For Private Circulation Only

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

DATA GEEK

SCHOOL OF BUSINESS AND MANAGEMENT

TEXT ANALYTICS

BY

BUSINESS ANALYTICS SPECIALIZATION

VOLUME 3 | ISSUE1 | APRIL 2021

FROM THE EDITOR'S DESK

Information is the oil of the 21st century, and analytics is the combustion engine.” – Peter Sondergaard, Senior Vice President, Gartner Research.

Text analytics is the technique to extract relevant information from unstructured textual data. It facilitates business to discover patterns and themes that aids in understanding the customers needs and expectations. Businesses can improve customer satisfaction by learning what their customers like and dislike about their products, detect product issues, conduct market research, and monitor brand reputation. Emails, online reviews, tweets, call centre agent notes, survey results, and other forms of written feedback all contain text data that can be used for analysis.

With this, we present an enthralling Volume 3 Issue 1 of DataGeek newsletter, centered on a pivotal theme of Text Analytics. It includes interesting articles and capstone projects done on text analytics by the BA specialization students. Team would like to extend sincere thanks and gratitude to Mr. Mohammad Shoaib, founder & CEO of Lumiq for his industry insights. It also has crossword on concepts of text analytics. For the first time, the aptitude section has been introduced to motivate and help students practice for upcoming placements.

I would like to extend gratitude to our Dean, Dr. Jain Mathew, Associate Deans Dr. Georgy Kurien and Dr. Jeevananda S, Head of Specialization – BA, Dr. Lakshmi Shankar Iyer for their guidance in making this issue a success. Also, a special appreciation to the newsletter team for the effort, time and inputs without which this issue would not have been possible. A thanks to all the students who have provided their valuable inputs. Once again congratulations to the entire team.

Please reach out to us for any queries or suggestions at datageek@mba.christuniversity.in

With regards,
Dr. Tripti Mahara

CONTENT

INTRODUCTION TO TEXT ANALYTICS

Humans have great ability to understand and derive patterns from the available text. We daily come across text data in diverse feeds, hashtags, trending topics across twitter, facebook, newspaper, reviews provided by customers on various sites etc. If this same pattern and insights have to be derived through an automated process, it is termed as Text Analytics. This automated process converts large volumes of textual unstructured data into quantitative data to uncover insights to aid in decision making.

Every business strives to provide the best to their customers. To achieve this, they are depending on text analytics to study and understand patterns, drifts in behaviour through the positive and negative feedback provided, buying trends, opinions of consumers, blogs etc. and modify the approachability to satisfy needs which can make a greater impact on business. By implementing text-based analytics, a business can bridge the gap to unlock the very needs and demands of the customers. Text analytics focuses on quantitative insights that give the essence of 'why' a particular problem arises and 'what' the reasons are and upon understanding, 'how' can a business overcome it in the most effective way. Various tools like HANA, Python, R, Microsoft excel etc can be used to achieve important tasks of Text analytics as discussed below.

Important Tasks in Text Analytics:

Information Extraction: It involves extracting the relevant information from large volumes of textual data. It centres on extracting attributes and entities. This information can be used for further analysis. (Rai, 2019)

Information Retrieval: Information Retrieval (IR) alludes to extricating relevant and related examples dependent on a particular arrangement of words or expressions. In this content mining strategy, IR frameworks utilize various calculations to track and screen client practices and find applicable information as needs are. Google and Yahoo web indexes are the two most famous IR frameworks. (3rdi, 2018)

Clustering: It looks to recognize characteristic constructions in text-based data and sort them into relevant subgroups or 'bunches' for additional examination. A critical test in the grouping interaction is to frame significant groups from the unlabelled text-based information without having any earlier data on them.

Summarisation: This content mining strategy helps to create a summary of a large volume of text in a way that the meaning and intent of the original document is preserved.

Categorization: This technique is used to classify text (review, paragraph, document) into a relevant category. The text could be the reviews provided by different users for a product and the reviews could be classified as positive or negative. Similarly, a mail can be classified into a spam or non spam email.

References:

1. 3rdi. (2018, September 12). 5 Common Techniques Used in Text Analysis Tools. Retrieved from 3rdi: https://www.3rdisearch.com/5-common-techniques-used-in-text-analysis-tools
2. Rai, A. (2019, June 01). What is Text Mining: Techniques and Applications.

Sai Sandeep Bhoslay
2027032

Mayank Bali
2027202

CAN TEXT ANALYTICS HELP INDIA IN COMBATTING MENTAL ILLNESSES?

World Health Organization (WHO) defines mental health as a state of well being in which one realizes one's own abilities and can cope up with stress and other difficulties in one's life. A mentally healthy individual works productively in order to contribute something to his or her community. October 10th every year is celebrated as World Mental Health Day so as to raise awareness about mental health, as it's often overlooked as compared to physical health. Different types of mental illnesses include depression, substance abuse, personality disorders, Post Traumatic stress disorder, anxiety disorder etc.

WHO estimates that about 7.5 percent Indians suffer from some mental disorder, out of which 56 million Indians suffer from depression and 38 million people suffer from anxiety disorder (Bhatia, 2020). This is crucial in a country like India where people are hesitant to talk about their mental issues in a fear of being mocked by the society and lack of awareness about the importance of the issue. The issue is worsened with the Corona Pandemic outbreak in 2020 which created a huge blow to overall mental health of the people across the nation.

Text Analytics can help India tackle this serious issue to a large extent. It is an AI technique which uses Natural Language Processing (NLP) to convert unstructured text data to structured format suitable for analysis. Text Analytics then work on these structured data to draw insights and represent it in charts and graphs.

