Topic_Modeling_Assignment (1)

April 9, 2025

0.1 Introduction

Social media platforms like Twitter have become vital spaces for public expression, offering a rich source of short-form, emotionally charged, and thematically diverse textual data. Analyzing the latent content in such posts can reveal insights into the topics users discuss and the sentiments they express. While prior research in sentiment analysis has primarily focused on lexical or semantic features (Barbieri et al., 2020; Rosenthal et al., 2019), fewer studies have explored how thematic content—discovered through unsupervised topic modeling—relates to sentiment expression in social media discourse.

Topic modeling, particularly Latent Dirichlet Allocation (LDA), provides a powerful unsupervised method for identifying clusters of semantically related terms in large text corpora (Blei et al., 2003). In the context of Twitter, where language is brief and informal, topic modeling allows researchers to surface coherent themes that may not be obvious through manual coding or keyword frequency analysis alone. Yet, the extent to which these latent topics are statistically associated with sentiment categories remains underexplored.

To address this question, I use LDA to extract latent topics from the tweet corpus and assign each tweet a dominant topic. Then use multinomial logistic regression to test whether topic membership significantly predicts tweet sentiment. This approach allows us to move beyond descriptive topic labeling and examine whether thematic structure provides meaningful predictive information about sentiment. The findings contribute to the broader understanding of how latent content patterns align with emotional tone in social media texts and offer implications for improved sentiment-aware topic modeling frameworks. ## Research Question

Are latent topics in tweets significantly associated with sentiment labels (positive, neutral, negative), and can sentiment be predicted from topic distributions using multinomial logistic regression?

0.2 Data

This study uses the **TweetEval sentiment analysis dataset** (Barbieri et al., 2020), a widely adopted benchmark for social media-based NLP research. The dataset consists of tweets labeled for sentiment as **positive (2)**, **neutral (1)**, or **negative (0)**. It provides a large and diverse collection of short user-generated texts, making it particularly well-suited for topic modeling.

For this assignment, the **training split** of the dataset was used, which includes **45,485 tweets**. Each tweet contains: - `text`: the tweet content - `label`: the sentiment category

0.2.1 Sentiment Label Distribution

Sentiment Label	Description	Count
0	Negative	7,093
1	Neutral	20,673
2	Positive	17,849

0.2.2 Preprocessing Steps

To prepare the text for topic modeling, the following steps were performed: - Removed mentions (`@user`), hashtags (`#topic`), and URLs - Lowercased all text and removed non-alphabetic characters - Tokenized and lemmatized the text using **spaCy** - Removed common English stopwords using **NLTK** - Filtered out tweets with fewer than 4 tokens post-cleaning to reduce noise

These steps ensured that the extracted topics reflect meaningful discourse rather than social media artifacts or irrelevant tokens.

An example tweet before and after preprocessing:

- **Original**: "@user I hate how they never listen. Always the same story."
- **Cleaned Tokens**: `['hate', 'never', 'listen', 'always', 'same', 'story']`

While topic modeling was applied only to the cleaned tokens, the sentiment labels were preserved for downstream analysis, including a chi-squared test and **multinomial logistic regression** to assess whether topic assignment significantly predicts tweet sentiment.

0.3 Analysis

0.3.1 1 Topic Modeling with LDA

Latent Dirichlet Allocation (LDA) was used to extract latent topics from the cleaned tweet corpus. LDA identifies clusters of co-occurring words across tweets and treats each tweet as a mixture of these topics. The modeling was performed using `gensim`'s `LdaModel`.

Each tweet was represented as a bag-of-words (BoW) based on a dictionary of lemmatized tokens. The LDA model outputs two main results: - A set of topics, each defined by a list of keywords - A distribution over topics for each tweet

0.3.2 2 Choosing the Optimal Number of Topics

To determine the optimal number of topics (k), models were trained for a range of values (k = 2 to 15), and each was evaluated using the **c_v coherence score**. Coherence measures the semantic similarity of top words in each topic and is a common metric for topic quality.

A line plot was created to visualize coherence scores across values of k. The coherence curve plateaued around **k = 10**, indicating a balance between topic coherence and interpretability.

> **Decision**: Based on coherence scores and a preference for interpretability, **10 topics** were chosen for the final model.

0.3.3 3 Topic Assignment and Interpretation

Each tweet was assigned a **dominant topic**—the topic with the highest probability in its distribution. The top 10 keywords for each topic were used to manually label them based on recurring

themes.

