | ADVAITA 2024 | ProcDNA

OBJECTIVES

Devise marketing strategies for **pharmaceutical companies** to engage with **physicians** based on their **digital activity**

Sentiment analysis
Analyze Physician Social Media Behavior

Marketing Strategies
Targeted marketing based on engagement

Cross-Platform Analysis

Referral Networks

NLP Techniques

Performance Metrics

Ensure scalable

Accuracy

EXECUTIVE SUMMARY OF DATA MODEL

Data Scraping
Collecting Healthcare Data from various online sources

Cleaning Data
Cleaning and preprocessing data for the model

Physician Tagging & Filtering
Segregating content posted by 'Dr's or 'Doctor's

Defining & Calculating KPI Scores
Finding relevant KPIs to rank the data

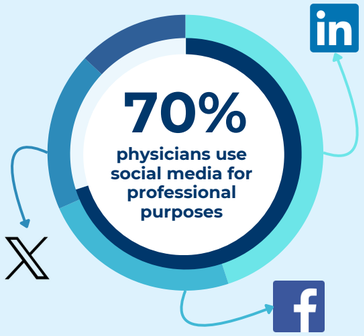
Calculating Final Weighted Score
Finding the digital footprint of each doctor

Classifying Key Influencers
Stratifying doctors based on footprints

DATA SOURCES & SCRAPING

239 M
US users actively seek health info on **YouTube** and **podcasts**

Niche Sites
like
doximity
Medscape
allow physicians to share opinions on specific topics



SCRAPING PIPELINE

Hashtags
50 to 60 Hashtags Generated by using GPT - 4o API

#MedTwitter #PatientAdvocacy #Telemedicine #ChronicIllness #HealthTech #PublicHealth

Check the top usernames associated with these hashtags

APIFY | Snscrape | Instaloader | tweepy

Scraping
Scrape Usernames

Scrape Posts, Comments, Likes, Followers and Mentions for each username (depending on the platform)

Parsing File
Parse the JSON file using Python script

Filtering Data
NLP techniques used-
- NER for extracting entities (locations, names)
- TF-IDF for keyword extraction.

Final Collected Data

METRICS

KEY PRINCIPAL INDICATORS

- Engagement Score
- Referral Score
- Comment Sentiment Score
- Domain Relevance Score
- Post Sentiment Score
- Affiliation Score
- Excitement Score

Decreasing Weightage

Comment Sentiment Scoring

Comments → Tokenization and Stemming → VADER sentiment analysis model → Generate Compound Score → Normalised Average Sentiment score

Affiliation Score

- Affiliation Score=Count of Doctors who have various Affiliation
- Key Affiliation Score=Count of Doctors in Key Partnership with companies

Weighted Affiliation Score = Affiliation Score × (1 + Key Affiliation Score / 10)

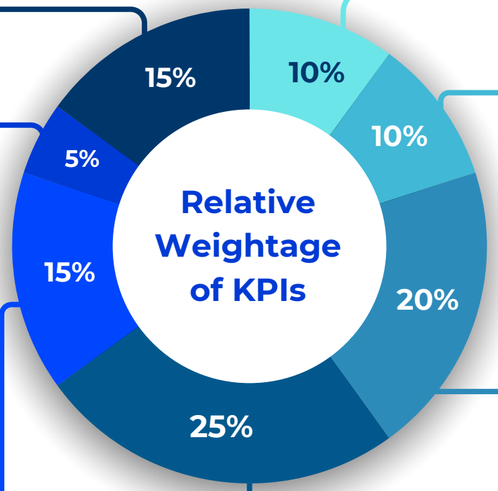
Post Perception Score

Post → VADER generated post sentiment score × Signum [Comment Sentiment Score] = Post Perception Score

Excitement Score

Excitement Score=(Keyword Count)×(Sentiment Score)

Keywords :- 'innovation' , 'breakthrough', 'advancement', 'cutting-edge', 'therapy', 'fda approval', 'drug development'



Referral Score

Doctor : Mention dictionary Map → Create Bidirectional Graph (NetworkX) → Calculate Betweenness Centrality

Domain Relevance Score

Cosine Similarity(A,B)= $\frac{A \cdot B}{|A| |B|}$

- A is the TF-IDF vector of the specialty.
- B is the TF-IDF vector of the posts.