One of the primary reasons for people not seeking help from professionals are the fear of opening up to another human with a fear of being judged. This issue can be solved to a significant extent by deploying chat bots for counselling. These bots can mine the conversations entered by the patients and identify the intent. It can be a general query or specific request seeking mental support or requires emergency responses. It can start by asking the person's name, location, contact number of the person and a beloved one too if in case of emergency. Machine Learning can be incorporated in these chat bots by using Natural Language Processing, where the machine learns from the input it receives.

The framework can be adopted by incorporating unsupervised learning to identify patterns and hidden intents in the text. For example, if a person who is undergoing an episode of depression converses with the chatbot, the machine can mine through the texts and identify the intent by matching it with symptoms of depression. The chat bot can then suggest measures to be undertaken by the person such as taking deep breaths or to consult a nearby psychiatrist by actually looking into the location of the person. Chat bots can also help in preventing suicides. This can be done by alerting emergency responses to the person's location when text mining tools pick up certain keywords related to suicide as in 'I want to kill myself', 'I am done with life', etc.

Text Analytics helps in creating a bigger image of the underlying issues to mental health. It helps in visualizing the data to draw insights from the same. The insights drawn can actually help decision makers like Government, hospitals, doctors, mental institutions etc. to combat such a serious issue. The data can cover aspects of age groups, gender, educational qualifications, income etc. and identify groups that are prone to mental illnesses. For instance, if we find a particular age group, say school children to be committing suicide, we can infer that it might be due to overload and stress. The decision makers can then adopt measures to prevent the issue from being worse.

Mental health is as equally important as physical health. There are prospects of incorporating new technologies in the field of mental healthcare just as in Cardiology or Oncology. Text Analytics is predicted to grow at a rate of 18% from 2019 to 2026 (Text Analytics Market Research, 2021). We can incorporate this growing field of Analytics into mental health care and reap the benefits out of it. It can contribute to the overall well-being of society as it caters to different mental requirements of different individuals suitably, thereby creating a society of mentally healthy people. An individual with a sound mind can be more productive and this can create a huge difference in her or his life. Thus, use of text analytics will be a boon to the entire community or nation and will help free from the hard clutches of mental illnesses.

NAADEEN ABDUL KARIEM
2027249

TEXT ANALYTICS IN SOCIAL MEDIA

Social Media sites like Twitter, Facebook, Instagram are places where people share their thoughts on various topics on the spur of the moment, supplying with a plethora of instant customer input. However, as the number of social media references grows, manually tracking them is no longer feasible. Hence, text analytics techniques are used to analyze and interpret these texts.

Text Analytics can help businesses to draw insights about various areas like understanding of the general sentiment or emotions expressed about a particular brand or a particular product, knowledge about what conversations (supporting arguments or controversies) regarding a particular brand, or a particular product is floating on social media, understand the purchase behavior of consumer, to comprehend the key issues in product feedback (topic detection).

It not only helps businesses but also other stakeholders. For instance, it can help to predict which political party is more likely to win the campaigns. Sentiment analysis and Topic Modeling can be applied here to analyze the people's opinion and perspective on apolitical party and can predict the possible outcome of the campaign.

During the campaign the impact generated by the candidate is monitored by taking the sample of people's response from social media and then computing impact percentage and sentiment index description.

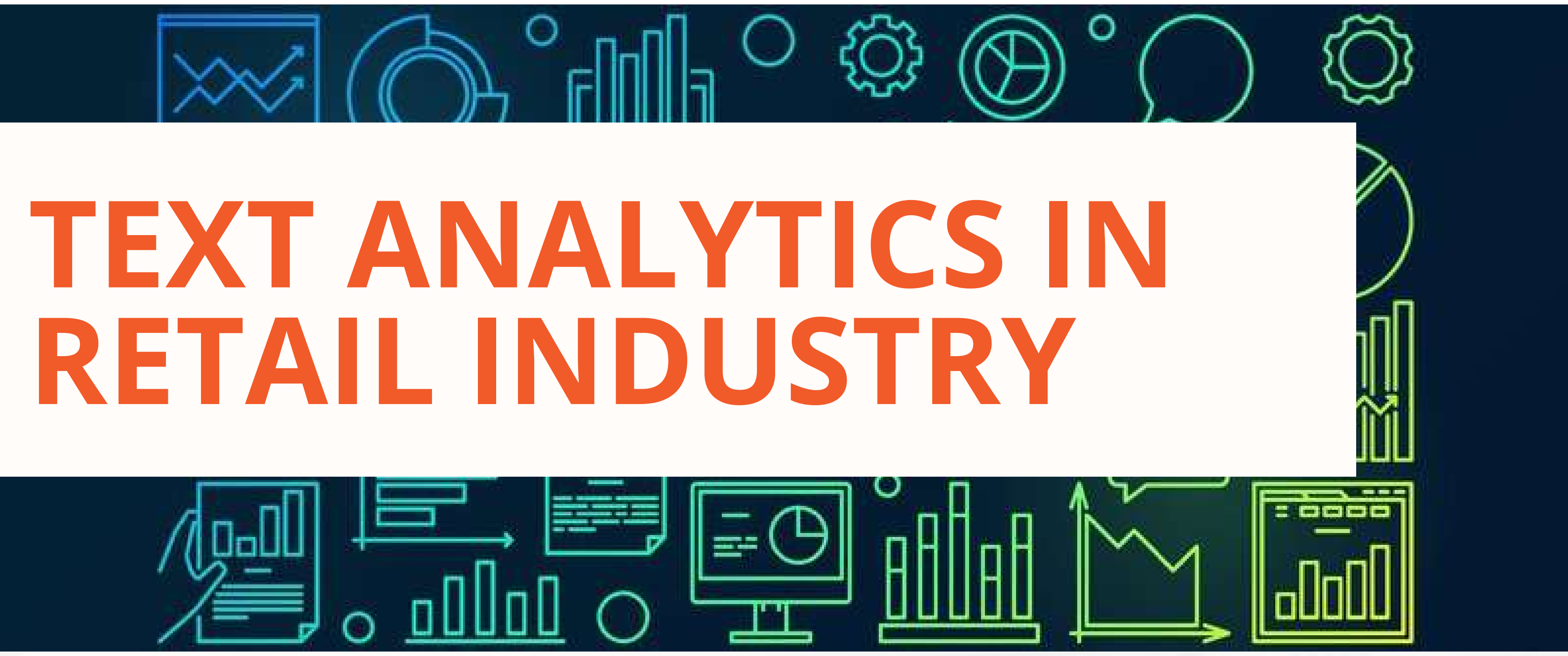
The collected samples contain information such as number of likes, comments, emojis, etc. on the campaigns. Then, insights on impact percentage is calculated using keywords and expressions. Alongside, the index description is also analyzed to obtain insights on people's perspective such as positive and negative sentiments on a particular party. From impact percentage and index description valuable visualization is achieved. It also enables federal agencies and national security authorities to monitor the behavior of the citizens by analyzing the potential threat topics and objectionable matters shared on social media and many more. Text analytics in Social media incorporates three-sort out steps: Gather, Analyze, and Visualize. "Gather" incorporates getting appropriate social media sources, collecting huge information and conveying important information.

