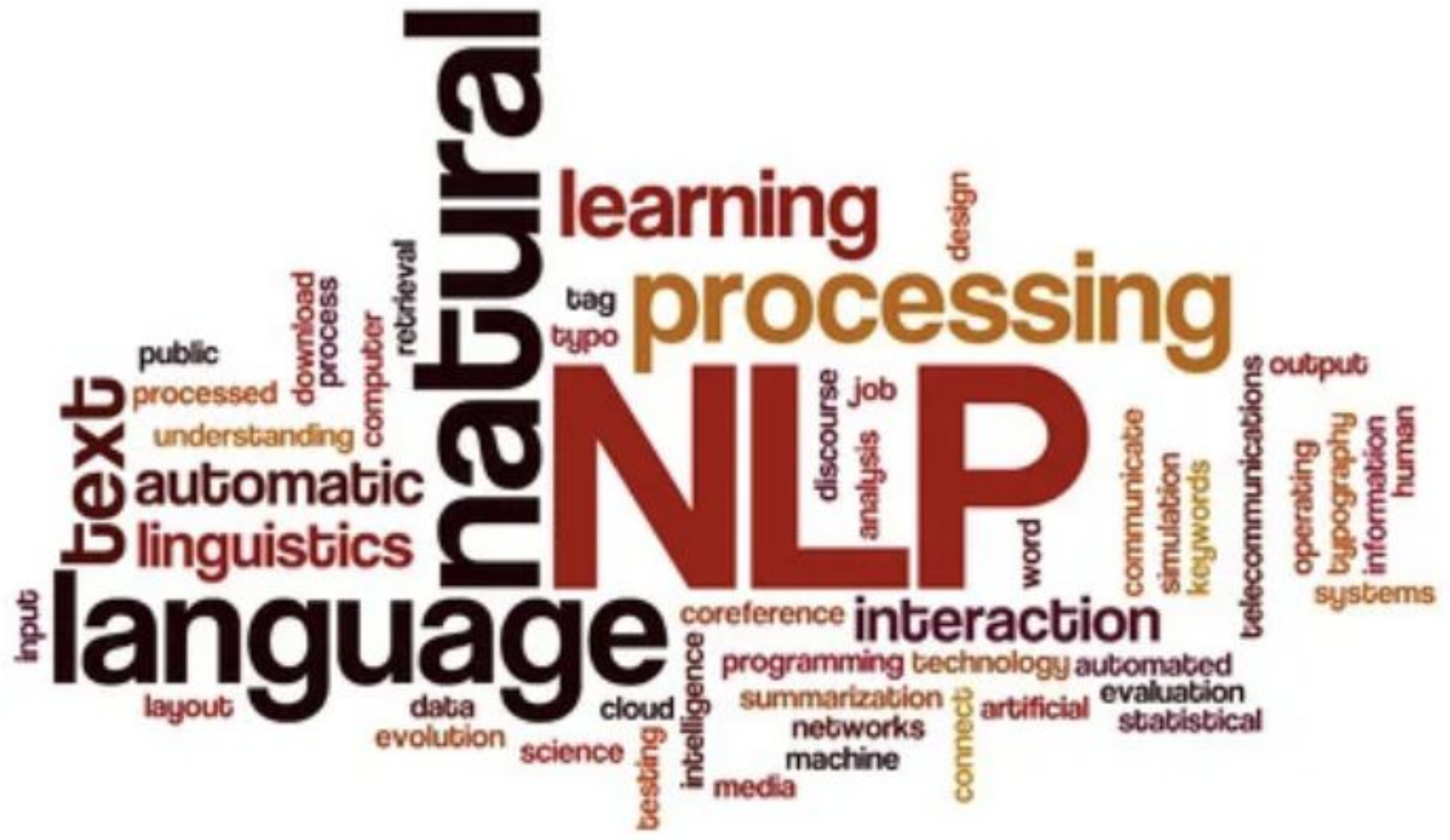
# Exploring Twitter Data with LDA: A Powerful Topic Modeling Technique