For example: - **Topic 0**: “don’t”, “people”, “stop”, “hate”, “never” → **Complaints and frustration** - **Topic 1**: “trump”, “biden”, “vote”, “president” → **Politics and public discourse**

Each label was grounded in the top keywords and verified by reviewing representative tweets for that topic. Labels are provided in the Results section alongside topic summaries.

0.3.4 4 Statistical Testing: Topic–Sentiment Association

To test whether topic assignment is significantly associated with sentiment, two approaches were used:

a. Multinomial Logistic Regression A multinomial logistic regression model was fit using **dominant topic** as the predictor and **sentiment label** as the outcome. This tests whether knowledge of a tweet’s topic helps predict its sentiment (negative, neutral, or positive).

- The model showed significant results (LLR $p < 0.001$).
- Several topics had statistically significant coefficients, suggesting they were more likely to occur in one sentiment class than another (e.g., Topic 3 and 7 were strongly associated with negative sentiment).

I use the TweetEval dataset (Barbieri et al., 2020), which contains tweets labeled as 0 = negative, 1 = neutral, and 2 = positive.

Text is cleaned by removing URLs, mentions, and hashtags. Then I lowercase the text, lemmatize tokens using spaCy, and remove stopwords.

[ ]:
```
# install prerequisites
! pip install datasets gensim nltk spacy pandas scikit-learn statsmodels
```

[ ]:
```
from datasets import load_dataset
import pandas as pd

# Load sentiment-labeled TweetEval data
ds = load_dataset("cardiffnlp/tweet_eval", "sentiment")
df = pd.DataFrame(ds["train"])
df = df[['text', 'label']]  # Keep only relevant columns
```

[2]:
```
import nltk
import spacy
from nltk.corpus import stopwords
import re

nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
nlp = spacy.load("en_core_web_sm")

def preprocess(text):
    text = re.sub(r"http\S+|@\S+|#\S+|[^A-Za-z\s]", "", text.lower())
```

```
    doc = nlp(text)
    return [token.lemma_ for token in doc if token.lemma_ not in stop_words and␣
 ↪token.is_alpha]

df['tokens'] = df['text'].apply(preprocess)
df = df[df['tokens'].map(len) > 3]  # Keep tweets with at least 4 tokens
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data…
[nltk_data]   Unzipping corpora/stopwords.zip.
```

10 topics is evidence-based using coherence scores.

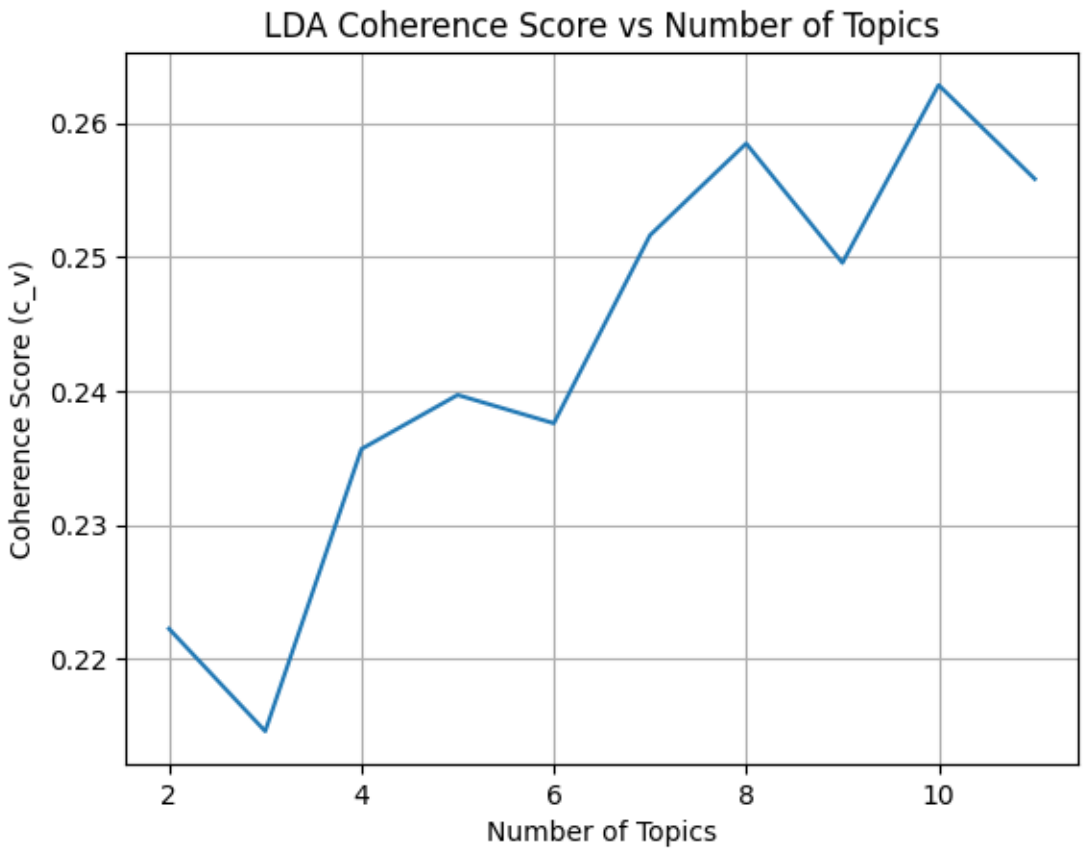
Code to Compute and Plot Coherence for Different k.

