Sentiment analysis using these lexicon-based methods, advanced sentiment analysis and the basic Textblob technique can yield different results for several reasons. While all methods aim to determine sentiment in text, they have distinct approaches and characteristics. Lexicon and Textblob are both tools used for text processing and natural language processing (NLP), but they serve different purposes and have different features:

1. Textblob [1]

The textblob library in python used for processing text data. Provides a very simple and consistent Application Programming Interface (API). Which is used for diving into common natural language processing tasks, such as part-of-speech tagging, noun phrase extraction, sentiment analysis, translation.

In Sentimental Analysis Textblob includes a built-in sentiment analysis module that uses a combination of a lexicon-based approach (Pattern) and a machine learning-based approach (NLTK's Naive Bayes classifier) to perform sentiment analysis on text. With its ease of use the textblob is ease to user and has as user friendly API. Its abstracts many complex Natural Process Language tasks into simpler method, which makes it possible of novices to NPL to use it without experience. And is customization using textblob allows user to train custom models for specific NLP tasks if needed, providing a degree of flexibility beyond prebuild lexicons.

1. The Lexicon-Based Sentiment Analysis

Lexicon refers to a python dictionary or database of words or phrases with associated linguistic or semantic information, such as sentiment scores, part-of-speech tags, or definitions. Lexicon-based methods rely on predefined sentiment lexicons or dictionaries containing words or phrases with associated sentiment scores. The overall sentiment of a text is calculated based on the scores of individual words or phrases in the text. Lexicon-based methods are relatively simple and straightforward but may not capture the nuances or context of sentiment well. The accuracy of lexicon-based methods depends on the comprehensiveness and quality of the sentiment lexicon being used. These methods may struggle with slang, idioms, or language variations not covered in the lexicon.

3. Advanced Sentiment Analysis Models

Some advanced models, such as machine learning models (e.g., Naive Bayes, Support Vector Machines, deep learning models) or pre-trained language models (e.g., BERT, GPT), can capture more complex patterns and context in text data. These models can generalize better to a wide range of sentiments and text types, including those with sarcasm, irony, or subtle sentiment. They can be fine-tuned on specific tasks or domains, making them adaptable to various applications and improving their accuracy.

Because of these differences, lexicon-based sentiment analysis and advanced sentiment analysis models may not always return the same results. Lexicon-based methods tend to be simpler and may produce different results, especially when dealing with non-standard language or highly context-dependent sentiment expressions.

The choice between lexicon-based and advanced methods depends on your specific needs and the characteristics of your text data. Lexicon-based methods are quick and straightforward for simple sentiment tasks, while advanced models are more versatile and may provide better accuracy for complex tasks. It's essential to consider the trade-offs and select the approach that best suits your use case. Additionally, you can combine both methods or use lexicon-based methods as a baseline for comparison when developing and evaluating advanced sentiment analysis models.

Lexicon and TextBlob are both tools used for text processing and natural language processing (NLP), but they serve different purposes and have different features.

Lexicon-based sentiment analysis uses sentiment lexicons to assign sentiment scores to words or phrases in text. It calculates the overall sentiment of a piece of text based on the scores of the words it contains. Common lexicons used for sentiment analysis include AFINN and VADER.

Lexicons are commonly used for simple sentiment analysis tasks, where the goal is to determine whether text expresses positive, negative, or neutral sentiment. They are also used in other NLP tasks, such as part-of-speech tagging or named entity recognition.

The main difference between Lexicon and TextBlob is that Lexicon typically refers to a dictionary or database of linguistic information, while TextBlob is a Python library that provides a range of NLP tools and includes a sentiment analysis module as one of its features. TextBlob 's sentiment analysis module combines both lexicon-based and machine learning-based approaches to provide a more comprehensive analysis of text sentiment.

|  |
| --- |
| A graph with blue rectangular bars  Description automatically generated |
| Figure 1 Textblob NPL sentiment analysis |
| A graph with blue bars  Description automatically generated |
| Figure 2 Lexicon based sentimental analysis |

The dataset spans from 2014 to 2019 and comprises over 500,000 entries. Upon filtering the data for the keyword "greyhound," it isolates 38 rows in the "contents" column where the term "greyhound" is present. The resulting analysis demonstrates the disparity between lexicon-based and sentiment analysis approaches.

Shown below is the Comparing both NPL and Lexicon-based sentiment analysis approaches on the data.

The results show that:

* There are a total of 38 rows or IDs in the dataset.

The count how often the two columns agree on the sentiment label:

The number of matches between the two columns

- The agreement rate can be calculated as the ratio of agreements to the total number of IDs:

- Agreement rate: 14 / 38 ≈ 0.3684 (approximately 36.84%)

- The disagreement rate is the complementary value of the agreement rate:

- Disagreement rate: 24 / 38 ≈ 0.6316 (approximately 63.16%)

From this comparison, we can see that the "Sentiment" and "Sentiment\_Label\_lex" columns do not always agree in assigning sentiment labels. There is a relatively high level of disagreement, with approximately 63.16% of the IDs having different sentiment labels between the two columns.