Domain Relevance = Cosine_Similarity [TF-IDF(speciality,post)]

Engagement Scoring

Engagement Score = $\frac{(\text{Likes} + \text{Comments\_no} + \text{Shares})}{\text{Followers}}$

Created a Normalised Engagement Score

MODEL IMPLEMENTATION

ISB · ADVAITA 2024 · ProcDNA

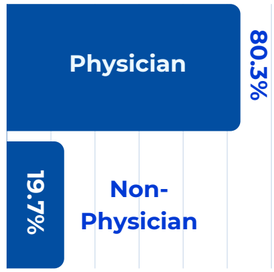
*Please find the data model and implementation on Python here

PREPROCESSING & PHYSICIAN TAGGING

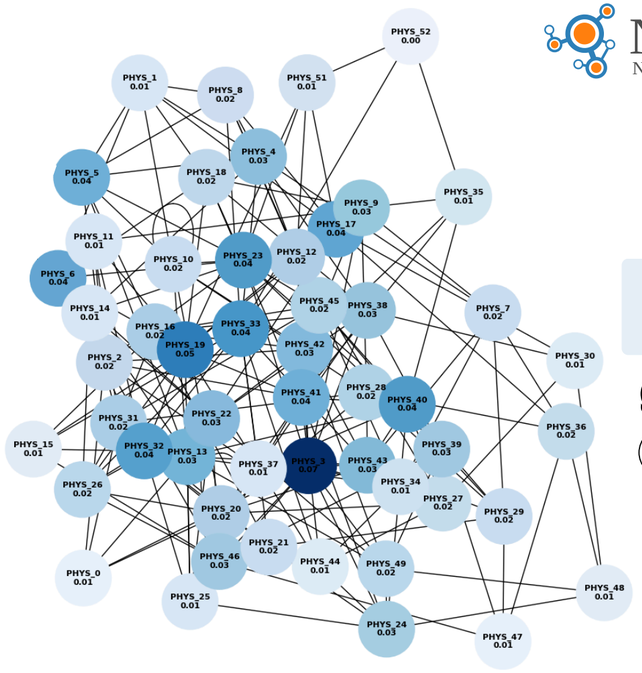
Representative tweets dataset scraped using Apify

Segregating the tweets by a doctor. Only these are used in the analysis

```
# Check for physician prefixes, specialties, or use NLP to identify person
 if any(prefix in name for prefix in ['Dr.', 'Doctor']) or \
    any(spec.lower() in specialty.lower() for spec in physician_specialties) or \
    any(result['entity'] == 'PER' for result in nlp(name)):
     return 'Physician'
 return 'Not Physician'
```

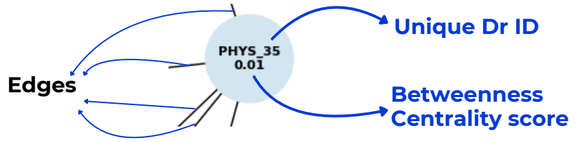


REFERRAL NETWORKS



NetworkX
Network Analysis in Python

has been used to create the adjoining graph



Betweenness centrality
is used for this purpose due to:

- Considers **Global Influence** instead of local influence
- Connects **Specialists** with **General Practitioners**

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

s, t: 2 distinct nodes
v: node under consideration
σ_{st}: *shortest paths between s & t*
$\sigma_{st}(v)$: *shortest paths between s & t which pass through v*

