Tweet Insight

Sandeep Naidu Mudili
smudili2@illinois.edu
University of Illinois
Urbana-Champaign, Illinois, USA

Taraka Vignesh Mullapudi
tvm3@illinois.edu
University of Illinois
Urbana-Champaign, Illinois, USA

Sai Venkata Abhijith Geda
geda2@illinois.edu
University of Illinois
Urbana-Champaign, Illinois, USA

Venkata Sai Ashrith Kona
kona2@illinois.edu
University of Illinois
Urbana-Champaign, Illinois, USA

ABSTRACT

Tweet Insight is an application designed to analyze and categorize social media content, with a focus on Twitter. It uses Latent Dirichlet Allocation (LDA) for topic modeling, identifying distinct topics within tweets and clustering them into broader categories. Data preprocessing, including punctuation removal, stopword filtering, lemmatization, and stemming, ensures high-quality input for the LDA model. The application also incorporates sentiment analysis via the SentimentIntensityAnalyzer, classifying tweets as positive, neutral, or negative. This combined approach provides a comprehensive view of social media content, enabling detailed insights into trends, topic distribution, and sentiment. Visualization through dashboards and Power BI allows users to interact with the data and gain actionable insights, useful for business intelligence, market research, and social behavior studies. Overall, Tweet Insight offers a robust framework for extracting valuable information from social media platforms, supporting data-driven decision-making.

KEYWORDS

Tweet Analysis, Social Media Analytics, Latent Dirichlet Allocation (LDA), Topic Modeling, Sentiment Analysis, Data Preprocessing, Data Visualization, Natural Language Processing (NLP), Business Intelligence, Social Behavior Analysis, Twitter Data.

ACM Reference Format:
Sandeep Naidu Mudili, Taraka Vignesh Mullapudi, Sai Venkata Abhijith Geda, and Venkata Sai Ashrith Kona. 2024. Tweet Insight. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Tweet Insight is a data analysis application designed to explore and categorize content from social media, focusing primarily on Twitter. As the volume of social media data continues to grow exponentially, businesses, researchers, and analysts require tools that can efficiently process this information to uncover trends, sentiments, and underlying themes. Tweet Insight meets this need by employing advanced natural language processing (NLP) techniques, including Latent Dirichlet Allocation (LDA), to identify and cluster topics from large datasets.

The application begins by preprocessing raw data to remove noise, such as punctuation and stopwords, ensuring a clean dataset for analysis. Once preprocessed, the LDA model is applied to extract key topics, which are subsequently clustered into broader thematic categories. Given the diversity of Twitter content, each topic can be associated with multiple clusters, allowing for a more flexible and comprehensive analysis.

In addition to topic modeling, Tweet Insight incorporates sentiment analysis through SentimentIntensityAnalyzer, enabling a deeper understanding of public opinion by classifying tweets as positive, neutral, or negative. This dual approach—combining topic modeling and sentiment analysis—offers users a multifaceted perspective on social media trends.

The insights derived from Tweet Insight are visualized using interactive dashboards and Power BI, allowing users to explore the data and uncover actionable insights. This makes Tweet Insight a valuable tool for business intelligence, market research, and social behavior analysis, providing a robust framework for data-driven decision-making in a rapidly evolving social media landscape.

Beyond business applications, Tweet Insight offers significant value to researchers and sociologists by providing insights into public discourse, emerging trends, and social movements. Policymakers can use this tool to gauge public opinion on critical issues, while news organizations can leverage it to understand audience sentiment around current events. Additionally, the application can support academic research by enabling studies on social behavior, communication patterns, and the spread of information or misinformation across social media platforms.

By encompassing a range of usages, Tweet Insight is positioned as a versatile and adaptable tool, serving a diverse set of needs in an era where social media plays a pivotal role in shaping public

2 METHODOLOGY

In this section, we describe the approach used to develop and implement Tweet Insight, focusing on the key steps involved in data collection, preprocessing, model training, and analysis. This section provides an overview of the processes that enable the application to analyze and categorize social media content, with an emphasis

Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

on topic modeling and sentiment analysis. Each subsection covers a specific stage of the workflow, outlining the techniques and tools used to ensure accurate and meaningful results.

