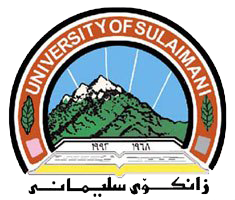
University of Sulaimani

College Of Science

Computer Department



****



**Social Media Text Preprocessing: Enhancing Data for Sentiment Analysis and Topic Modeling on Twitter**

Supervisor Dr. Miran Taha

Prepared By:

Hawraz Abdalla Abubakr

Year 2023-2024

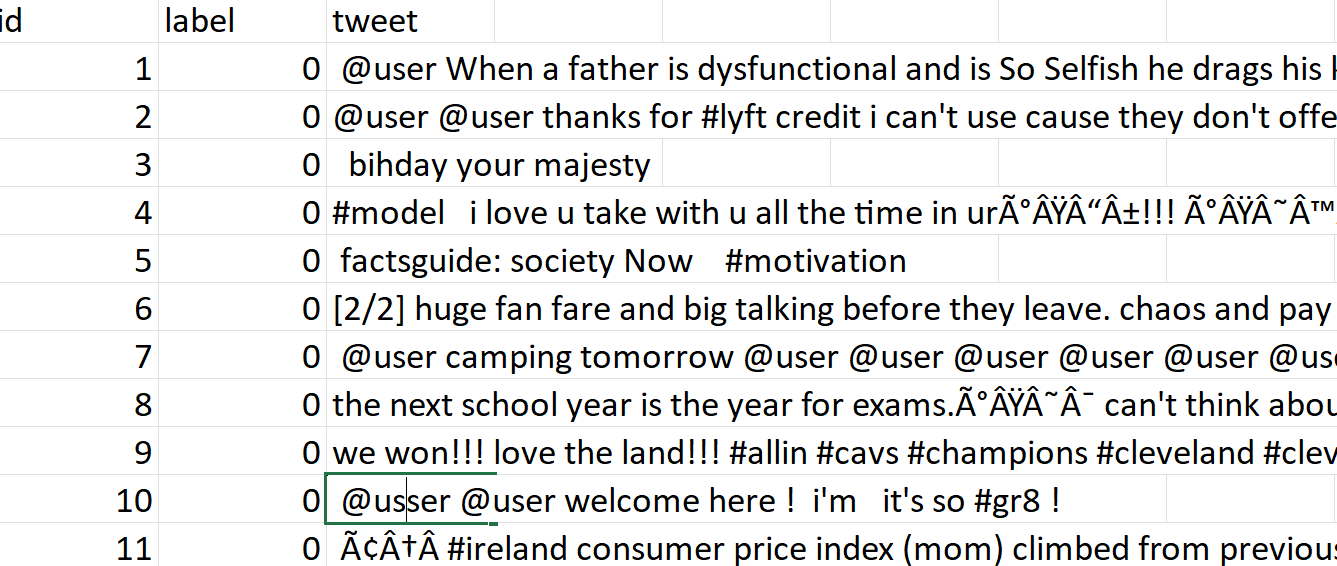
**what is social media text preprocessing**?

Social media text preprocessing is a crucial step in natural language processing (NLP) tasks, as social media data often contains noise, informal language, and various challenges that need to be addressed before analysis.

Social media is a place where people share quick updates. It's like a big library with lots of up-to-date information. We can use it to understand how people feel (sentiment analysis) and figure out what they're talking about (topic modeling). But, because the words people use on social media can be a bit messy, we need to tidy them up to get useful information. This project explains how we make Twitter data better using different tricks and tools to analyze feelings and discover hot topics.

**The Dataset of the Project**

I recently obtained a Twitter dataset from GitHub, but it appears to be noisy. In my preprocessing efforts, I've been working to clean and prepare the data. Although the dataset isn't real-time, its contents still hold valuable insights into various topics. By addressing the noise and optimizing the dataset, I aim to extract meaningful patterns and trends from the Twitter data for my analysis.



This figure(1) is Example of Dataset before we perform cleaning process.

**Requirement of project:**

**Install Python:**

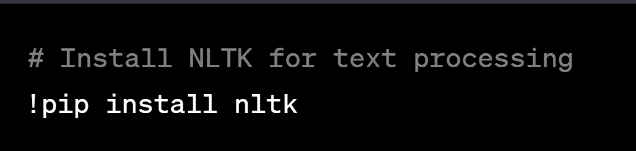
Visit the official Python website to download and install the latest version of Python.