These text information is created by people through posts, tweets, browsing, feedback, etc. Not all data that will be collected from social media will be significant. "Analyze" picks significant data from immense quantities of text information for illustrating by removing noise, inferior quality data, and uses distinctive data analytics techniques to explore the data, rearrange it and get some valuable insights from it. "Visualize" presents the data collected in the "Analyze" step in a more significant manner.Thus, text analytics on data available of social media helps in enhancing the brand value, retain customers and increase customer acquisition by understanding their concerns and addressing them, boosting campaign performance.

KARTHIK M
2027011

MOUMITA DUTTA
2027036

BAKKI AKHIL
2028229

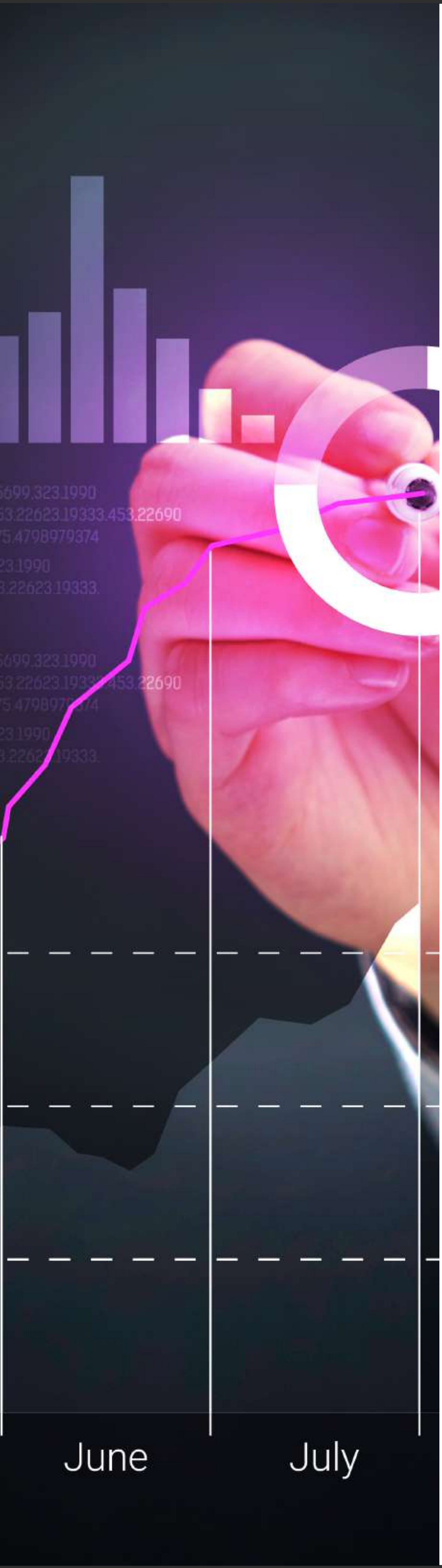


TEXT ANALYTICS IN RETAIL INDUSTRY

E-commerce industries are primarily customer-driven. An e-commerce business must optimize the shopping experience to enhance customer satisfaction. A customer describes his/her shopping experience by providing reviews and ratings. The use of mere ratings to understand and analyze customer satisfaction isn't enough as each customer delight level varies for perspectives. Thus, organizations would have to uncover customer thoughts about the brand further, as expressed by the customers online through written feedback or reviews.

Successful companies listen, understand, prioritize and cater to customer feedback as provided through reviews. About 80-90% of this data is unstructured. As quoted by Peter Drucker, "If you can't measure it, you cannot improve it", which implies how crucial it is for the brands to measure these customer text datasets.

Text analytics powered by AI, Machine Learning and Natural Learning Processing includes techniques to analyze unstructured data to describe, interpret and understand text, and provide solutions. Successful e-commerce brands use text analytics to gain insights into the customer text dataset, determine customer trends, identify patterns, verify the hypothesis and discover new topics that require stakeholder's attention.