Anjali Shahi, Chaitanya Attarde, Mikhail Pinto, Mohit Kambli

## Introduction

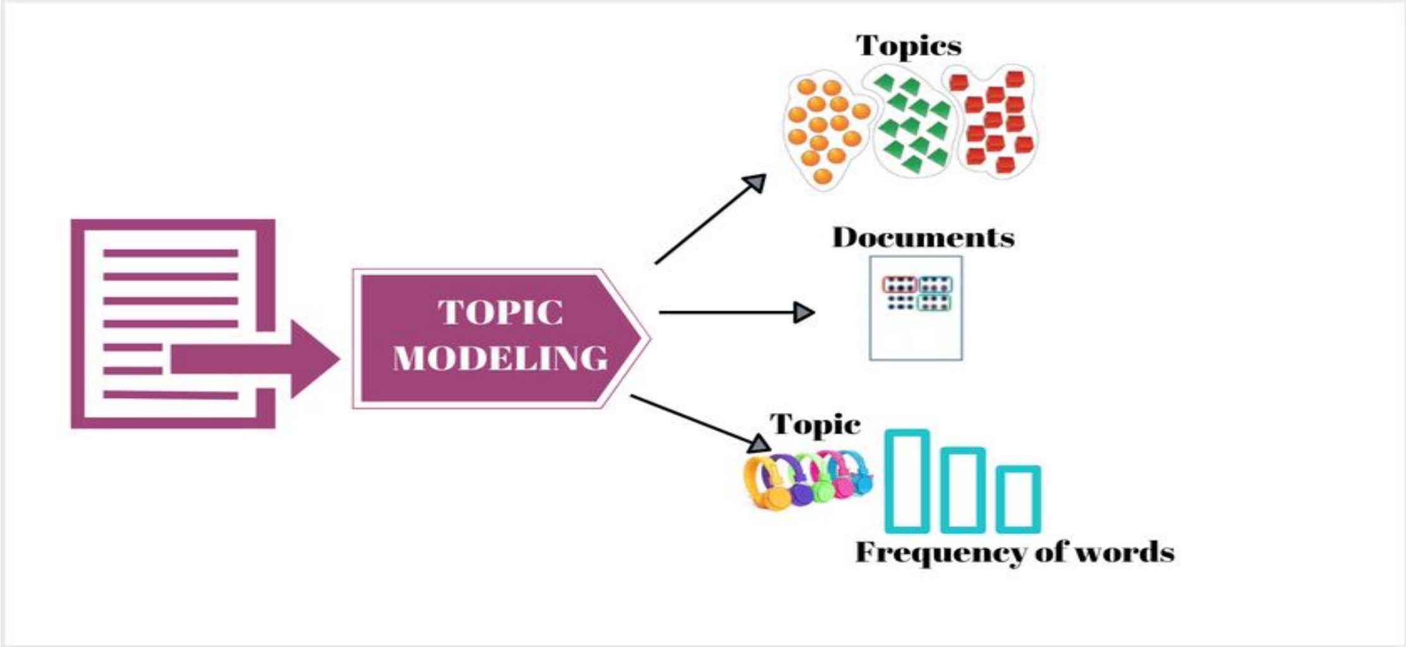
**Natural Language Processing:**

- Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language.
- It is used in a wide range of applications, including language translation, sentiment analysis, text summarization, and speech recognition.
- NLP techniques include tokenization, stemming, part-of-speech tagging, named entity recognition, topic modeling and sentiment analysis.