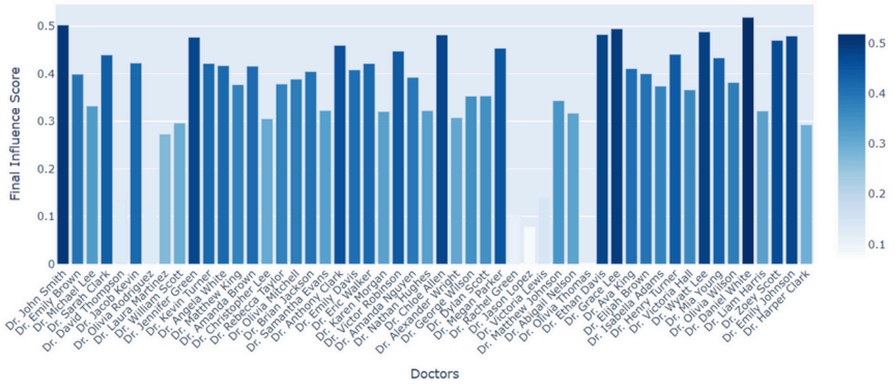
FINAL SCORING AND CLASSIFICATION

Final Influence Score =

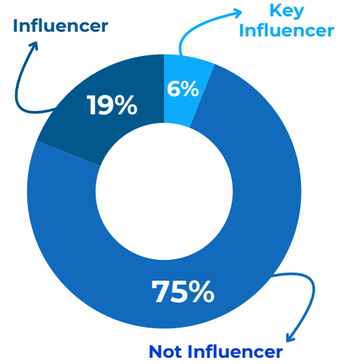
$$\sum_{i=1}^{n} (\text{Metric}_i \times \text{Weight}_i)$$

Top 25% Influencers
ie they have significant influence through digital footprint

Top 5% Key Influencers
ie they have *top* influence in their domian through digital footprint



Final Influence Score of Doctors



Classification of Doctors

```
def calculate_final_influence_score(row):
    """Calculates the final influence score based on weighted metrics."""
    return sum(
        pd.to_numeric(row.get(metric, 0), errors='coerce') * weight
        for metric, weight in weights.items()
    )
```

IMPROVEMENTS

EXTRA INSIGHTS

Data Driven insight on finding optimal Key Influencers for a particular drug launch

Doctor-Drug Association

- **Domain Analysis**
- **Post Analysis**
- **Publication Analysis**

Pharma → **Use these insights** → **Optimal recommendation of doctors for a drug**

```python
from drug_named_entity_recognition import find_drugs
# Function to extract drug names from posts associated with each doctor
def extract_drug_names_from_posts(posts):
    """Extracts drug names from posts using the find_drugs function."""
    drug_names_list = [[drug for text in post if isinstance(text, str) for drug in find_drugs(text.split())]
        if isinstance(post, list) else find_drugs(post.split()) if isinstance(post, str) else []
        for post in posts]
    return drug_names_list
```

Sponsor influential doctors with patents in a domain to drive visibility and credibility for your product through their research and online presence.

Sentiment Spread and Sentiment Shift Scores of comments to analyse the consistency of sentiments on a post

Forecasting Emerging key influencers by analysis temporal changes in KPI Scores

ADDITIONAL DATA SOURCES

- **Healthcare Journals & Publication Databases** — Google Scholar, PubMed
- **Health Review Platforms** — Teladoc HEALTH
- **Telemedicine and Online Consulting** — healthgrades, amwell
- **Podcasts and Videos**

CHALLENGES VS SOLUTIONS

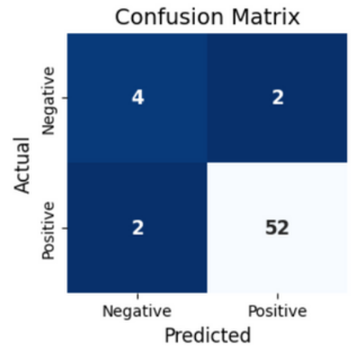
CHALLENGES	SOLUTIONS
API Restriction	Queue System and Paid subscription
Data preprocessing for large volume of data	Hadoop Apache Spark
Data Fragmentation Across Platforms	ETL Pipelines (Extract, Transform, Load) e.g., Apache NiFi, AWS Glue and Entity Matching Algorithms
Multilingual problem in comments	Google Translate API, BERT-based multilingual models

ACCURACY IMPROVEMENT

Accuracy Metrics:

- **Precision**
- **Recall**
- **F1 Score**

Benchmarked sentiment scores by pre-existing sentintment analyser



Accuracy: 71.67%

- **MAE measures the average error**
- **Accuracy is the proportion of predictions with errors below a set threshold (e.g., 0.1).**

Reducing Bias

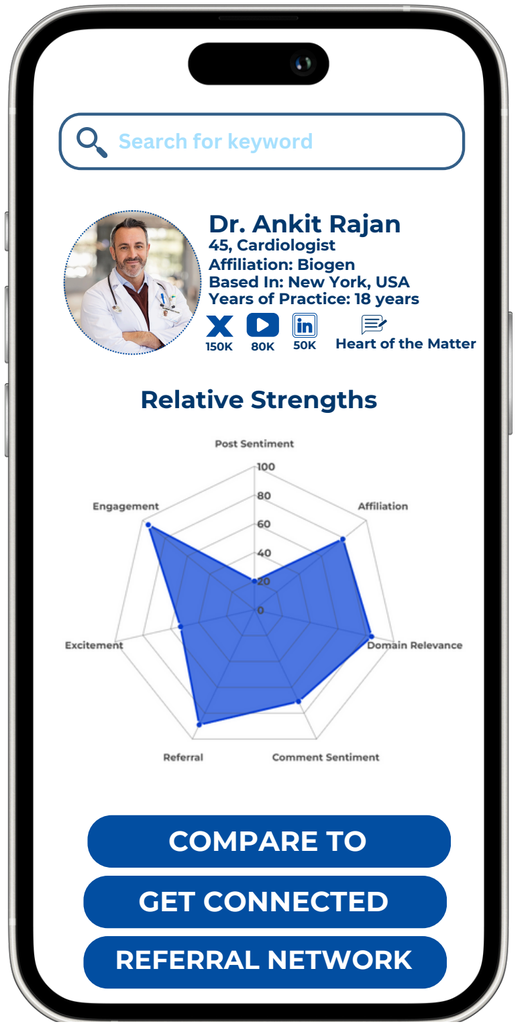
Fairlearn

Fairlearn Toolkit: Mitigates bias by applying fairness constraints such as demographic parity, equal opportunity

- Balanced Data
- Delocalized Collection
- Diverse Features

UI & DASHBOARD FLOW

USER INTERFACE



DASHBOARD FLOW

SCALABILITY & MARKETING STRATEGY

SCALABILITY

- **Handling High Data Volume - AWS, Azure, Google Cloud**
- **Ensuring Data Privacy and Compliance- BlockChain**
- **Processing Unstructured, Multilingual Data - Google mBERT**
- **Automating Analysis for Large-Scale Insights- AutoML**
- **Large Scale Data Engineering - Apache Spark and Hadoop**

Scalable Data Collection
Integrate LLM to optimize queries for snscrape and Apify, scrape data, process it with LLM, and store insights in cloud storage for scalable analysis

Scalable platform for pharma, offering key doctor engagement and influence metrics to drive targeted collaborations and boost product adoption.

SAMPLE SUBSCRIPTION MODEL

Pay-Per-Use Plan
- Custom Report Generation
- Market Access Strategy
- Competitive Benchmarking
- Campaign Engagement Analysis
- Real-World Evidence Reports (RWE)

Choose Plan

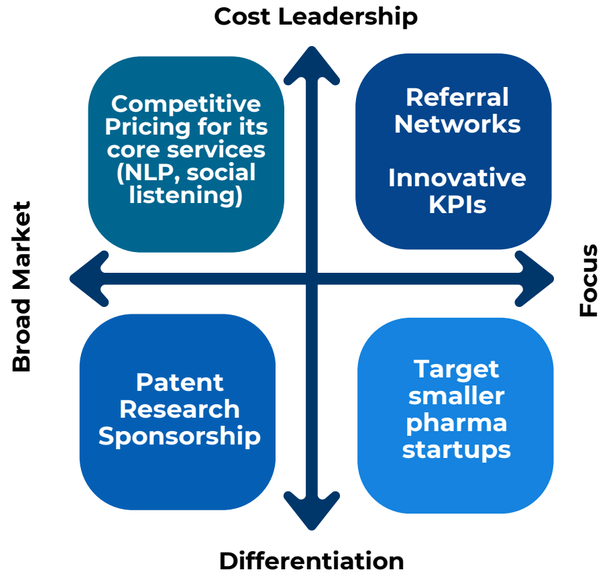
Subscription-Based Pricing
- Full Dashboard Access and Sentiment Analysis Reports
- Basic Referral Network Tracking
- Basic KPI Monitoring
- Regular Physician Segmentation
- Standard Marketing Strategies