Further analysis could involve examining specific examples of disagreement to understand why these differences occur and which sentiment analysis method might be more accurate or appropriate for the given data and context. Additionally, it might be helpful to assess the performance of each method in terms of accuracy, precision, recall, or other evaluation metrics to determine which one aligns better with the expected sentiment labels.

1. Disagreement Between Sentiment Labels:

There is a significant level of disagreement between the sentiment labels assigned by the two methods. Approximately 63.16% of the IDs have different sentiment labels in the "Sentiment" and "Sentiment\_Label\_lex" columns.

This discrepancy suggests that the two methods have different interpretations or criteria for assigning sentiment labels, which could be due to variations in the underlying algorithms or lexicons used.

2. Agreement in Sentiment Labels:

While there is substantial disagreement, there are still cases where both methods agree on the sentiment label. Approximately 36.84% of the IDs have the same sentiment label in both columns.

In these instances of agreement, it could be an indication that both methods are capturing a clear and unambiguous sentiment signal in the text.

3. Further Analysis Required:

To better understand the reasons for disagreement, it would be beneficial to examine specific examples where the two methods differ in their sentiment labels.

Analyzing these examples could help identify cases where one method might be more accurate or suitable than the other, and it could provide insights into the strengths and limitations of each method.

4. Sentiment Analysis Evaluation:

It's essential to consider the overall accuracy and performance of both sentiment analysis methods. Evaluation metrics such as accuracy, precision, recall, F1-score, or ROC-AUC can be calculated to assess their effectiveness in assigning sentiment labels.

Additionally, the choice between these methods may depend on the specific objectives of the sentiment analysis and the context of the data being analysed.

5. Context Matters:

Sentiment analysis is highly context-dependent, and the choice of method should align with the goals of the analysis. Different methods may excel in different contexts or for specific types of text data.

The results also show that on the lexicon based sentimental analysis the data in the twitter dataset contains over ?? results to part:

Limitations:

Twitter data.

Twitter data scraping has become more challenging over the years due to several factors, including Twitter's efforts to protect user data, privacy concerns, and technical countermeasures to prevent automated scraping. Here are some reasons why Twitter data scraping has become difficult:

1. \*\*API Limitations\*\*: Twitter has imposed rate limits and restrictions on its API (Application Programming Interface), which is the preferred and legitimate way to access Twitter data. Free access to the Twitter API has become more limited, making it difficult to collect large volumes of data without a paid subscription.

2. \*\*Data Access Authorization\*\*: To access certain types of Twitter data, especially user-specific data such as tweets, you often need proper authorization and authentication. This requires developers to adhere to Twitter's guidelines and use OAuth authentication, which adds complexity to the scraping process.

3. \*\*Rate Limiting\*\*: Twitter enforces strict rate limits on API requests, including the number of requests per minute and the number of tweets you can retrieve in a single request. This makes it challenging to collect large datasets quickly.

4. \*\*Data Anonymization\*\*: Twitter has taken steps to anonymize and generalize certain data, making it more challenging to track individual user activities and behaviors.

5. \*\*Content Encryption\*\*: Some parts of Twitter's website use encryption (HTTPS), which makes it harder to scrape data from those pages directly using traditional web scraping techniques.

6. \*\*Anti-Scraping Measures\*\*: Twitter employs various anti-scraping measures to detect and block automated scraping bots. These measures can include rate limiting, CAPTCHAs, and IP blocking.

7. \*\*Legal and Ethical Concerns\*\*: There are legal and ethical considerations when scraping data from websites like Twitter. Unauthorized scraping can potentially violate Twitter's terms of service and data usage policies, leading to legal consequences.

8. \*\*User Privacy\*\*: Privacy concerns have prompted platforms like Twitter to take measures to protect user data. As a result, some user-specific data may be restricted or limited in accessibility.

9. \*\*Evolving Technology\*\*: As technology evolves, so do the methods used to protect websites and data. Twitter and other social media platforms continually update their systems to counter scraping attempts.

Given these challenges, it's essential to approach Twitter data scraping with caution and adhere to Twitter's terms of service and API usage policies. If you need access to Twitter data for research or analysis, consider using the official Twitter API, which provides a legitimate and sanctioned way to retrieve data while adhering to Twitter's guidelines and limitations. Additionally, always respect user privacy and ensure that your data scraping efforts comply with relevant laws and regulations. Art

Conclusions:

In conclusion, the data demonstrates that sentiment analysis results can vary significantly depending on the method used. It highlights the importance of carefully selecting the appropriate sentiment analysis approach based on the specific requirements of the analysis and the characteristics of the data. Further investigation into specific cases of disagreement can provide valuable insights into the strengths and weaknesses of each method.

References:

<https://textblob.readthedocs.io/en/dev/>