**Topic Modeling:**

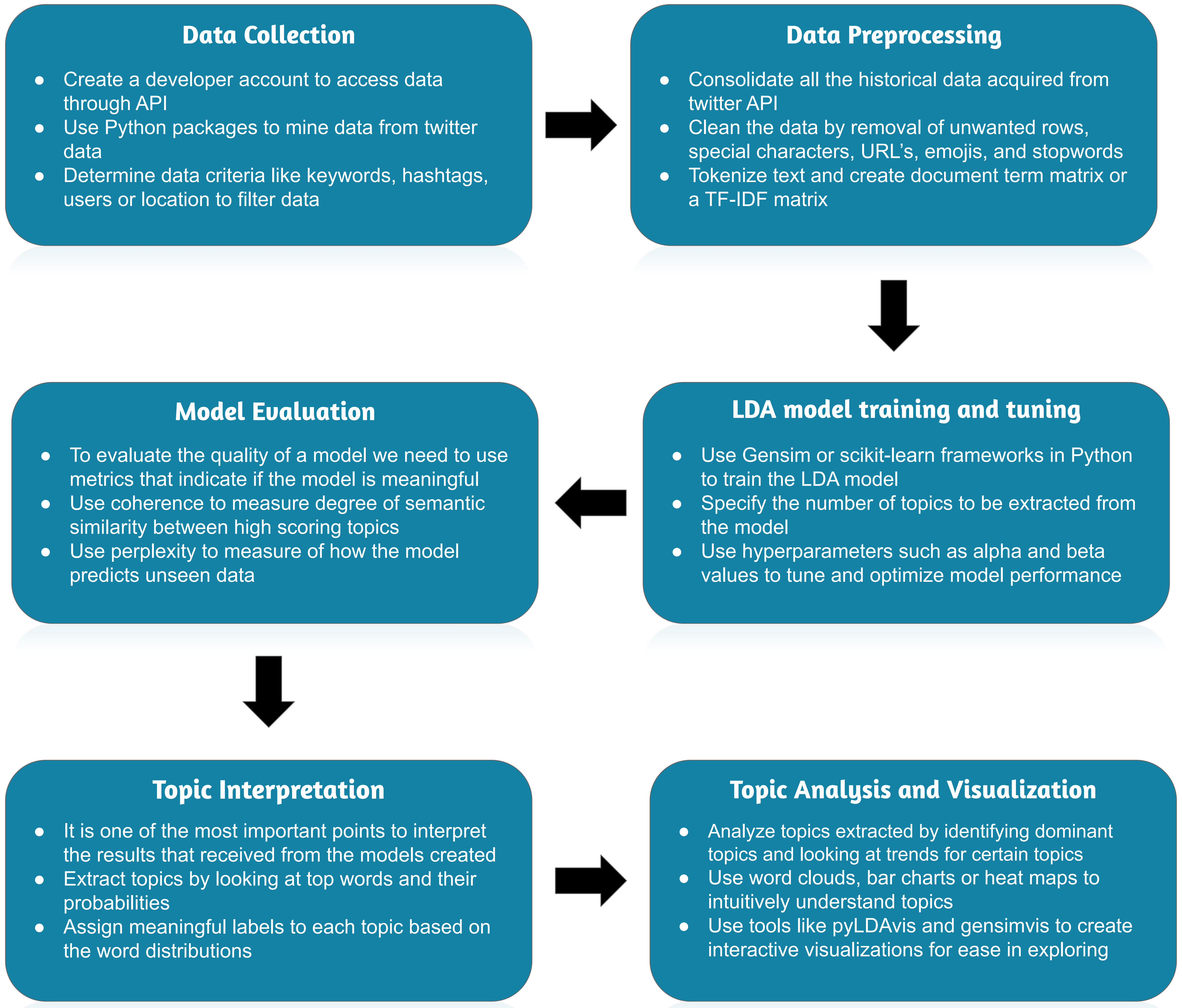
- Topic modeling is a process of discovering hidden patterns or themes within a collection of documents.
- It is used in a wide range of applications, including text classification, information retrieval, and content analysis.
- Techniques for topic modeling include LDA, Non-Negative Matrix Factorization (NMF), and Hierarchical Dirichlet Process (HDP).