2.1 Data Collection

Data collection for "Tweet Insight" involved obtaining large-scale datasets from Kaggle, focusing on datasets that contain tweets from various contexts. We used a combination of multiple Kaggle datasets to ensure diversity and breadth, totaling approximately 5 million tweets. This large volume of data provided a robust foundation for training our topic modeling and sentiment analysis processes.

The datasets contained tweets from different time periods, geographies, and subject matters, contributing to a comprehensive representation of Twitter content. By using multiple datasets, we aimed to minimize biases and enhance the generalizability of our findings. Each dataset was carefully examined to ensure data quality and integrity, checking for completeness, duplicate entries.

2.2 Preprocessing of Data

Preprocessing is a critical step in preparing raw text data for analysis, ensuring that the input for the Latent Dirichlet Allocation (LDA) model is clean, normalized, and free of noise. This section outlines the process used to preprocess the 5 million tweets in our dataset, detailing the operations carried out to transform the raw text into a suitable format.

The preprocessing begins with the removal of punctuation and unwanted characters. Using a regular expression pattern, any character not in the alphanumeric set (letters and digits) or whitespace is removed. This ensures that only valid words remain in the text, reducing potential distractions for the LDA model.

Next, stopwords are removed to focus on meaningful content. The stopword list is compiled from multiple sources, including the standard list from the Natural Language Toolkit (nltk) and a custom list from a JSON file. By unionizing these sets, we create a comprehensive collection of common words that do not contribute significant semantic information.

Following stopword removal, the text is normalized through lemmatization. Lemmatization reduces words to their base or root form, allowing the model to recognize different inflections of the same word as equivalent. We use the WordNetLemmatizer from nltk to accomplish this, which helps improve topic modeling accuracy.

Stemming is the final step, where each word is further reduced to its stem, providing a simplified version of the text. The PorterStemmer is employed to achieve this, further refining the input for analysis. This preprocessed text is then ready for use in the LDA model and subsequent analysis steps.

By thoroughly preprocessing the data, we ensure a high-quality input for topic modeling and sentiment analysis, leading to more reliable and coherent results from Tweet Insight.

2.3 Training the LDA Model

Training the Latent Dirichlet Allocation (LDA) model is a key step in topic modeling for Tweet Insight. Using the preprocessed data, we employed the LDA algorithm to extract latent topics from the dataset of 5 million tweets. This process involves configuring the model with various parameters to optimize its performance in terms of the number of topics to generate and the number of words to represent each topic.

To achieve the best results, we trained the LDA model multiple times, adjusting the number of topics and the number of words per topic to find the ideal combination. Each iteration of the model was analyzed for coherence and relevance, with attention given to the clarity of the topics and the meaningfulness of the words within each topic. We evaluated whether the topics made sense in terms of clustering and thematic grouping.

After several iterations, the optimal configuration was found with 20 topics and 8 words in each topic. This configuration provided a clear structure to the topics, allowing for meaningful clustering and analysis. The selection of 20 topics struck a balance between granularity and generalization, enabling effective categorization of the diverse content in the tweets. The choice of 8 words per topic provided sufficient context to understand the main theme without introducing too much complexity.

Once the ideal LDA model was trained, we saved it for future use, ensuring consistency across different analyses. This saved model can be applied to new data, enabling continued topic extraction and categorization without re-training the LDA from scratch.

The flexibility of the LDA model allows for further adjustments if needed, but the current configuration has proven effective in capturing relevant topics and providing a solid foundation for clustering and sentiment analysis in Tweet Insight.

2.4 Using the Trained Model on New Data

After training the Latent Dirichlet Allocation (LDA) model and achieving the ideal configuration with 20 topics and 8 words per topic, we proceeded to use this trained model on new data. This section describes the process of collecting new tweets through Twitter APIs, preprocessing the data, and applying the trained LDA model to classify the tweets into topics.

First, we wrote Twitter API scripts to collect tweets from the news feeds of different accounts, focusing on a diverse range of users to capture various perspectives and topics. These scripts gathered approximately 500,000 tweets, which were then stored in an Excel file. The Excel sheet was structured with two primary columns: 'username' for the Twitter handle of the user, and 'tweet-content' for the text of the tweet.

Before applying the LDA model, the tweets were preprocessed to remove noise and standardize the text. This preprocessing followed the same steps outlined in the earlier section, including converting text to lowercase, removing punctuation and stopwords, and applying lemmatization and stemming. This ensured that the data fed into the LDA model was consistent and suitable for topic classification.