Choose Plan

Premium Add-Ons
- Custom AI & ML Models
- Advanced Referral Mapping
- Clinical Trial Programs
- Custom Market Intelligence Reports
- Advanced KPI & Predictive Analytics

Choose Plan

MARKETING



INDIRECT MARKETING
- Targeted Email Campaigns
- Sales Representatives
- Webinars and Virtual Events
- Social Media Engagement

DIRECT MARKETING
- Influencer Collaborations
- Public Relations
- Content Marketing
- Educational Partnerships

STRATEGIC REQUIREMENTS

- Real-time insights into their marketing and engagement efforts.
- Flexible Pricing Model for both enterprises and startups
- Training of clients for ensuring smooth use of services

COMPETITOR ANALYSIS

INDUSTRY ANALYSIS

Global Pharma Market

$1.48T USD Trillion 2021

$1.85T USD Trillion 2026(p)

4.2 % CAGR expected till 2026

Latest Technology evolution (AI, Digital Health)

SWOT ANALYSIS

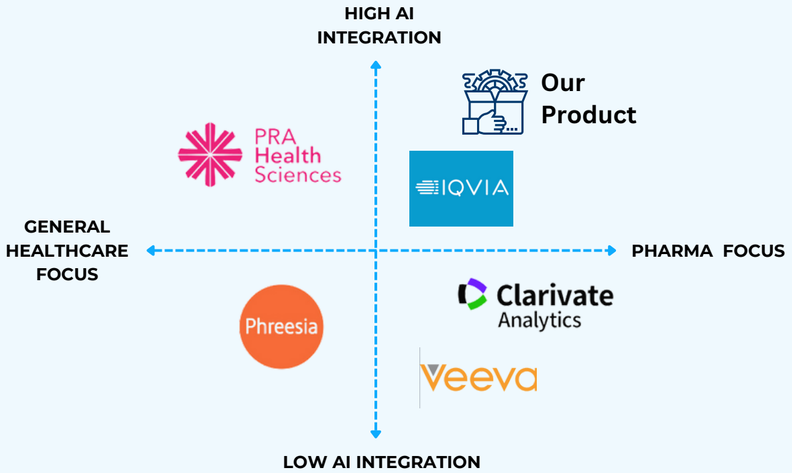
Strength
- Extensive use of state-of-the-art AI
- Scalable Infrastructure
- Cross-Platform Capabilities
- Physician Segmentation

Weaknesses
- Reliance on Social Media Data
- Data Privacy Concerns
- Complexity for Clients

Opportunities
- Expansion into New Markets
- Augmented and Virtual Reality (AR/VR)
- Sponsored Research and Clinical Trials
- Collaboration with Healthcare Organizations

Threats
- API Access Restrictions
- Physician Pushback
- Regulatory Scrutiny

MARKET POSITIONING



COMPETITOR ANALYSIS

	OTHERS	US
USP	Engagement Rates Patient Reviews	Social Media analytics based
Influencer Strategy	CRM based physician engagement	Key Influencer prediction with KPIs
Marketing Strategy	Focuses on approaching KOLs	Focus on Key Influencer Marketing
Technology	Sentiment Analysis of comments	Unique data engineering approaches and NLP techniques

APPENDIX

- Data Model
- Datasets and Notebook
- KOL Sentiment Analysis
- Pharmaceutical Company
- Important Resources

Thank you!