**LDA (Latent Dirichlet Allocation):**

- LDA is a statistical model used for topic modeling, which is the process of identifying hidden patterns in a collection of documents.
- The LDA algorithm uses statistical inference to estimate the parameters of the model, including the distribution of topics and the distribution of words within each topic.

## Methodology



### Data Collection

- Create a developer account to access data through API
- Use Python packages to mine data from twitter data
- Determine data criteria like keywords, hashtags, users or location to filter data

### Data Preprocessing

- Consolidate all the historical data acquired from twitter API
- Clean the data by removal of unwanted rows, special characters, URL's, emojis, and stopwords
- Tokenize text and create document term matrix or a TF-IDF matrix

### LDA model training and tuning

- Use Gensim or scikit-learn frameworks in Python to train the LDA model
- Specify the number of topics to be extracted from the model
- Use hyperparameters such as alpha and beta values to tune and optimize model performance

### Model Evaluation

- To evaluate the quality of a model we need to use metrics that indicate if the model is meaningful
- Use coherence to measure degree of semantic similarity between high scoring topics
- Use perplexity to measure of how the model predicts unseen data

### Topic Interpretation

- It is one of the most important points to interpret the results that received from the models created
- Extract topics by looking at top words and their probabilities
- Assign meaningful labels to each topic based on the word distributions

### Topic Analysis and Visualization

- Analyze topics extracted by identifying dominant topics and looking at trends for certain topics
- Use word clouds, bar charts or heat maps to intuitively understand topics
- Use tools like pyLDAvis and gensimvis to create interactive visualizations for ease in exploring

## Benefits

- Flexible and can be applied on any type of data.
- Identifies hidden and latent topics.
- Provides insights into underlying patterns and structures.
- Scalable and suitable for processing big data.

## Limitations

- Requires fixed number for topics to be specified in advance.
- Assumes topics are independent of each other.
- Does not consider temporal aspect of data.
- Does not provide a clear measure of the quality of identified topics.

## Applications

- In natural language processing and speech recognition, to identify the underlying themes in spoken or written language.
- In image analysis and object recognition, to extract meaningful topics from image captions.
- Can be used to identify fake news and also Spam detection.
- Can be used in Recommender systems, to suggest similar products or articles on the topic of interest.
- Can be used in sentiment analysis, to identify sentiment of specific topic in text data.
- Can be used in topic modelling of text data from various domains like finance, healthcare, social media and politics.

## Conclusion

In conclusion, natural language processing (NLP) is a powerful tool for analyzing and interpreting human language. In today's fast-paced business world, corporate layoffs are a common occurrence that often generate significant public attention and discussion, especially on social media platforms, like Twitter. To get better insight of the sentiments expressed in tweets related to layoffs, we conducted a sentiment analysis using NLP techniques, specifically topic modeling. By leveraging the power of topic modeling, we were able to answer research questions such as the predominant sentiments expressed in tweets related to layoffs and how they vary depending on the specific company and industry. Our project highlights the value of NLP techniques in analyzing unstructured data sources such as social media, and the potential insights that can be gained through sentiment analysis in fields such as business and finance. With the continuous evolution and improvement of NLP technologies, we can look forward to even more advancements in analyzing and interpreting human language.

## References

- Liu, Y., Zhao, W., Jiang, J., & Zhang, Y. (2021). Mining opinion topics and dominant words from online reviews via deep neural networks and latent Dirichlet allocation. Information Processing & Management, 58(5), 102565. doi: 10.1016/j.ipm.2021.102565
- Wang, J., Zhao, Y., & Lu, X. (2021). A comparative study on deep neural network and LDA in sentiment analysis. Journal of Ambient Intelligence and Humanized Computing, 12(6), 5279-5291. doi: 10.1007/s12652-021-03489-3
- Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics, 80-88. Retrieved from https://www.researchgate.net/publication/261064842_Empirical_study_of_topic_modeling_in_Twitter
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. Proceedings of the 33rd European Conference on Information Retrieval, 338-349. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-20161-5_29
- Hong, L., & Davison, B. D. (2011). Predicting popular messages in Twitter. Proceedings of the 20th International Conference Companion on World Wide Web, 57-58. Retrieved from https://dl.acm.org/doi/10.1145/1963192.1963253