Sentiment Analysis is one of the key applications of Text Analytics in the retail domain. While analyzing the text data for sentiment analysis, the focus is on finding the positive and negative words that describe the shopping experiences and customer support. Negative reviews suggest a scope of improvement by the organization to meet the expectations of customers. Similarly, positive reviews increase the popularity of the product and lead to an increase in sales. This identification helps develop the experience and fills the process gaps, helps maintain the business's SLA by giving on-call resolution or on chat at the first time of customer reaching to the company. Customer feedback in third party websites can be obtained and analyzed for understanding the customer trend. Several similar patterns and trends in the data help organizations prioritize customer issues, understand the essential topics and observe new topics to improvise customer satisfaction which is the core for any business within any industry.

Collecting relevant reviews from various eCommerce platforms, followed data processing are the initial steps of Sentiment Analysis. Basic pre processing tasks include tokenization (breaking sentences into words), stop words removal and lemmatization

This is followed by vectorization and building a classification model that categorizes a review as negative or positive. Visualization techniques like word cloud, interactive maps and various charts can be used to present the findings of sentiment analysis. Along with the review data, an organization can also make use of social media data, emails, text survey, blogs to understand the feelings of customers. This helps to bring out insights that will aid organizations in future decision making.

AS QUOTED BY CLIVE HUMPY, "DATA IS THE NEW OIL" AND 80% OF THIS DATA IS UNSTRUCTURED COMPRISING OF TEXT, AUDIO, VIDEO. THEREFORE, TEXT ANALYTICS WILL CREATE VALUE FROM THE TEXTUAL DATA AND HELP BUSINESSES IN CUSTOMER RETENTION LEADING TO AN INCREASE IN PROFIT.

ARS YUVANESH
2027827

ARPITA VINAY GAONKAR
2028244

CORPORATE CONNECT

MOHAMMAD SHOAIB

Mr. Shoaib is the founder & CEO of Lumiq. Prior to founding Lumiq, he has worked with corporate giants like Google, Cisco, NASA, Yamaha, Wincor Nixdorf, NTUC FairPrice, and Tangs. He has global experience in building million-dollar enterprise products. He is an alumnus of IIM Ahmedabad and NIT Kurukshetra.

CORPORATE CONNECT

Could you please tell us about your role in the company and something about the company?

I am the founder and CEO of Lumiq. Lumiq is an end-to-end data analytics company that enables enterprises to make better sense of their data and monetize it. Through a unique blend of data science, data engineering, and intelligent automation, we help customers to leverage their data to drive actionable insights and deliver tangible business outcomes.
We work with leading enterprises in the Financial Services domain.

Can you talk about the company and its focus areas?

Lumiq has product offerings across the value chain specifically targeted for the financial services industry. We are a full stack data science company and our expertise lies in Data Science, Data Engineering, DevOps and ML ops. Following are some of our offerings:

Drishti - A complete Document Intelligence & Document Understanding engine with capability to process handwritten docs. It is very versatile and can be plugged in, to automate any process where document processing is an integral part, like Customer Onboarding, KYC, PII Handling, Fraud/Forgery detection to name a few.

AUTOM8 - An omni channel friction less customer interaction bot that can deflect volume on all your digital channels (email, bots, skype, slack and so on) without compromising on customer experience.

Aurum - An NLP driven real-time enterprise wide, evolution friendly customer deduplication product which can be used across lead qualification, underwriting, credit appraisal workflow and cross-sell & upsell cycles.

smartUW - A comprehensive AI/ML-driven end-to-end underwriting assistant to look deeper into individual customer attributes.

emPower - Our financial services focused cloud data platform and AI warehouse that helps enterprises to organise and make all data easily available.

Does your company use Text Analytics to build any product or services that it offers? If yes, please elaborate on it.

Though many of our products leverage Text analytics Drishti has the biggest share of the text analytics under the hood. Drishti is a document AI product built with the objective of extracting information from the documents and organizing it to enable automation of end to end business process decisioning. For example, it is very common for KYC documents to be collected for validation during the buying journeys for most financial products (insurance, loans, credit cards and so on).

The supporting documents are often uploaded and more often than not a human or at best OCR driven processes are used to extract and pre-fill the information from the forms. Drishti removes the need for human intervention barring oversight.

Drishti follows an end to end approach whereby it allows a three way matching of the application form data along with the relevant information extracted from supporting documents. It then alerts users only in case it sees discrepancies thereby eliminating the need for manually processing or even clicking on the buttons.

Only documents with exceptions or where the algorithms have low confidence are sent for manual eyeballing and review - the data collected as part of this process is further used to refine the models.

Drishti processes the document via a series of steps whereby pre-processing takes care of issues like reorientation, doc/color corrections, boundary detections, enhancements of scan quality and so on. The second step is the actual data extraction using textract and the third leg of the process is the insights generation where by POST OCR corrections, layout information, extraction, additional insights, signature/photo matching and so on are performed. The last leg is usually as per the needs of the underlying business process.

We have been able to get as high as 95% plus accuracy for certain use cases. Drishiti is offered as a SaaS solution and works straight out of the box for most documents. As with all models with more data the models become smarter and performance keeps on improving.

What do you think is the future of text analytics and which sectors have the potential to use it for business growth?

Text Analytics Market size exceeded USD 6 billion in 2020 and is poised to register gains at around 20% CAGR between 2021 and 2027. These are unprecedented times in terms of traction for text analytics and next two years will see even more success stories.

The biggest challenge/opportunity in text analytics is around inferring conceptual metaphors and narrative patterns in a text. Another challenge is to generalize extraction of information from documents (text) requiring minimum data.

One of the future trends in text analytics would be improving on the foundation of text categorization, increasing the capabilities to classify intent, context, category, nature, sentiments of the text. Another potential future trend is suggestive AI, even if the model is not confident about taking an action it should at least make suggestions to the end user so that he/she can take proper steps.

What should an ideal career path in the text analytics domain be?