**Install Jupyter:**

Open a terminal or command prompt and run the following command to install Jupyter using pip.

**Install Libraries:**

Once Python and Jupyter are installed, then we have installed the libraries that require for the project:



# Install Spacy for advanced NLP tasks

!pip install spacy

!python -m spacy download en

# Install TextBlob for simplified text processing

!pip install textblob

# Install regex for regular expression-based text cleaning

!pip install regex

# Install Pandas for data manipulation and analysis

!pip install pandas

And many other libraries my not include all of them here ..

**Techniques for preprocessing social media text:**

**Lowercasing:** Convert all text to lowercase to ensure uniformity and simplify further analysis.

**Tokenization:** Break the text into individual words or tokens. This helps in analyzing the content at a finer level.

**Removing Punctuation and Special Characters:** Remove unnecessary punctuation and special characters as they may not contribute significantly to the analysis.

**Removing Stopwords:** Eliminate common words (such as "and," "the," "is") that don't carry much meaning and can be considered noise.

**Stemming and Lemmatization:** Reduce words to their root form. Stemming involves removing suffixes, while lemmatization considers the meaning of the words and transforms them to their base or dictionary form.

**Handling Emoticons and Emoji:** Depending on your analysis, you may choose to replace emoticons and emojis with a corresponding representation or remove them altogether.

**Handling URLs and User Mentions:** Replace or remove URLs and user mentions (e.g., @username) as they might not add much value to the analysis.

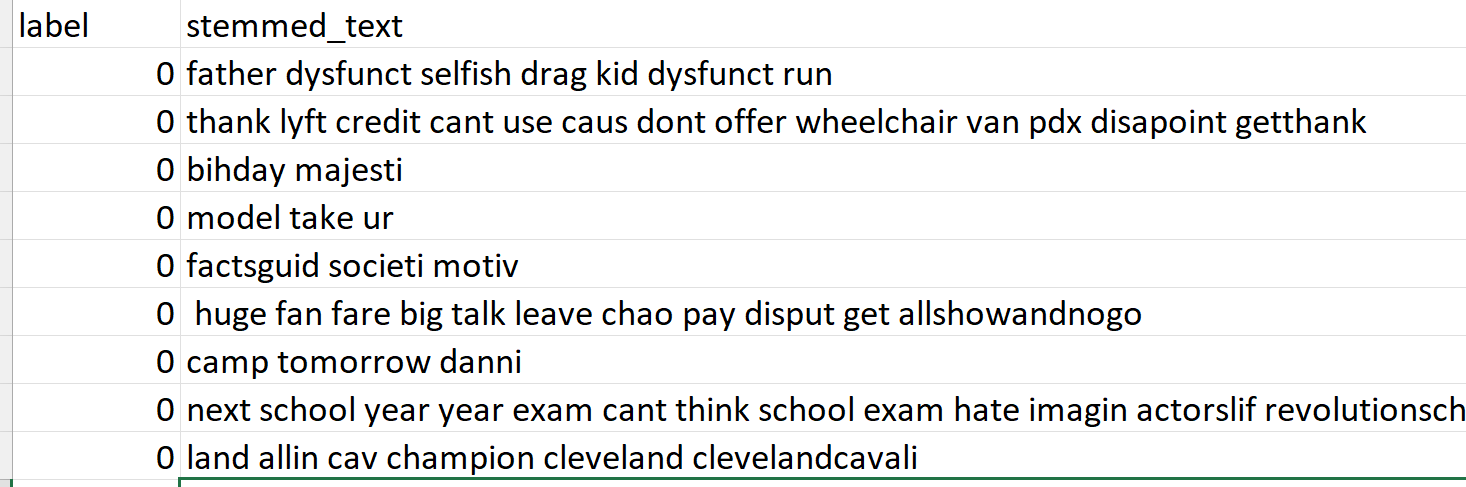
**Handling Hashtags:** Depending on your analysis, you might want to extract and keep hashtags to understand trending topics or remove them if they don't contribute to your specific task.

**Spell Checking and Correction:** Correct spelling mistakes to ensure accurate analysis, especially in the case of informal language commonly found on social media.

**Handling Abbreviations and Acronyms:** Expand abbreviations and acronyms to enhance the interpretability of the text.

**Removing Duplicate Characters:** Address instances where characters are repeated unnecessarily, such as "loooove" or "sooooo."