Analysis

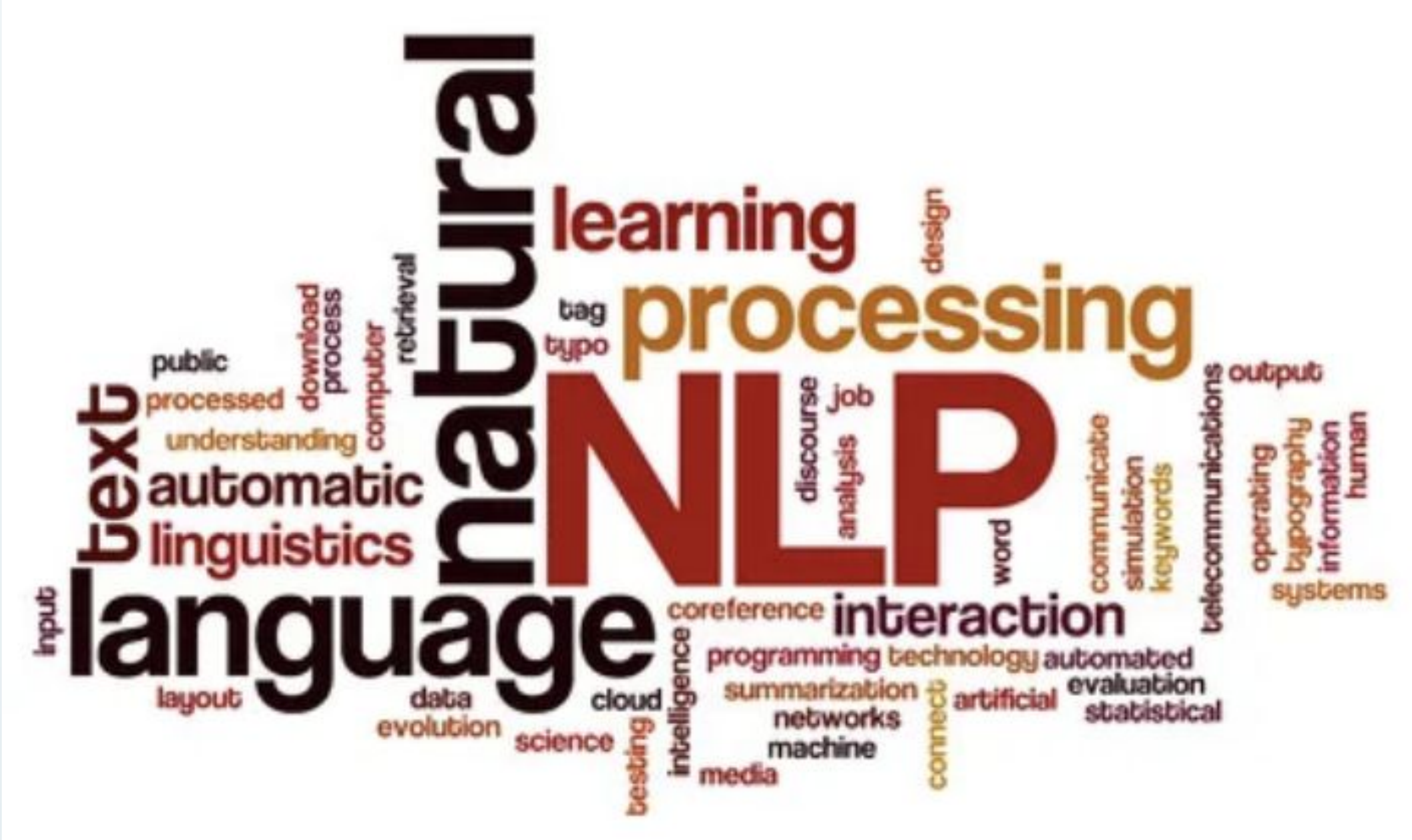
# Exploring Twitter Data with LDA: A Powerful Topic Modeling Technique