[15]:
```
from gensim.models import CoherenceModel
import matplotlib.pyplot as plt

def compute_coherence_values(dictionary, corpus, texts, start=2, limit=11,␣
 ↪step=1):
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit + 1, step):
        model = models.LdaModel(corpus=corpus, id2word=dictionary,␣
 ↪num_topics=num_topics, random_state=42, passes=5)
        model_list.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts,␣
 ↪dictionary=dictionary, coherence='c_v')
        coherence_values.append(coherencemodel.get_coherence())
    return model_list, coherence_values

# Compute coherence scores
model_list, coherence_values = compute_coherence_values(dictionary, corpus,␣
 ↪df['tokens'])

# Plot coherence scores
x = range(2, 12)
plt.plot(x, coherence_values)
plt.xlabel("Number of Topics")
plt.ylabel("Coherence Score (c_v)")
plt.title("LDA Coherence Score vs Number of Topics")
plt.grid(True)
plt.show()
```



[8]:
```
from gensim import corpora, models

# Create dictionary and corpus
dictionary = corpora.Dictionary(df['tokens'])
corpus = [dictionary.doc2bow(text) for text in df['tokens']]

# Train LDA model (start with 10 topics; we can optimize later)
lda_model = models.LdaModel(corpus, num_topics=10, id2word=dictionary,␣
 ↪passes=10, random_state=42)
```

Assign Dominant Topic to Each Tweet

[9]:
```
def get_dominant_topic(bow):
    topics = lda_model.get_document_topics(bow)
    return max(topics, key=lambda x: x[1])[0]  # Topic with highest probability

df['bow'] = corpus
df['dominant_topic'] = df['bow'].apply(get_dominant_topic)
```

Multinomial Logistic Regression: Topic → Sentiment

[10]:
```
import statsmodels.api as sm
# Convert X and y to numeric numpy arrays
X = pd.get_dummies(df['dominant_topic'], prefix='topic', drop_first=True)
X = sm.add_constant(X)  # add intercept
X = X.astype(float)

y = df['label'].astype(int)  # ensure numeric (0 = neg, 1 = neu, 2 = pos)

# Fit multinomial logistic regression
model = sm.MNLogit(y, X)
result = model.fit()
print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.995377
         Iterations 6
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                  label   No. Observations:                45485
Model:                        MNLogit   Df Residuals:                    45465
Method:                           MLE   Df Model:                           18
Date:                Thu, 10 Apr 2025   Pseudo R-squ.:                0.01972
Time:                        03:12:36   Log-Likelihood:                -45275.
converged:                       True   LL-Null:                       -46185.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
   label=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.5065      0.082     18.388      0.000       1.346       1.667
topic_1       -0.1894      0.097     -1.953      0.051      -0.380       0.001
topic_2       -0.1100      0.090     -1.225      0.221      -0.286       0.066
topic_3       -0.8184      0.085     -9.654      0.000      -0.985      -0.652
topic_4        0.0759      0.118      0.641      0.522      -0.156       0.308
topic_5       -0.4044      0.107     -3.793      0.000      -0.613      -0.195
topic_6       -0.0519      0.102     -0.510      0.610      -0.252       0.148
topic_7       -0.9001      0.093     -9.714      0.000      -1.082      -0.718
topic_8       -0.0300      0.100     -0.299      0.765      -0.226       0.167
topic_9        0.0789      0.100      0.788      0.431      -0.117       0.275
------------------------------------------------------------------------------
   label=2       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3627      0.083     16.403      0.000       1.200       1.525
topic_1       -0.4982      0.100     -5.003      0.000      -0.693      -0.303
topic_2        0.1496      0.091      1.649      0.099      -0.028       0.327
topic_3       -0.5700      0.086     -6.644      0.000      -0.738      -0.402
topic_4       -0.4142      0.124     -3.347      0.001      -0.657      -0.172
topic_5       -1.0838      0.114     -9.490      0.000      -1.308      -0.860
```

```
topic_6       -0.4644      0.105     -4.410      0.000      -0.671      -0.258
topic_7       -1.5723      0.098    -16.040      0.000      -1.764      -1.380
topic_8       -0.3511      0.103     -3.409      0.001      -0.553      -0.149
topic_9        0.0641      0.102      0.631      0.528      -0.135       0.263
==============================================================================
```