With the cleaned data, we used the trained LDA model to classify each tweet into one of the 20 topics. The model outputs a probability distribution across all topics for each tweet, indicating which topics are most relevant. The tweet is then assigned to the topic with the highest probability, allowing us to categorize it accordingly.

The classified data was then written back to the Excel file, with an additional column indicating the assigned topic for each tweet. This output file provides a clear record of the username, tweet content, and topic classification, facilitating further analysis and visualization. The use of Excel ensures compatibility with various data analysis tools and platforms.

By applying the trained LDA model to new data, we created a process for continuously updating our analysis with fresh content from Twitter. This approach allows for ongoing monitoring of social media trends and topics, contributing to a dynamic and flexible framework for tweet categorization in Tweet Insight.

2.5 Clustering and sentiment Analysis

After classifying tweets into topics using the Latent Dirichlet Allocation (LDA) model, the next step is to cluster these topics into broader categories and perform sentiment analysis. This subsection outlines the process of topic clustering and sentiment analysis, leading to the generation of additional data for visualization and further analysis.

2.5.1 ***Topic Clustering.*** Once the LDA model was trained, the resulting topics were assigned to one or more of the following five clusters:

(1) **Emotions and Expressions**
(2) **Daily Life and Activities**
(3) **Relationships and Social Interactions**
(4) **Negative Sentiments and Anger**
(5) **Changes and Transitions**

Each topic was analyzed to determine which cluster(s) it belonged to. This was done by examining the key words and context in each topic, allowing for accurate categorization. Given the flexibility of this system, a topic could belong to multiple clusters if its content was relevant to more than one category.

2.5.2 ***Sentiment Analysis.*** Following topic clustering, sentiment analysis was conducted for each tweet. Using a built in library `SentimentIntensityAnalyzer` from `nltk.sentiment.vader`, we classified tweets as **positive**, **neutral**, or **negative**. This sentiment score offered additional context for each tweet, providing a deeper understanding of the overall mood and social dynamics within the data.

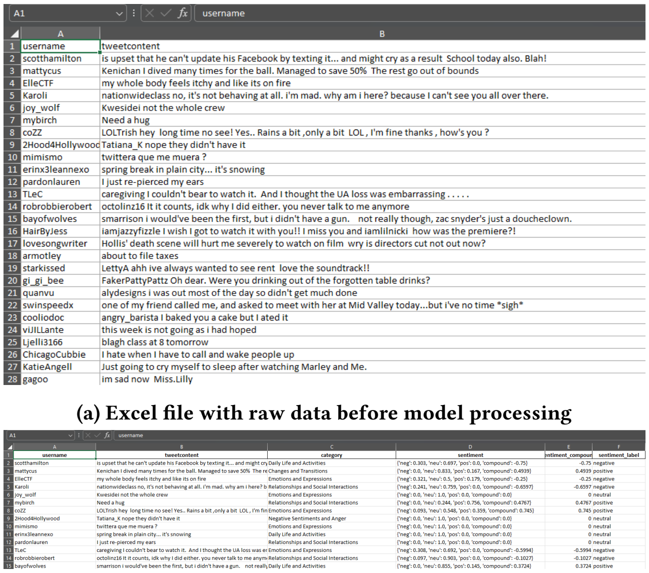
2.5.3 ***Data Sheet Augmentation.*** To integrate the results of clustering and sentiment analysis, we added two new columns to the Excel output data sheet:

- **Cluster**: The cluster to which the tweet's topic belongs.
- **Sentiment**: The sentiment classification (positive, neutral, or negative).

This augmentation allows for a comprehensive view of each tweet, combining the assigned cluster and its sentiment analysis. This structure facilitates further visualization and analysis, forming a foundation for exploring social media trends and topic correlations.

2.6 Visualization

Upon completion of the code execution, an excel file is created automatically with the output data. The input data file initially contains raw tweet data, containing user handles and tweet content. Additional columns indicating the assigned topic, sentiment classification, and topic clusters are appended after processing with our LDA model and sentiment analysis. Below, we present to figures: Figure X shows a snapshot of the excel file containing raw data before model processing. Figure Y shows a snapshot of the excel file which is generated post processing.