Anjali Shahi, Chaitanya Attarde, Mikhail Pinto, Mohit Kambli

## Introduction

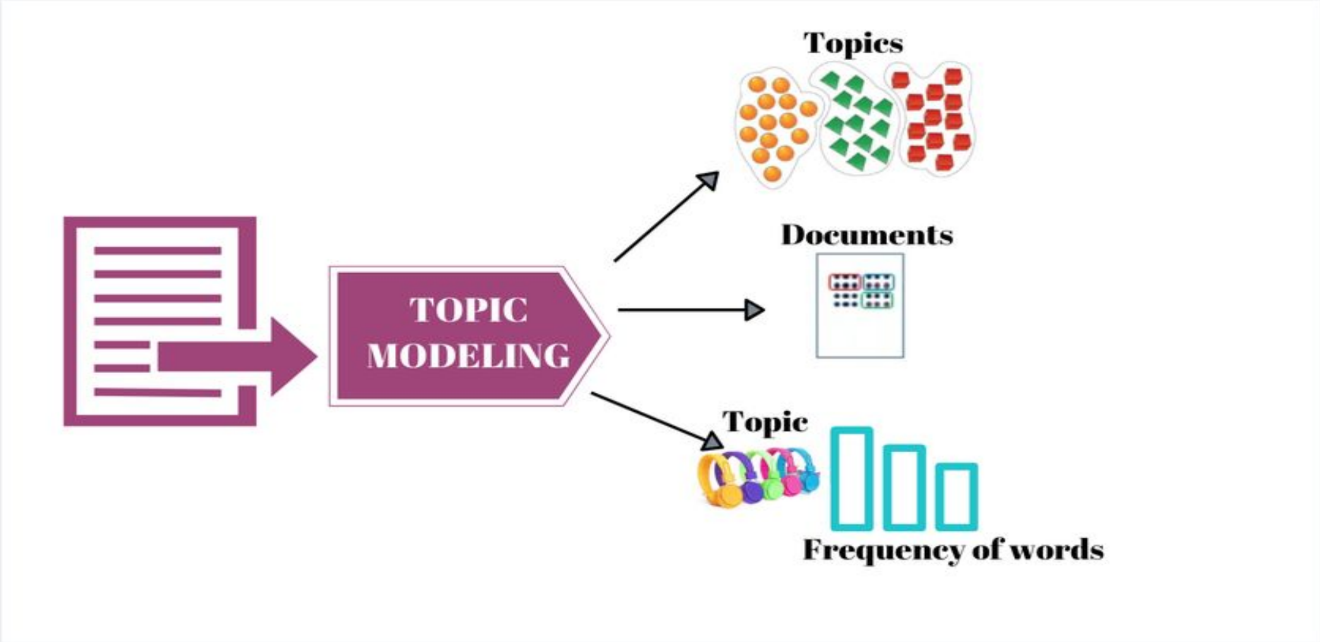
**Natural Language Processing:**

- Natural Language Processing (NLP) is a subfield of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language.
- It is used in a wide range of applications, including language translation, sentiment analysis, text summarization, and speech recognition.
- NLP techniques include tokenization, stemming, part-of-speech tagging, named entity recognition, topic modeling and sentiment analysis.