As a fresher you need to be familiar with data analytics tools in general to begin with text analytics. Tools such as python, Excel, R, Tableau are good to begin with. Get well versed with topics like Tokenization, Normalization, Stemming, Lemmatization, Corpus, Stop Words etc. After basics are clear more advanced topics like Parts-of-speech (POS) Tagging, Statistical Language Modeling, Bag of Words, n-grams,
Regular Expressions, Zipf's Law, Similarity Measures, Syntactic & Semantic Analysis, Sentiment Analysis and information retrieval should be worked upon. At an experienced level some research oriented topics like sequence to sequence, reinforcement learning, auto encoders, one shot learning should be worked upon. Begin with implementing research papers in these domains and later on can implement your own research work and propose new architectures.

What kind of skill set do you think a person should have other than the knowledge of tools to have a successful career in the Analytics domain?

A great business analytics professional must have skills including good communications, problem solving, critical thinking, visualizing. Most importantly the ability to seamlessly & iteratively move from a detail oriented view to a macro view and back.

Apart from the core data science skills, exposure to SQL, APIs, big data tooling in general and exposure to Amazon AWS, Microsoft Azure, or Google Cloud (especially the ML tools like Sagemaker and so on) make a very focussed professional who has understanding of modelling, deployments and governance.

What would be your advice to aspiring Analytics Professionals?

I guess the first and foremost thing would be to be empathetic to the end user. Wear their hats - look at the problem from their point of views and how they experience it - this would help you identify the problem correctly and make the right design choices and decisions. Also, while making decisions think as if it's your personal money being invested to solve the problem. Decompose the problem into smaller chunks and never lose sight of the end customer. Collaborate, seek feedback, validate assumptions so that you identify any gaps and you build what solves the problem. Remember it's all about generating tangible value. Also, question and challenge the data - understand it. The biggest challenge on ground in data science is the data. Be open and be pragmatic.

Secondly, analytics professionals must be a lifelong learner. If you look back at the pace of democratization and developments in the analytics space be it sharing of the research by academics or by companies - just keeping up to date could be a demanding task. To be a problem solver one must know what is the right arrow to shoot at the problem. Keep working to develop your technical and non-technical skills.

LEADING COMPANIES IN TEXT ANALYTICS

RAPIDMINER

RapidMiner brings AI to enterprise to rapidly create and operate AI solutions to drive immediate business impact to positively shape the future. RapidMiner provides an end-to-end platform that integrates data preparation, machine learning, and model operations, as well as a user experience that provides depth for data scientists while simplifying complex tasks.

CLIENTS:
Fidelity,Goodhill,DHG,Alliant,Alcoa,Michelin,denso,ArcelorMittal,Abbott,SANOFI,MMC,HP,Domino’s

AYLIEN

AYLIEN is an AI, NLP & Machine Learning that provide Text Analysis & News API's which allow users to make sense of human-generated content at scale. The company provides a wide range of content analysis solutions catering to businesses, data scientists, marketers, developers and academics. AYLIEN mission is to make sense of the incomprehensible by leveraging AI to extract insights and enable businesses to make valuable decisions.
CLIENTS:
Wells Fargo, MOODY’S, AON, Revolut, DECK, AMPLYFI, IHS Markit

LUMINOSO

UMINOSO TECHNOLOGIES IS AN ARTIFICIAL INTELLIGENCE (AI) AND NATURAL LANGUAGE UNDERSTANDING (NLU) COMPANY THAT HELPS BUSINESSES TO RAPIDLY DISCOVER VALUE IN THEIR UNSTRUCTURED DATA. LUMINOSO'S AI ANALYZES TEXT-BASED DATA ACCURATELY FOR ANY INDUSTRY WITHOUT REQUIRING LENGTHY SETUP OR TRAINING. THEIR SOFTWARE CAN ANALYZE UNSTRUCTURED DATA NATIVELY IN 14 DIFFERENT LANGUAGES. LUMINOSO'S SOLUTIONS PROVIDE COMPANIES VALUABLE INSIGHTS INTO BUSINESS PROCESS IN ORDER TO STREAMLINE THEIR CONTACT CENTER PROCESSES, MONITOR BRAND PERCEPTION, AND OPTIMIZE THE CUSTOMER EXPERIENCE.

CLIENTS:
ATHENAHEALTH, SWISS POST SOLUTIONS, MICROCHIP, FANCY, HALO8, HILTON, C_SAPCE, DENSO

SCIBITE

SCIBITE IS PAVING THE WAY BY PIONEERING THE COMBINATION OF THE LATEST IN MACHINE LEARNING WITH AN ONTOLOGY-LED APPROACH. SCIBITE'S SEMANTIC INFRASTRUCTURE PROVIDES REAL-TIME ANSWERS TO BUSINESS-CRITICAL QUESTIONS BY UNLOCKING THE VALUE AND FULL POTENTIAL OF UNSTRUCTURED DATA. SCIBITE'S API TECHNOLOGIES ARE FAST, FLEXIBLE, AND DEPLOYABLE, MAKING THEM A CRUCIAL COMPONENT IN SCIENTIFIC, DATA-DRIVEN STRATEGIES.