A1 | username

	A	B
1	username	tweetcontent
2	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah!
3	mattycus	Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
4	ElleCTF	my whole body feels itchy and like its on fire
5	Karoli	nationwideclass no, it's not behaving at all. i'm mad. why am i here? because I can't see you all over there.
6	joy_wolf	Kwesidei not the whole crew
7	mybirch	Need a hug
8	coZZ	LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fine thanks , how's you ?
9	2Hood4Hollywooc	Tatiana_K nope they didn't have it
10	mimismo	twittera que me muera ?
11	erinx3leannexo	spring break in plain city... it's snowing
12	pardonlauren	I just re-pierced my ears
13	TLeC	caregiving I couldn't bear to watch it. And I thought the UA loss was embarrassing . . . . .
14	robrobbierobert	octolinz16 It it counts, idk why I did either. you never talk to me anymore
15	bayofwolves	smarrison i would've been the first, but i didn't have a gun. not really though, zac snyder's just a doucheclown.
16	HairByJess	iamjazzyfizzle I wish I got to watch it with you!! I miss you and iamlilnicki how was the premiere?!
17	lovesongwriter	Hollis' death scene will hurt me severely to watch on film wry is directors cut not out now?
18	armotley	about to file taxes
19	starkissed	LettyA ahh ive always wanted to see rent love the soundtrack!!
20	gi_gi_bee	FakerPattyPattz Oh dear. Were you drinking out of the forgotten table drinks?
21	quanvu	alydesigns i was out most of the day so didn't get much done
22	swinspeedx	one of my friend called me, and asked to meet with her at Mid Valley today...but i've no time *sigh*
23	cooliodoc	angry_barista I baked you a cake but I ated it
24	viJILLante	this week is not going as i had hoped
25	Ljelli3166	blagh class at 8 tomorrow
26	ChicagoCubbie	I hate when I have to call and wake people up
27	KatieAngell	Just going to cry myself to sleep after watching Marley and Me.
28	gagoo	im sad now Miss.Lilly

(a) Excel file with raw data before model processing

A1 | username

	A	B	C	D	E	F
1	username	tweetcontent	category	sentiment	entiment_compoun	sentiment_label
2	scotthamilton	is upset that he can't update his Facebook by texting it... and might cry	Daily Life and Activities	{'neg': 0.303, 'neu': 0.697, 'pos': 0.0, 'compound': -0.75}	-0.75	negative
3	mattycus	Kenichan I dived many times for the ball. Managed to save 50% The re	Changes and Transitions	{'neg': 0.0, 'neu': 0.833, 'pos': 0.167, 'compound': 0.4939}	0.4939	positive
4	ElleCTF	my whole body feels itchy and like its on fire	Emotions and Expressions	{'neg': 0.321, 'neu': 0.5, 'pos': 0.179, 'compound': -0.25}	-0.25	negative
5	Karoli	nationwideclass no, it's not behaving at all. i'm mad. why am i here? b	Relationships and Social Interactions	{'neg': 0.241, 'neu': 0.759, 'pos': 0.0, 'compound': -0.6597}	-0.6597	negative
6	joy_wolf	Kwesidei not the whole crew	Emotions and Expressions	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0	neutral
7	mybirch	Need a hug	Relationships and Social Interactions	{'neg': 0.0, 'neu': 0.244, 'pos': 0.756, 'compound': 0.4767}	0.4767	positive
8	coZZ	LOLTrish hey long time no see! Yes.. Rains a bit ,only a bit LOL , I'm fin	Emotions and Expressions	{'neg': 0.093, 'neu': 0.548, 'pos': 0.359, 'compound': 0.745}	0.745	positive
9	2Hood4Hollywood	Tatiana_K nope they didn't have it	Negative Sentiments and Anger	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0	neutral
10	mimismo	twittera que me muera ?	Emotions and Expressions	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0	neutral
11	erinx3leannexo	spring break in plain city... it's snowing	Daily Life and Activities	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0	neutral
12	pardonlauren	I just re-pierced my ears	Relationships and Social Interactions	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0	neutral
13	TLeC	caregiving I couldn't bear to watch it. And I thought the UA loss was er	Emotions and Expressions	{'neg': 0.308, 'neu': 0.692, 'pos': 0.0, 'compound': -0.5994}	-0.5994	negative
14	robrobbierobert	octolinz16 It it counts, idk why I did either. you never talk to me anym	Relationships and Social Interactions	{'neg': 0.097, 'neu': 0.903, 'pos': 0.0, 'compound': -0.1027}	-0.1027	negative
15	bayofwolves	smarrison i would've been the first, but i didn't have a gun. not really	Daily Life and Activities	{'neg': 0.0, 'neu': 0.855, 'pos': 0.145, 'compound': 0.3724}	0.3724	positive