**Topic Modeling:**

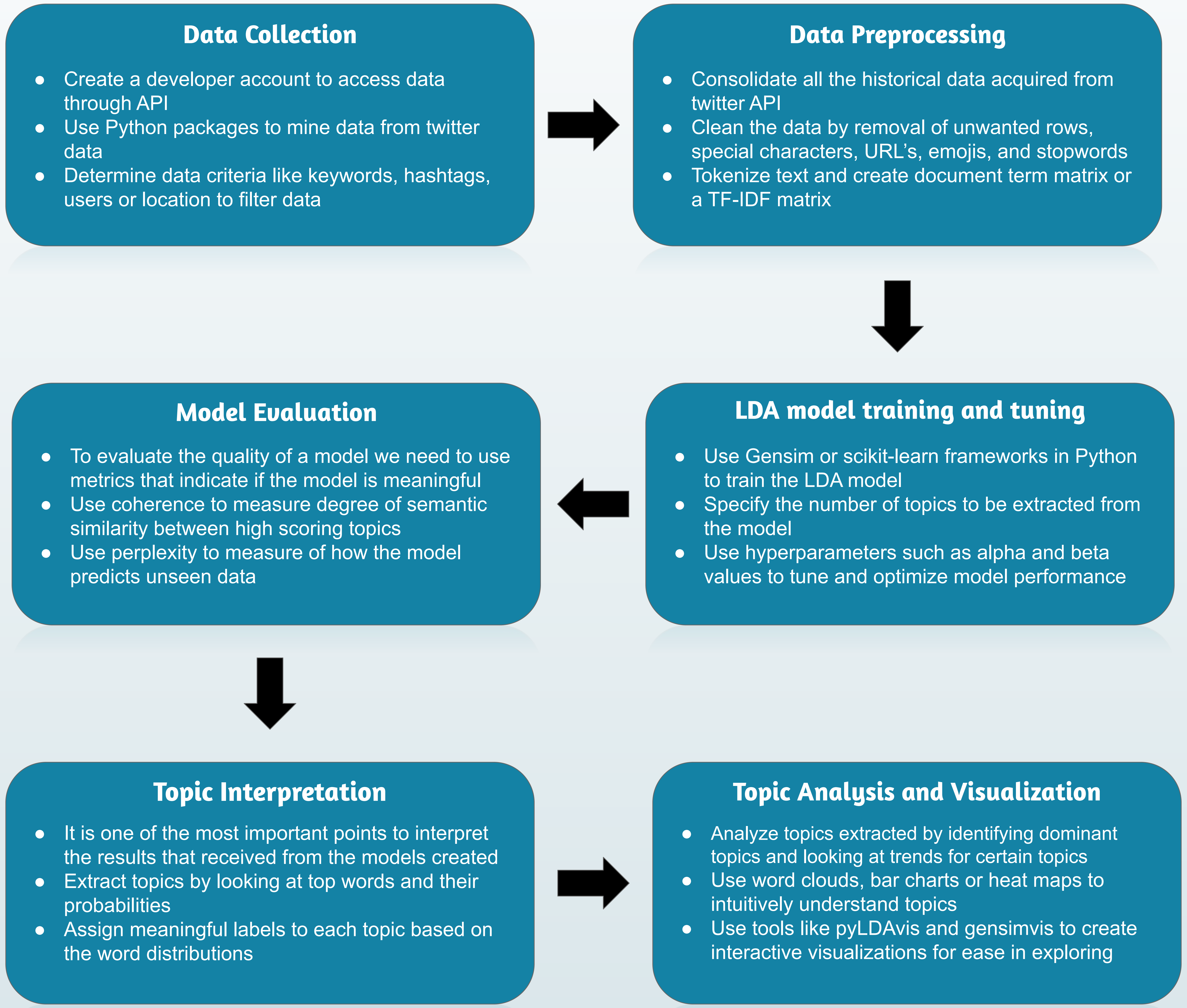
- Topic modeling is a process of discovering hidden patterns or themes within a collection of documents.
- It is used in a wide range of applications, including text classification, information retrieval, and content analysis.
- Techniques for topic modeling include LDA, Non-Negative Matrix Factorization (NMF), and Hierarchical Dirichlet Process (HDP).