CLIENTS: UNILEVER, SANOFI, TAKEDA, NOVARTIS, PFIZER, GSK, ABBVIE, ASTRAZENECA, BIOGEN, CELGENE, BOEHRINGER INGELHEIM

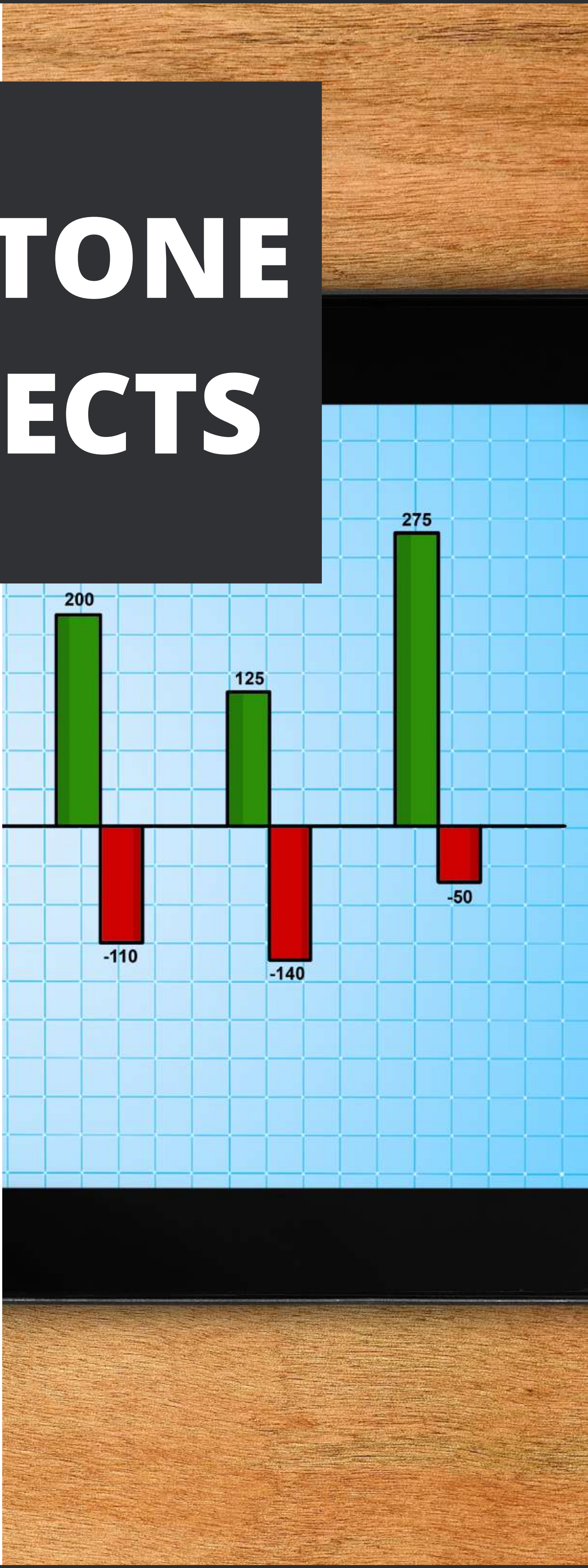
CONVERSEON.AI:

Converseon is a pioneer in AI-powered social listening analysis and insights, as well as associated voice of customer data. The company provides the highest quality social listening data and associated insights, and assist clients in amalgamating these across the organisation to drive business results. Conversus, a SaaS Machine-Learning-As-A-Service platform offered by Converseon that integrates with select leading social listening/management, pre-built machine learning classifiers for NLP deployment, and a suite of social/AI powered programmatic "Insights-on-Demand" solutions.

CLIENTS:
Forrester, 3M, DOW, Hilton

Komal Nagarajan
2027654

CAPSTONE PROJECTS



SENTIMENT ANALYSIS AND DETECTION OF SUICIDAL IDEATION IN MENTAL HEALTH SUBREDDITS

The project aims to cluster posts on Reddit especially in the subreds related to mental health in order to identify posts that are indicative of suicidal intent. Traditional methods of assessing suicidal intent in messages and social media posts were by manually categorising the text and building a supervised learning model. This project hopes to cut down on the work required to classify the texts by using emotions as a foundation to identify suicidal intent. The project used VADER and NRC Lexicons to find the polarity and emotions associated with text. These were then classified into three categories namely suicidal intent, positive intent and negative intent based on the emotions and the polarity of the emotions.

The words associated with the emotions clearly indicated that fear is being associated with the suicidal intent as most occurring words in this emotion are "kill' and 'myself'. In both positive and negative emotions the people posting are asking people who have faced similar issues to approach them via direct messages. The practical implications of the project is that it can be used to build further models which can track social media posts of individuals and with the help them if there changes in their emotional state over prolonged periods of time.

Aditya Manoj Tiwari
1928161

SOCIAL MEDIA ANALYSIS FOR NATURAL SKIN CARE PRODUCTS USING TEXT ANALYTICS

Skincare industry is observing a behavioral change as customer preference has changed from using harsh chemicals to preferring toxin free, vegan, cruelty free organic products for their daily skincare routine. This project aims to analyze the customer reviews from amazon for a skincare brand named Mamaearth and study the sentiments of customers in order to observe the shift in the skincare industry towards natural toxin free products. Sentiment analysis helps business to understand sentiment of their customers, their behavior and analyze customers brand perception, trending words, positive or negative attitudes and a variety of feelings-based summaries for a product. By performing text analytics on amazon reviews, Mamaearth can know its customers preference better. The objective of this research paper is to identify and classify reviews into positive, negative and neutral categories. Then further, we will be drilling down positive and negative reviews separately.

Once we have as set of positive and negative customer reviews, we will use machine learning techniques- Topic Modelling to group in similar kind of reviews.

The data set used for the research purpose is extracted from amazon. Extraction of reviews is done using scrapy library in python. After extraction, data preprocessing is done on extracted amazon reviews using NLTK library in python. For pre-processing process, we have removed punctuations, converted data into lower case, removing rare words, removing stop words, performed stemming and lemmatization on the dataset.

Once the reviews are pre-processed, we have performed Exploratory data analysis to plot frequency of words and word cloud. After data pre-processing and exploratory data analysis. Further, we have used Topic modelling technique to identify phrase patterns within reviews, then grouped them to find meaningful insights that will be helpful for the company.
Sentimental analysis on amazon reviews helps the company to understand their customer's liking and behaviour which in further helps company to improve their products. At the end of this project, we can conclude positive and negative aspects about Mamaearth products, this can be used by the decision makers of the company to improve customer experience.