(b) Excel file generated post processing

Figure 1: Comparison of raw data and processed data

The provided visualization, compiled using Power BI, well illustrates the sentiment analysis and topic classification results from Tweet Insight's processing. The multi component visualization provides a comprehensive view of the data by integrating multiple graphical elements to illustrate the sentiment analysis across tweet categories.

Figure Z here presents a bar chart measuring sentiments (positive, negative, and neutral) in relation to various categories like Changes and Transitions, Daily Life and Activities, Emotions and Expressions, Negative Sentiments and Anger, and Relationships and Social Interactions is displayed in the upper section.

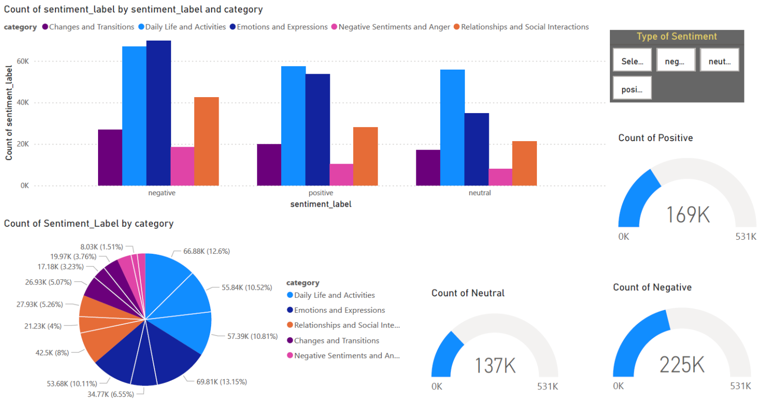


Figure 2: Multi Component Visualization Figure Z

A pie chart showing the distribution of tweets in each of the categories is located beneath the bar chart as can be seen in Figure Z. Each slice is color coded to match the categories in the above chart and is proportionately sized based on the volume of tweets in that category. This gives an indication of the categories that are more common in the dataset visually.

Moreover, three gauges on the right side in Figure Z give a brief overview of the total number of tweets categorized as positive, neutral, and negative. These gauges, which also display the total number of tweets along with the count for each sentiment, are helpful for quickly understanding the overall sentiment landscape in the dataset.

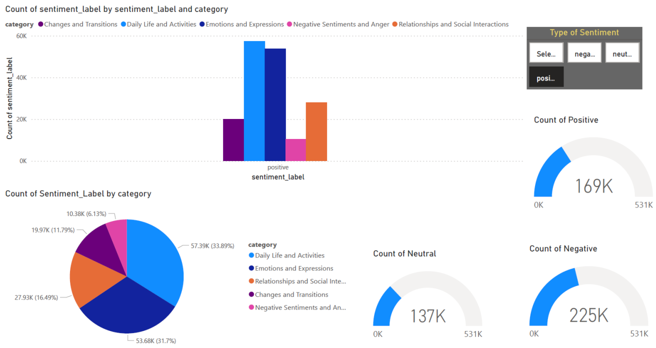

Count of sentiment_label by sentiment_label and category
category ● Changes and Transitions ● Daily Life and Activities ● Emotions and Expressions ● Negative Sentiments and Anger ● Relationships and Social Interactions
60K
40K
20K
0K
Count of sentiment_label
positive
sentiment_label
Type of Sentiment
Sele...
nega...
neut...
posi...
Count of Positive
169K
0K
531K
Count of Sentiment_Label by category
10.38K (6.13%)
19.97K (11.79%)
57.39K (33.89%)
27.93K (16.49%)
53.68K (31.7%)
category
● Daily Life and Activities
● Emotions and Expressions
● Relationships and Social Inte...
● Changes and Transitions
● Negative Sentiments and An...
Count of Neutral
137K
0K
531K
Count of Negative
225K
0K
531K