**LDA (Latent Dirichlet Allocation):**

- LDA is a statistical model used for topic modeling, which is the process of identifying hidden patterns in a collection of documents.
- The LDA algorithm uses statistical inference to estimate the parameters of the model, including the distribution of topics and the distribution of words within each topic.

## Methodology



### Data Collection

- Create a developer account to access data through API
- Use Python packages to mine data from twitter data
- Determine data criteria like keywords, hashtags, users or location to filter data

### Data Preprocessing

- Consolidate all the historical data acquired from twitter API
- Clean the data by removal of unwanted rows, special characters, URL's, emojis, and stopwords
- Tokenize text and create document term matrix or a TF-IDF matrix

### LDA model training and tuning

- Use Gensim or scikit-learn frameworks in Python to train the LDA model
- Specify the number of topics to be extracted from the model
- Use hyperparameters such as alpha and beta values to tune and optimize model performance

### Model Evaluation

- To evaluate the quality of a model we need to use metrics that indicate if the model is meaningful
- Use coherence to measure degree of semantic similarity between high scoring topics
- Use perplexity to measure of how the model predicts unseen data

### Topic Interpretation

- It is one of the most important points to interpret the results that received from the models created
- Extract topics by looking at top words and their probabilities
- Assign meaningful labels to each topic based on the word distributions

### Topic Analysis and Visualization

- Analyze topics extracted by identifying dominant topics and looking at trends for certain topics
- Use word clouds, bar charts or heat maps to intuitively understand topics
- Use tools like pyLDAvis and gensimvis to create interactive visualizations for ease in exploring

## Benefits

- Flexible and can be applied on any type of data.
- Identifies hidden and latent topics.
- Provides insights into underlying patterns and structures.
- Scalable and suitable for processing big data.

## Limitations

- Requires fixed number for topics to be specified in advance.
- Assumes topics are independent of each other.
- Does not consider temporal aspect of data.
- Does not provide a clear measure of the quality of identified topics.

## Applications

- In natural language processing and speech recognition, to identify the underlying themes in spoken or written language.
- In image analysis and object recognition, to extract meaningful topics from image captions.
- Can be used to identify fake news and also Spam detection.
- Can be used in Recommender systems, to suggest similar products or articles on the topic of interest.
- Can be used in sentiment analysis, to identify sentiment of specific topic in text data.
- Can be used in topic modelling of text data from various domains like finance, healthcare, social media and politics.

## Conclusion

Corporate layoffs are a common occurrence in today's business world, and they often generate significant public attention and discussion. With the rise of social media platforms, individuals now have a powerful tool to express their opinions and emotions about these events. In this project, we would be conducting a sentiment analysis of tweets related to layoffs, which would in turn help us in  answer the following research questions:

- What are the predominant sentiments expressed in tweets related to layoffs?
- How do sentiments vary depending on the specific company and industry?

In order to meet the requirements, we would be bringing one of the NLP's technique in action, known as Topic Modeling.

## References

- Liu, Y., Zhao, W., Jiang, J., & Zhang, Y. (2021). Mining opinion topics and dominant words from online reviews via deep neural networks and latent Dirichlet allocation. Information Processing & Management, 58(5), 102565. doi: 10.1016/j.ipm.2021.102565
- Wang, J., Zhao, Y., & Lu, X. (2021). A comparative study on deep neural network and LDA in sentiment analysis. Journal of Ambient Intelligence and Humanized Computing, 12(6), 5279-5291. doi: 10.1007/s12652-021-03489-3
- Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. Proceedings of the First Workshop on Social Media Analytics, 80-88. Retrieved from https://www.researchgate.net/publication/261064842_Empirical_study_of_topic_modeling_in_Twitter
- Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. Proceedings of the 33rd European Conference on Information Retrieval, 338-349. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-20161-5_29
- Hong, L., & Davison, B. D. (2011). Predicting popular messages in Twitter. Proceedings of the 20th International Conference Companion on World Wide Web, 57-58. Retrieved from https://dl.acm.org/doi/10.1145/1963192.1963253