Sampada Pandey
1927745

FOOD RECIPE RECOMMENDATION SYSTEM

The list of foods and their health benefits was compiled using information from www.livestrong.com and www.nutritiondata.self.com. The majority of the data was collected on foods that contain milk or fish as the primary ingredient. Since they are both solid foods, the various fixings that are added to them during the food preparation process are the donors that increase the estimates of other nutritional benefits such as calories. The dataset contains 2,31,637 rows and 12 columns, but the variables under investigation for the food recommendation are recipe name, recipe id, tags, nutrition, description, number of steps, and what are the steps.

DATA PREPARATION:

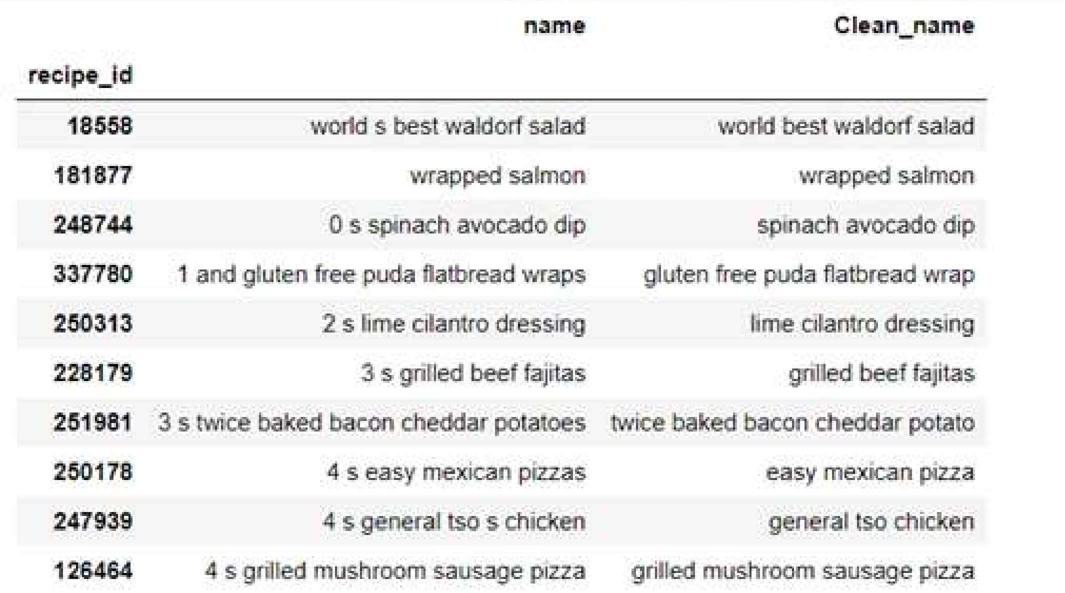
Data isn't flawless all of the time. Missing values, raw data and data anomalies are the most common data errors. As a result, the above-mentioned issues were resolved as part of the data cleaning process. The data types were translated into the appropriate formats. Data integration was also done in order to produce a final data set that could be used to construct models.

Since the dataset was huge (219789) and the name of the recipes was not appropriate, text cleaning the various text preprocessing steps are:

- Tokenization
- Lower casing
- Stop words removal
- Stemming
- Lemmatization
- Punctuation removal

The aim is reducing the word's inflectional and often derivationally related forms to a common base form and reducing the number of total terms to a few “root” terms

The project was carried out in a systematic manner using the CRISP-DM methodology. Understanding the variables under investigation is essential for developing a solution to the defined problem and achieving the set goals.