Figure 3: Positive Sentiment Visualization Fig Z_a

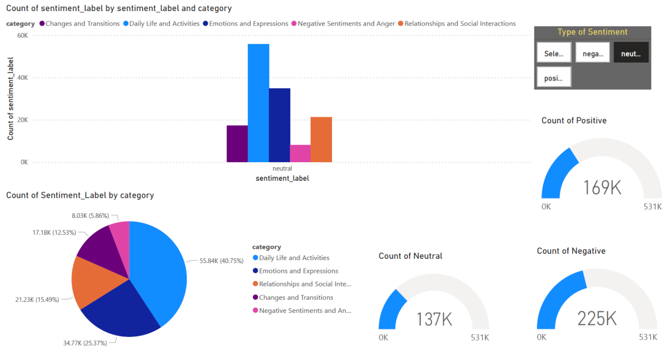

Count of sentiment_label by sentiment_label and category
category ● Changes and Transitions ● Daily Life and Activities ● Emotions and Expressions ● Negative Sentiments and Anger ● Relationships and Social Interactions
60K
40K
20K
0K
Count of sentiment_label
neutral
sentiment_label
Type of Sentiment
Sele...
nega...
neut...
posi...
Count of Positive
169K
0K
531K
Count of Sentiment_Label by category
8.03K (5.86%)
17.18K (12.53%)
55.84K (40.75%)
21.23K (15.49%)
34.77K (25.37%)
category
● Daily Life and Activities
● Emotions and Expressions
● Relationships and Social Inte...
● Changes and Transitions
● Negative Sentiments and An...
Count of Neutral
137K
0K
531K
Count of Negative
225K
0K
531K

Figure 4: Neutral Sentiment Visualization Fig Z_b

A key feature of this dashboard is the interactive filter labeled "Type of Sentiment" which lets users refine display depending on a selected sentiment type. As observed in figure a, b, c we can observe the analysis specific to the sentiment positive, neutral and negative respectively. By enabling customized analyses that concentrate on specific areas of interest, this feature improves dashboard's usefulness and helps stakeholders make decisions based on targeted insights.

This suite of visualizations is an effective tool for analyzing how tweet categories and sentiment interact, highlighting possible patterns and insights that can be used to improve sociological studies, business intelligence and market research.

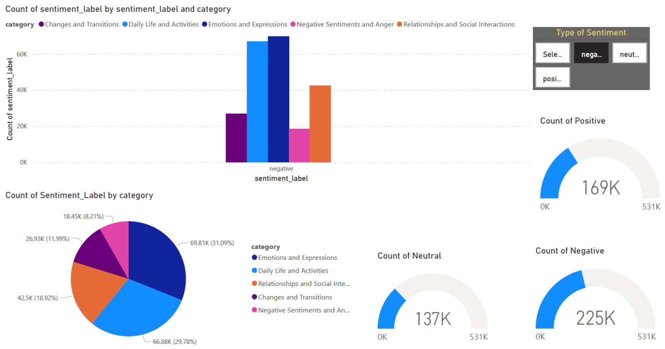

Count of sentiment_label by sentiment_label and category
category ● Changes and Transitions ● Daily Life and Activities ● Emotions and Expressions ● Negative Sentiments and Anger ● Relationships and Social Interactions
60K
40K
20K
0K
Count of sentiment_label
negative
sentiment_label
Type of Sentiment
Sele...
nega...
neut...
posi...
Count of Positive
169K
0K
531K
Count of Sentiment_Label by category
18.45K (8.21%)
26.93K (11.99%)
69.81K (31.09%)
42.5K (18.92%)
66.88K (29.78%)
category
● Emotions and Expressions
● Daily Life and Activities
● Relationships and Social Inte...
● Changes and Transitions
● Negative Sentiments and An...
Count of Neutral
137K
0K
531K
Count of Negative
225K
0K
531K

Figure 5: Negative Sentiment Visualization Fig Z_c

3 MODEL WORKFLOW

The model workflow for Tweet Insight illustrates the sequence of steps taken to analyze social media content, focusing on Twitter data. This section describes the flow from data collection and preprocessing to topic modeling, clustering, sentiment analysis, and visualization. The workflow provides a comprehensive view of the process used to transform raw tweets into meaningful insights.