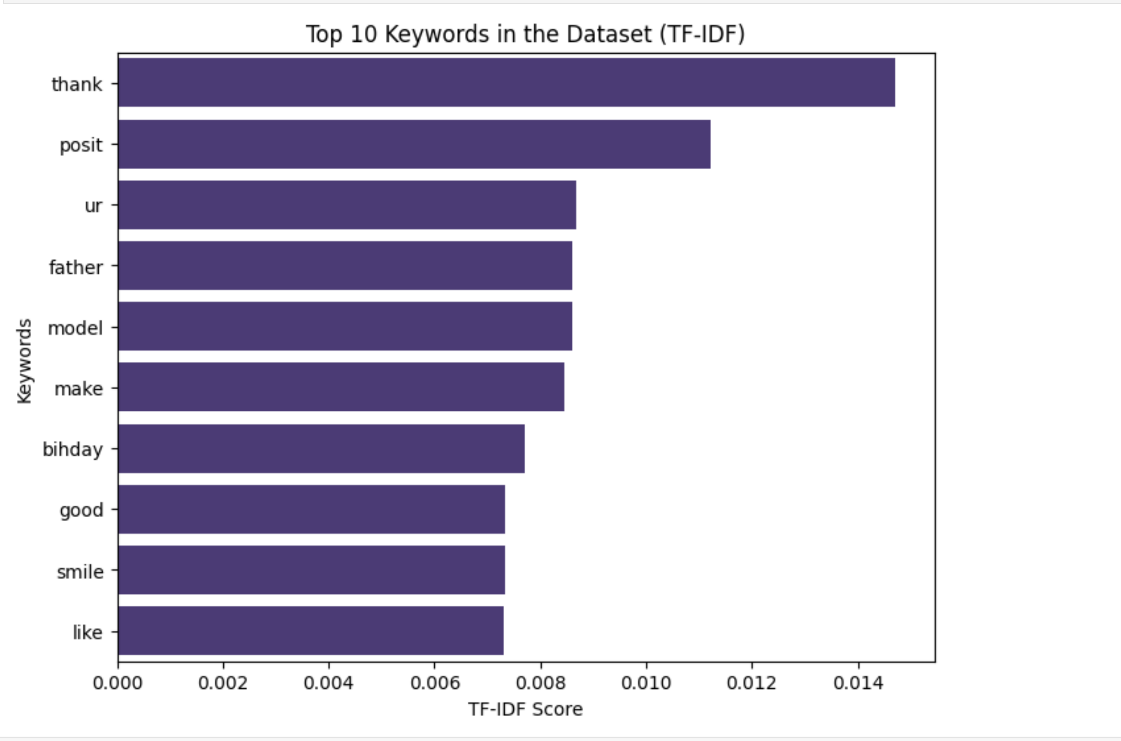
**Result of Dataset After all above steps:**



**TF-IDF features**

represent the importance of words in a dataset by considering both the frequency of the word in the Dataset and its rarity across the entire collection of Dataset (corpus).

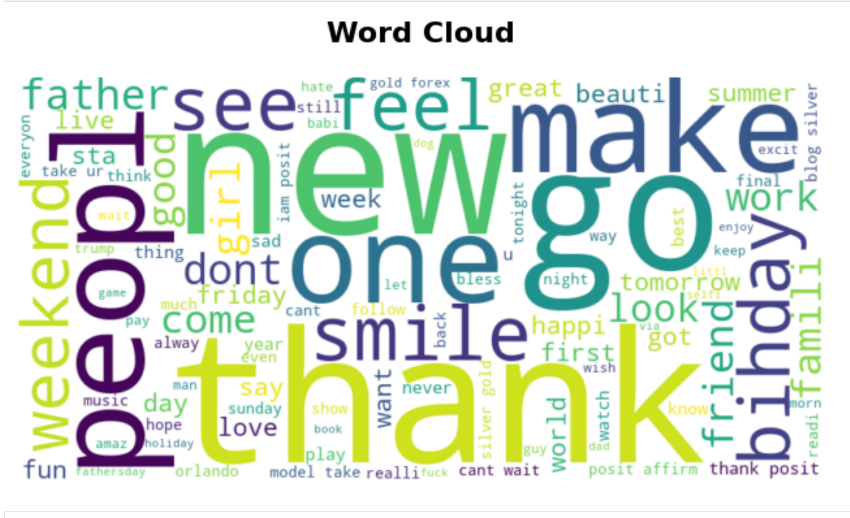
TF-IDF is essential for determining the importance of words in a Dataset , aiding in tasks like information retrieval, Dataset representation, and text mining.