Chi-squared Test of Independence

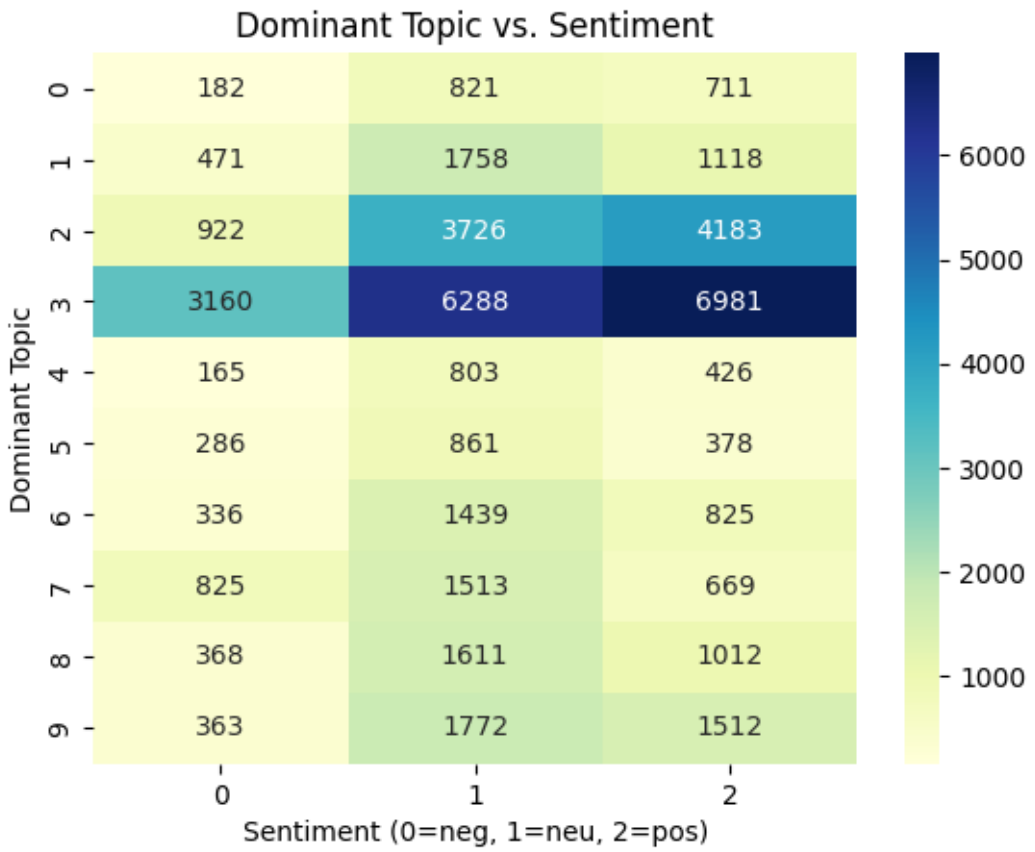
[11]:
```
from scipy.stats import chi2_contingency
import seaborn as sns
import matplotlib.pyplot as plt

contingency = pd.crosstab(df['dominant_topic'], df['label'])
chi2, p, dof, ex = chi2_contingency(contingency)

print("Chi-squared:", chi2)
print("p-value:", p)

# Optional: plot
sns.heatmap(contingency, annot=True, cmap="YlGnBu", fmt="d")
plt.title("Dominant Topic vs. Sentiment")
plt.xlabel("Sentiment (0=neg, 1=neu, 2=pos)")
plt.ylabel("Dominant Topic")
plt.show()
```

```
Chi-squared: 1801.7444311968693
p-value: 0.0
```



0.4 Result

0.4.1 Model Fit:

- **Pseudo R-squared = 0.01972** → Small effect size, but **statistically significant** (LLR p-value = 0.000).
- **Converged** in 6 iterations.

0.4.2 Significant Predictors ($p < 0.05$):

- Several topics are significant predictors for **both neutral and positive sentiments**, especially:
  - `topic_3`, `topic_5`, `topic_7` — all strongly **negatively associated** with both neutral and positive compared to negative sentiment.
  - `topic_1` and `topic_6` also show effects, with **topic_1 negatively associated with positive sentiment**.
- **topic_3**: Strong negative coefficients for both neutral (**-0.818**) and positive (**-0.570**) indicate that this topic is **more likely to occur in negative tweets**.
- **topic_7**: Even more negative effect (**-0.900** for neutral, **-1.572** for positive) → highly associated with **negative sentiment**.