The workflow begins with collecting data from Twitter APIs. This involves capturing tweets from various accounts and saving them to an Excel file. The raw tweets are then preprocessed to remove noise, such as punctuation and stopwords, and normalize the text using lemmatization and stemming.

Next, the preprocessed data is fed into the trained Latent Dirichlet Allocation (LDA) model to classify each tweet into one of the predefined topics. Once classified, the topics are clustered into broader categories for better context. Sentiment analysis is applied to each tweet to determine whether the sentiment is positive, neutral, or negative. The results are added to the data sheet, which is then used to create visualizations that offer insights into social media trends and patterns.

3.1 Pseudocode

Below is the pseudocode representing the core steps in the workflow for Tweet Insight. This pseudocode outlines the major operations and their sequence, providing a simplified representation of the entire process.

Step 1: Data Collection Collect data from Twitter APIs Store tweets in Excel file with columns 'username' and 'tweetcontent'

Step 2: Data Preprocessing each tweet in the Excel file Convert text to lowercase Remove punctuation and stopwords Apply lemmatization and stemming

Step 3: Apply Trained LDA Model Load the pre-trained LDA model each cleaned tweet Classify the tweet into one of the LDA topics Assign the appropriate cluster based on topic classification Add the cluster assignment to the Excel file

Step 4: Sentiment Analysis Initialize SentimentIntensityAnalyzer each tweet in the Excel file Perform sentiment analysis Classify the sentiment as positive, neutral, or negative Add the sentiment classification to the Excel file

Step 5: Data Sheet Augmentation Augment the Excel file with new columns for 'Cluster' and 'Sentiment'

Step 6: Visualization Load the augmented Excel file into visualization tools (e.g., Power BI)

4 EVALUATION AND RESULTS

To evaluate the accuracy of the Tweet Insight model, our team conducted a manual classification exercise to serve as a ground truth for comparison with the model's outputs. Each of the four team members manually classified 50 tweets (a total of 200) into one of the five predefined clusters and determined the sentiment for each tweet. The clusters were "Emotions and Expressions," "Daily Life and Activities," "Relationships and Social Interactions," "Negative Sentiments and Anger," and "Changes and Transitions." The sentiments were categorized as positive, neutral, or negative.

After the manual classification, the results were compared with the output from the trained model to determine accuracy levels in clustering and sentiment analysis.

4.1 Clustering Accuracy

To evaluate clustering accuracy, we compared the model's topic classifications with the manual ground truth. The model achieved a 91% accuracy rate in clustering, indicating that the Latent Dirichlet Allocation (LDA)-based topic modeling and subsequent clustering approach were effective in capturing the thematic structure of the tweets.

4.2 Sentiment Analysis Accuracy

The SentimentIntensityAnalyzer was used to classify each tweet into positive, neutral, or negative. By comparing the model's output with manual classification, we found that the model achieved a 93% accuracy rate in sentiment analysis. This high accuracy demonstrates that the sentiment analysis component is reliable and can effectively capture the mood of the tweets.

4.3 Discussion

These results suggest that Tweet Insight is a robust application for analyzing social media content. With a 91% accuracy in topic clustering and 93% in sentiment analysis, the model proves effective in categorizing tweets into meaningful clusters and assessing sentiment. Given the subjective nature of manual classification, these accuracy rates reflect the reliability of the approach.

5 CONCLUSIONS AND FUTURE SCOPE

The "Tweet Insight" project aimed to analyze and categorize Twitter content using topic modeling and sentiment analysis. Through the use of Latent Dirichlet Allocation (LDA) for topic extraction and SentimentIntensityAnalyzer for sentiment analysis, the model demonstrated a high degree of accuracy in classifying tweets into predefined clusters and assessing sentiment. With a 91% accuracy rate for clustering and a 93% accuracy rate for sentiment analysis, the results suggest that the approach is robust and effective.

One of the significant conclusions from this project is that the workflow and methodology used for "Tweet Insight" can be applied to other social media platforms and text-based data sources. The underlying techniques of data preprocessing, topic modeling, and sentiment analysis are not limited to Twitter, allowing for broader applications across different types of social media and text content.

The versatility of this approach means that it can be adapted for various purposes, such as business intelligence, market research, social behavior studies, and public opinion analysis. By applying these techniques to other platforms, such as Facebook, Instagram, or LinkedIn, we can gain insights into different demographics, industries, and social trends.