This figure is showing important words in the dataset.

**Word Cloud**

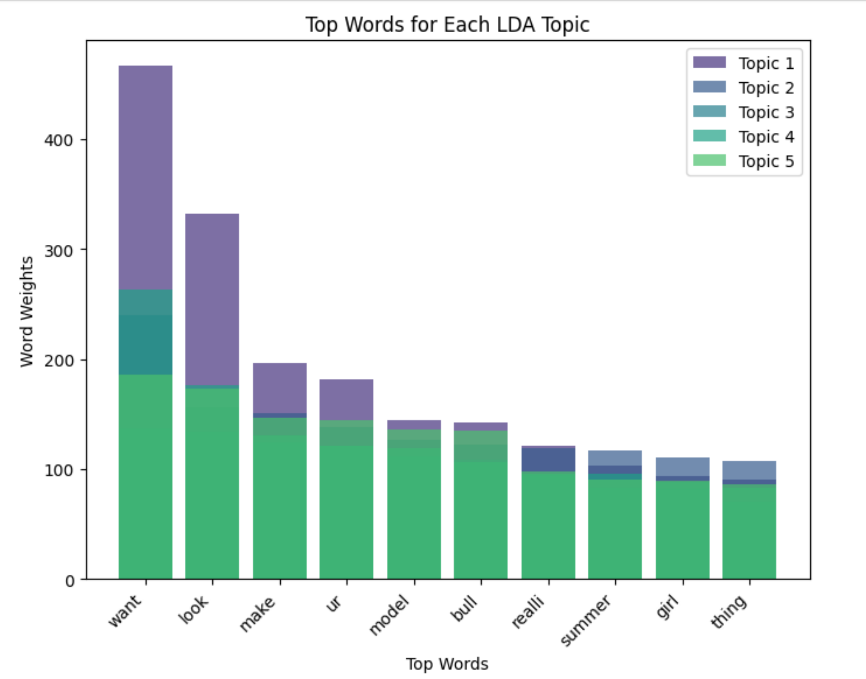
A word cloud is a visual representation of text data where the size of each word corresponds to its frequency.



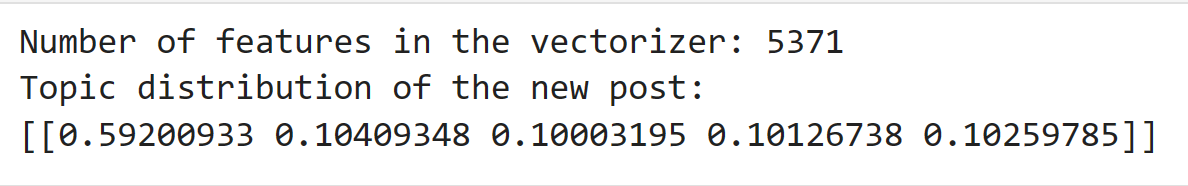
This figure is showing WordCloud of Dataset

**Latent Dirichlet Allocation (LDA)**

Performing Latent Dirichlet Allocation (LDA) on a social media dataset is valuable for uncovering latent topics or themes present in user-generated content. It aids in understanding the prevalent discussions, categorizing posts into relevant topics, and gaining insights into the diverse content shared on social platforms, essential for social media analysis and content recommendation.

****

When a new post is introduced, the LDA model assigns topics to it based on the words it contains. The model doesn't inherently "know" the topic; instead, it utilizes probabilities to estimate the likelihood of the post belonging to different topics based on its content and the learned topic distributions from the training data. The result of model for new text is like this:



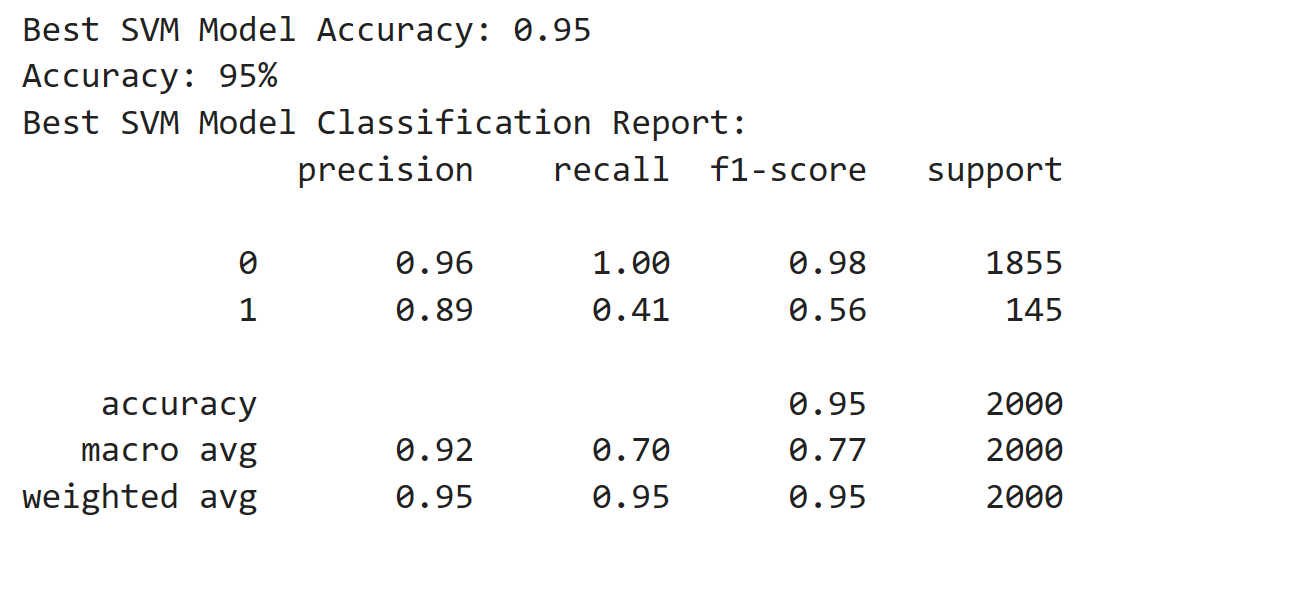
**Machine Learning**

We have tested two model on Dataset and we have Different Result   
**Support Vector Machine (SVM) Classifier for Social Media Sentiment Analysis:** SVM is a classification algorithm used in social media sentiment analysis to effectively categorize posts or tweets into different sentiment classes (such as positive, negative, or neutral) by finding the optimal hyperplane that separates these sentiments in a high-dimensional feature space.

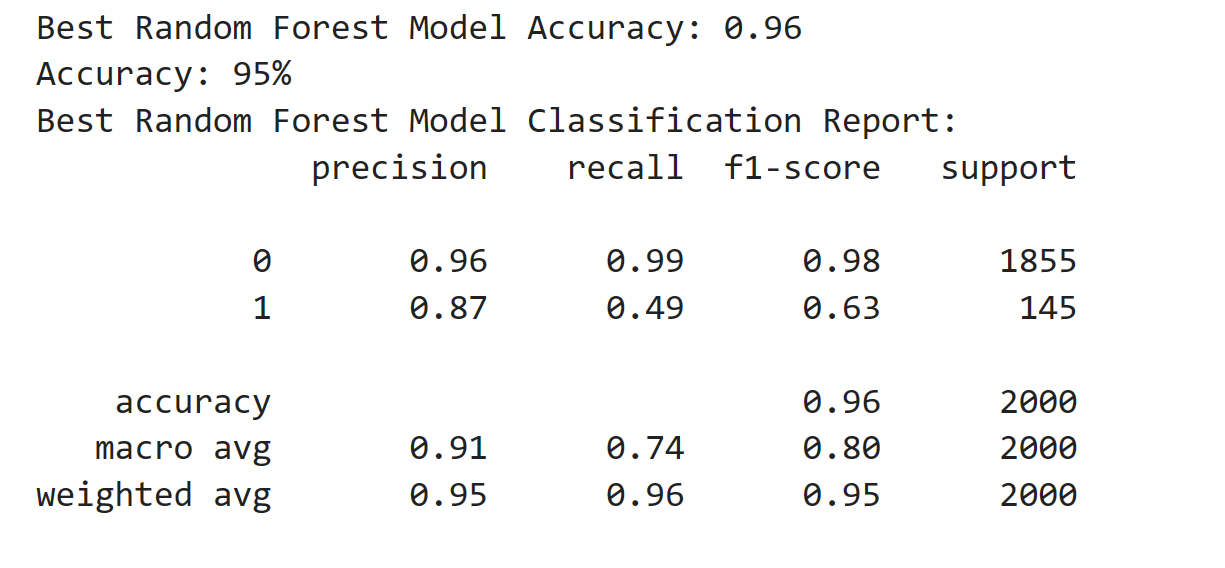
**Random Forest (RF) Classifier for Social Media Sentiment Analysis:** RF is an ensemble learning method employed in social media sentiment analysis to analyze and classify user-generated content. By constructing multiple decision trees and combining their outputs, RF enhances the accuracy of sentiment predictions, making it robust for handling the diverse and dynamic nature of social media language.