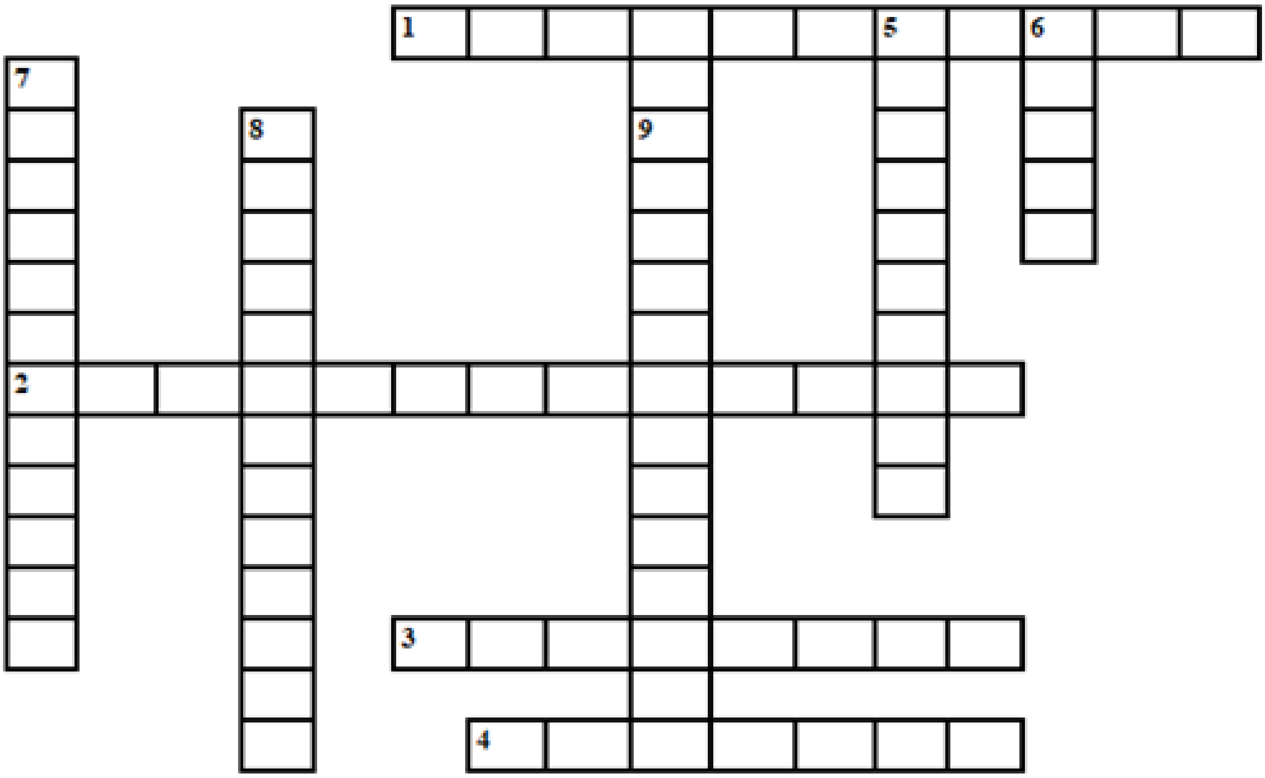


recipe_id	name	Clean_name
18558	world s best waldorf salad	world best waldorf salad
181877	wrapped salmon	wrapped salmon
248744	0 s spinach avocado dip	spinach avocado dip
337780	1 and gluten free puda flatbread wraps	gluten free puda flatbread wrap
250313	2 s lime cilantro dressing	lime cilantro dressing
228179	3 s grilled beef fajitas	grilled beef fajitas
251981	3 s twice baked bacon cheddar potatoes	twice baked bacon cheddar potato
250178	4 s easy mexican pizzas	easy mexican pizza
247939	4 s general tso s chicken	general tso chicken
126464	4 s grilled mushroom sausage pizza	grilled mushroom sausage pizza

Python libraries such as numpy, pandas, scikit-learn, and matplotlib are critical to the project's success. Optimization of memory, The techniques used in the analysis to achieve the goals and objectives were one hot encoding and Collaborative dependent filtering. Finally, although the analysis has drawbacks, they can be interpreted as potential research areas that can be investigated for new ideas in the future.

Feni Martina Fernando
1927958

CHECK YOUR KNOWLEDGE



ACROSS:

1. Name the technique that breaks the raw text into words, characters or subwords.

2. Name the function that finds the structural relationship between tokens adhering to the grammatical rules.

3. Number of times a token occurs in a text.

4. Two or more words that tend to appear together frequently.

DOWN :

5. Important words/ topics that exist in a text

6. A statistical measure that assesses the relevance of a term to a text in a set of documents.

7. Name the process of breaking the text documents apart into pieces.

8. In NLP, name the technique that can be applied for keyword normalization to transform a keyword to its base configuration.

9. Name of the text analysis technique that determines the most commonly appearing words or ideas in a given text.

CHECK YOUR SCORE

1. TOKENISATION
2. SYNTAXPARSING
3. WORDFREQUENCY
4. COLLOCATION
5. KEYWORD
6. CLUSTERING
7. CHUNKING
8. LEMMATIZATION
9. TFIDF

Anchal Aneja
2027245

CHECK YOUR APTITUDE

1) A mother is twice as old as her son. If 26 years ago, the age of the mother was 10 times the age of the son, what is the present age of the mother?

a) 57 years
b) 62 years
c) 61 years
d) 59 years

2) A boat can travel with a speed of 14 km/hr. in still water. If the speed of the stream is 3 km/hr., find the time taken by the boat to go 68 km downstream.

a) 6 hours
b) 3 hours
c) 4 hours
d) 5 hours

3) A pipe can fill a tank in 7 hours and another pipe can empty the tank in 14 hours. If both the pipes are opened at the same time, the tank can be filled in

a) 10 hours
b) 12 hours
c) 14 hours
d) 16 hours

4) Two ships are sailing in the sea on the two sides of a lighthouse. The angle of elevation of the top of the lighthouse is observed from the ships are 30° and 45° respectively. If the lighthouse is 120 m high, the distance between the two ships is:

a) 173 m
b) 273.96 m
c) 310.7 m
d) 327.6 m

5) In this series 8, 11, 9, 12, 10, 13, ... What number should come next?

a) 11
b) 12
c) 9
d) 13

6) A person crosses a 900 m long street in 5 minutes. What is his speed in km per hour?

a) 10.8
b) 7.2
c) 8.4
d) 10

7) If in a certain language, NOIDA is coded as OPJEB, how is MUMBAI coded in that language?

a) CDKGH
b) NVNCBJ
c) FGNJK
d) IHLED

8) Peter is in the west of Tom and Tom is in the south of John. Mike is in the north of John then in which direction of Peter is Mike?

a) South-East
b) South-West
c) South
d) North-East

9) I. Seema is older than Ritesh.
II. Suresh is older than Seema.
III. Ritesh is older than Suresh.
If the first two statements are true, the third statement is

a) False
b) True
c) Uncertain

10) If January 1, 1996, was Monday, what day of the week was January 2, 1997?

a) Thursday
b) Wednesday
c) Friday
d) Sunday

CHECK YOUR SCORE

1) d
2) c
3) c
4) d
5) a

6) a
7) b
8) d
9) a
10) a

Shweta Seth
2027661

MEET OUR CREW

Shweta Seth
2027661

Komal Nagarajan
2027654

Anchal Aneja
2027245

Aishwarya Jayakumar
2028150

Shiv Kumar Patil
2027564

Ishita Bhatnagar
2027541

Vishnu C R
2027115