However, there is always room for improvement and further exploration. Future work could focus on the following:

- **Expanding to Other Social Media Platforms**: Apply the model to different social media platforms to validate its effectiveness across various contexts.
- **Improving Clustering Accuracy**: Refine the clustering approach to improve accuracy, especially for topics with overlapping or ambiguous content.
- **Enhancing Sentiment Analysis**: Explore additional sentiment analysis techniques to capture more nuanced emotions and sentiments in text.
- **Handling Ambiguous Content**: Develop strategies to manage tweets or posts with mixed or unclear meanings to improve classification accuracy.
- **Scaling to Larger Datasets**: Test the model with larger and more diverse datasets to ensure scalability and robustness.

Overall, the Tweet Insight project demonstrated that the combination of topic modeling and sentiment analysis provides a powerful framework for social media analysis. By building on these results and exploring new avenues for application, we can continue to advance the field of social media analytics and derive meaningful insights from vast amounts of text-based data.

6 CONTRIBUTIONS

The success of the "Tweet Insight" project relied on the collaboration and diverse skill sets of our team members. Each individual brought unique talents and expertise to the project, contributing to its overall effectiveness and high accuracy rates.

6.1 Team Member Contributions

- **Taraka Vignesh Mullapudi (tvm3)**: Taraka was responsible for data fetch using APIs, collecting large datasets from Kaggle, and managing data preprocessing. His expertise in API integration and data manipulation ensured a steady flow of data for the project.
- **Sai Venkata Abhijith Geda (geda2)**: Sai took the lead on visualizing results, error analysis, and testing. His attention to detail and strong analytical skills helped identify discrepancies and ensure the accuracy of the model's output. His work on visualizations made it easier to understand and interpret the results.
- **Venkata Sai Ashrith Kona (kona2)**: Ashrith focused on sentiment analysis and report writing. His deep understanding of sentiment analysis techniques and proficiency in technical writing contributed to the project's comprehensive documentation and sentiment analysis accuracy.
- **Sandeep Naidu Mudili (smudili2)**: Sandeep was responsible for model training and clustering. His expertise in machine learning and topic modeling was crucial in training

the Latent Dirichlet Allocation (LDA) model and clustering the topics into broader categories.

6.2 Collaboration and Unique Talents

The diverse talents and collaborative spirit of the team were key factors in the project's success. Each team member brought a unique perspective, allowing the project to benefit from a variety of skills and experiences. The combination of data science, machine learning, sentiment analysis, and visualization expertise enabled the team to address complex challenges and achieve high accuracy in clustering and sentiment analysis.

Overall, the collaborative approach and diverse skill set of the team ensured that the "Tweet Insight" project was completed successfully, delivering meaningful insights and a solid framework for future work in social media analysis.

7 SOURCE CODE AND PRESENTATION VIDEO

- Source code of the **Tweet Insight** project is available on GitHub: **Team 15_CS 510 Project GitHub Link**
- Comprehensive slides detailing the **Tweet Insight** project can be found here: **Team 15_CS510_FinalProject_PPT**
- For a detailed video explanation of the **Tweet Insight** project, please visit: **CS 510_Team15_TweetInsight_Video**

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Languages in Social Media*, Department of Computer Science, Columbia University, New York, 2009.

[2] Mandava Geetha Bhargava, Kuppa Tara Phani Surya Kiran, & Duvvada Rajeswara Rao, "Analysis and Design of Visualization of Educational Institution database using Power BI Tool," *Global Journal of Computer Science and Technology*, vol. 18, no. C4, pp. 1–8, 2018. Available: https://computerresearch.org/index.php/computer/article/view/1776

[3] Farkhod, A.; Abdusalomov, A.; Makhmudov, F.; & Cho, Y.I., "LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model," *Applied Sciences*, vol. 11, article 11091, 2021. DOI: 10.3390/app112311091

[4] M. Wankhade, A.C.S. Rao, & C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, pp. 5731–5780, 2022. DOI: 10.1007/s10462-022-10144-1

[5] X. Fang & J. Zhan, "Sentiment analysis using product review data," *Journal of Big Data*, vol. 2, article 5, 2015. DOI: 10.1186/s40537-015-0015-2