**Result**

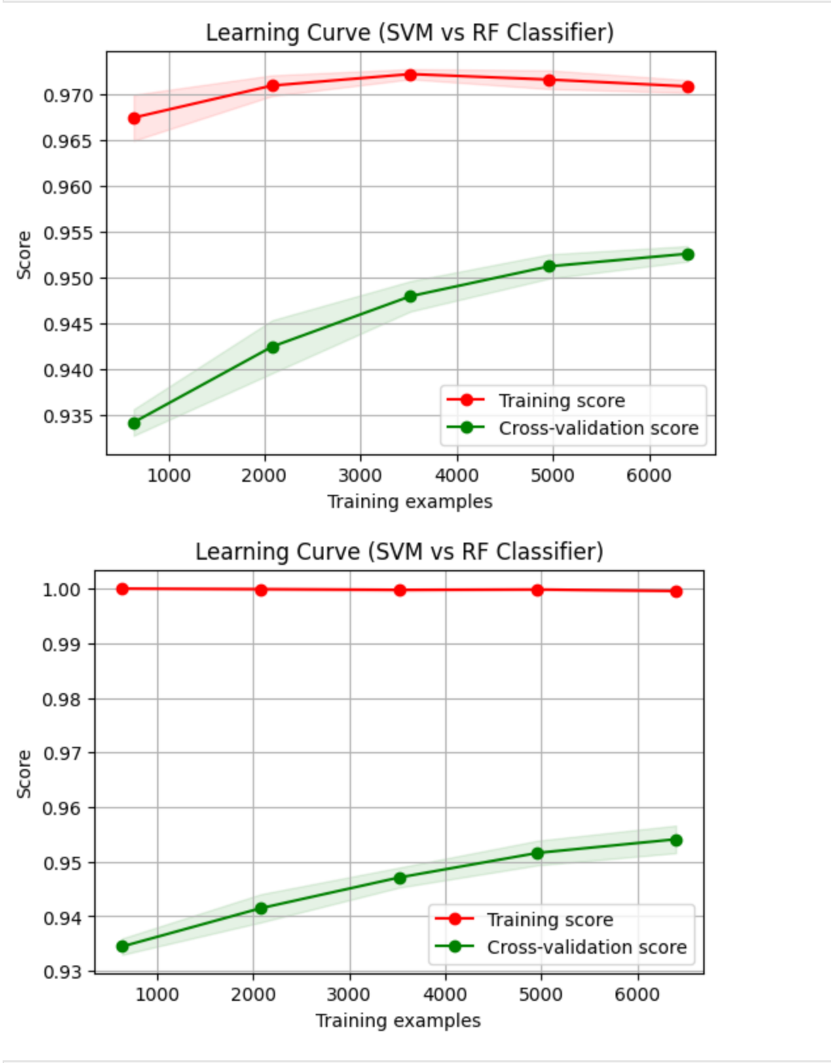
**1-SVM Model**

****

**2-RF Model**

****

**The result is not very difference**



**Training Score and Cross-validation score**

**High Training Score (Red Line):**

The model performs very well on the training data, achieving a high accuracy or score.

**Lower Cross-Validation Score (Green Line):**

The performance on the cross-validation set is not as high as on the training set, and it might not improve significantly with more data.

**Gap Between Lines:**

The gap between the two lines suggests that the model is overfitting to the training data. Overfitting occurs when the model captures noise or specific patterns in the training set that do not generalize well to new, unseen data.

Now we need **Regularization**

**It's up to you. Can you improve the results?**

**Conclusion**

While the SVM and Random Forest models demonstrated commendable accuracy in sentiment analysis, the observed overfitting in the learning curves highlights the importance of addressing generalization challenges. The iterative process of refining the models, incorporating regularization techniques, and fine-tuning hyperparameters will be crucial to achieving a robust sentiment analysis solution capable of providing reliable predictions on diverse datasets.

**References**

Cielen Davy Meysman Arno D. B and Ali Mohamed. 2016. Introducing Data Science : Big Data Machine Learning and More Using Python Tools. Shelter Island NY: Manning.

Russell, M. A. (2013). Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more. " O'Reilly Media, Inc.".