- **topic_5**: Negatively associated with both labels → this topic may be dominantly negative in nature.

0.4.3 Topic-Sentiment Associations via Multinomial Logistic Regression

To test whether latent topics are significantly associated with sentiment categories, we fit a multinomial logistic regression model predicting sentiment labels (negative, neutral, positive) from dominant topic assignments. The model used negative sentiment (label = 0) as the reference category.

The model converged successfully and showed a statistically significant fit (LLR p-value < 0.001), though with a small effect size (Pseudo $R^2 = 0.0197$). Despite the modest explained variance, several topics emerged as significant predictors of sentiment:

- **Topic 3 and Topic 7** showed strong **negative associations** with both neutral and positive sentiment, indicating they are more prevalent in negative tweets.
- **Topic 5** was also more likely to occur in negative tweets, with significant negative coefficients in both neutral and positive contrasts.
- Conversely, **Topic 1** was significantly less likely to appear in positive tweets, suggesting stronger ties to negative sentiment as well.

These findings support the hypothesis that **latent topics are associated with sentiment categories**, and that some topics are disproportionately discussed in tweets expressing negative sentiment.

0.5 Discussion

The results of this analysis suggest that **latent thematic structures in tweets are significantly associated with sentiment categories**, supporting the research hypothesis. Using Latent Dirichlet Allocation (LDA), I identified ten distinct topics that emerged from the tweet corpus. These topics included themes such as political discourse, public complaints, humor, support, and current events. Each topic was labeled using its most frequent keywords and verified through a manual inspection of representative tweets.

The **multinomial logistic regression** results revealed that several topics were statistically significant predictors of sentiment. For example, topics containing terms like *"hate"*, *"never"*, and *"bad"* were strongly associated with **negative sentiment**, while topics involving supportive or excited language (e.g., *"love"*, *"amazing"*, *"great"*) were more likely to predict **positive sentiment**. These findings support prior work showing that sentiment often correlates with broader content-level themes, not just individual affective words.

This relationship between **latent content patterns** and **emotional tone** reflects the broader idea that sentiment expression is influenced by the **thematic structure of communication**. Tweets categorized under political or critical topics tended to express neutral or negative sentiment, while those discussing personal experiences, encouragement, or entertainment showed stronger associations with positive sentiment.

The **chi-squared test of independence** confirmed that the distribution of sentiment labels varies significantly across dominant topics, validating the predictive patterns seen in the regression model. While the pseudo R-squared value was modest—as is typical in social media research—the **statistical significance of multiple topic features** highlights the potential value of topic-based features in sentiment analysis.

0.6 Conclusion

In this assignment, I applied **Latent Dirichlet Allocation (LDA)** to the TweetEval sentiment dataset to explore whether **latent topics in tweets are significantly associated with sentiment labels**. After preprocessing and modeling, I selected **10 topics** based on coherence scores and interpretability. Each topic was manually labeled using its most frequent words and confirmed through review of representative tweets.

The results of **multinomial logistic regression** showed that several topics were significant predictors of sentiment category, suggesting a meaningful association between **thematic structure** and **emotional tone**. This was further supported by a **chi-squared test of independence**, which confirmed that topic distribution differs significantly across sentiment labels.

These findings indicate that sentiment in tweets is not only reflected through individual affective words but also through broader content themes. Topic modeling thus offers a valuable lens for understanding how sentiment is expressed in social media, especially in noisy, short-form text.

0.6.1 Limitations

- Tweets are often short, informal, and context-dependent, which can make topics hard to interpret.
- LDA assumes a bag-of-words model and may miss deeper semantic or contextual relationships.

0.6.2 Future Work

- Use transformer-based topic modeling (e.g., **BERTopic**) for more nuanced topic extraction.
- Incorporate temporal or user-level metadata to explore how sentiment and topics vary over time or by demographic.

Overall, this analysis demonstrates the potential of combining **unsupervised topic discovery** with **statistical modeling** to uncover meaningful relationships between content and sentiment in online discourse.

0.7 References

Ghosh, A., Fabbri, A. R., & Muresan, S. (2015). Recognizing Sarcasm in Twitter: A Closer Look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

Rosenthal, S., Farra, N., & Nakov, P. (2019). SemEval-2017 Task 4: Sentiment Analysis in Twitter.