Ecole Supérieure
d'Informatique et du Numérique
COLLEGE OF ENGINEERING & ARCHITECTURE

Université Internationale de Rabat
THE INNOVATIVE UNIVERSITY

# Integrated Project's Report

## Sentimental Analysis
## Celebrities: Saad LAMJARED
## Simo Life

**Realized by:**
Imane NACIRI
Oussama OUTELHA
Aymane JOUHARI
Achraf EZZABDI
Ismail ASSIM

# Dedications:

To our beloved parents, who have selflessly given so much for our sake; for their boundless patience, unconditional love, and unwavering belief in us. They have gone above and beyond to ensure our happiness and accomplishments. May this humble endeavor serve as a token of our profound affection and unwavering loyalty. No dedication can adequately convey the depth of our gratitude towards them. May God grant them good health and a long, fulfilling life.

# Acknowledgement:

We wish to express our gratitude through these few lines of appreciation, first and foremost to our mentors, Mrs. Nada Sbihi, Mr. Mounir GHOGHO, Mr. Hakim HAFIDI, Mr. Youness MOUKAFIH, and Mr. Abdelghani GHANEM, for their invaluable advice and guidance. Their support and mentoring have played a vital role in the realization of this project.

We cannot conclude this project without expressing our deep gratitude to all the professors at the International University of Rabat, whose dedication and assistance have been pivotal throughout this year. Their commitment to education and their willingness to assist have greatly contributed to our growth and achievements.

We extend our sincere appreciation to all who have contributed to our project, and we are profoundly grateful for their unwavering support and guidance.

# Abstract

This sentiment analysis project aims to analyze and understand public sentiment towards two prominent celebrities, Saad Lamjarred and SimoLife. By applying advanced natural language processing and machine learning techniques to social media data, we will uncover valuable insights into how these celebrities are perceived by the public.

The project will involve collecting and preprocessing a comprehensive dataset of social media posts, comments, and discussions related to Saad Lamjarred and SimoLife. Through sentiment analysis algorithms, we will classify the sentiment of each post as positive, negative, or neutral.

By examining the sentiment distribution and identifying prevalent themes, emotions, and sentiments associated with each celebrity, we will gain a deep understanding of public perception. Additionally, we will explore potential factors influencing sentiment, such as significant events, controversies, or collaborations.

# Table of content:

# Figures list

# Introduction

This report focuses on a sentiment analysis project that specifically examines the sentiments expressed towards two popular celebrities, Saad Lamjarred and SimoLife. The project encompasses web scraping and data preprocessing, model training and evaluation, testing on unlabeled data, as well as analysis and visualization of the results.

By gathering data from various online sources, such as social media platforms, the project aims to understand the sentiment associated with Saad Lamjarred and SimoLife. This data is then processed and cleaned to ensure its accuracy and consistency.

Using machine learning or deep learning algorithms, sentiment analysis models are trained and evaluated to classify the collected text into positive, negative, or neutral sentiments. The models' performance is assessed using appropriate metrics to select the most effective one.

The chosen model is then tested on unlabeled data to evaluate its generalization capabilities and its ability to handle new, unseen text related to Saad Lamjarred and SimoLife. This testing phase provides insights into how well the model performs in real-world scenarios.

To enhance the interpretation of the sentiment analysis results, word cloud visualizations are employed. These visual representations highlight the most significant and frequently occurring words associated with Saad Lamjarred and SimoLife, offering a concise overview of the prevailing sentiments.

In summary, this sentiment analysis project aims to gain insights into the sentiments expressed towards Saad Lamjarred and SimoLife. By employing web scraping, data preprocessing, model training and evaluation, testing on unlabeled data, and visualizations, the project provides a comprehensive analysis of public sentiment towards these celebrities.

# Celebrities:

**Saad Lamjarred** is a Moroccan singer, songwriter, and actor who gained international recognition for his music, particularly in the genre of Moroccan pop (Raï). He is known for his unique musical style that combines traditional Moroccan sounds with contemporary pop elements.

Saad Lamjarred has faced legal issues and accusations of sexual assault in the past. In 2016, he was arrested in France on charges of sexually assaulting a woman. However, it's important to note that the charges against him were later dropped due to insufficient evidence. This incident has generated significant public debate and discussion surrounding Saad Lamjarred's personal life and image.

*Figure 9: Sadd Lamjared*

**Mohamed Baabit,** also known as Simo Life, is a popular Moroccan social media influencer and content creator. Simo Life gained fame through his comedic and entertaining videos on platforms such as Instagram and YouTube. Simo Life often incorporates relatable situations and engaging storytelling in his content, which has helped him amass a large fan base. His entertaining videos and charismatic personality have made him a well-known figure among Moroccan internet users.

*Figure 10: Simo LIFE*

# Scrapping & Preprocessing:

**Scrapping:** The first step of our project was to gather data. For that we ve used a python script and the selenium librairy automate the process of logging into Facebook, navigating to specific posts URL, and retrieving the comments in a list that will be used to create a pandas dataframe then saved in a csv file

- Getpass module: secure way to input passwords without displaying it on the screen.
- Time module: introduce time related fonctions for delays and pauses.
- Pandas library for data manipulation and analysis
- Selenium library for automating web browsers, interacting with web elements navigating through web pages.

**Labeling**: In the excel file, those comments were each manually assigned the appropriate class label : 1 if positive and -1 if negative that represent the sentiment we want to predict and analyze.

**Deleting stop words**: common words that don't contribute much to the overall meaning, removing them reduces the dimensionality focus on more important words and so improves the efficiency of the following steps.

**Deleting special characters:** irrelevant and can introduce noise, removing them enhance the quality of data and simplifies the following steps.

**Tokenization:** breaking down the text into individual small units called tokens, typically words or phrases. Tokenization is a fundamental step in natural language processing as it provides a structured representation of the text, making it easier to analyze and process. This step helps in counting word frequencies, understanding the structure of the text and analyzing the meaning of each word in the context.

**Stemming:** a technique that reduces words to their base or root form by removing suffixes and prefixes. For example, stemming can convert "running," "runs," and "ran" to the common stem "run." It helps in reducing the dimensionality of the data, consolidating related words, reduce sparsity in text data and achieving better generalization.

**TF-IDF:**

a numerical representation of the importance of each term (word) in a document within a collection of documents.

It considers both how frequent a term appears in a document (TF) [ appears more: high value] + its rarity across all documents (IDF) [ appears in fewer documents : high value ].

It helps to identify important and discriminative terms while downplaying common terms that appear in many documents.

**Importance**:

The importance of these preprocessing steps lies in cleaning and structuring the data, enhancing its quality, reducing noise, and transforming the raw text data into a suitable format for machine learning algorithms. They contribute to improving the efficiency, accuracy, and interpretability of the machine learning algorithms TF-IDF (t, d) = TF (t, d) * IDF(t)

# Splitting, training and model evaluation

## Algorithm Used:

In the sentiment analysis project focused on Saad Lamjarred and SimoLife, several algorithms can be considered for training the sentiment analysis model. Here are the algorithms that we utilized:

- **Naive Bayes:** is a probabilistic algorithm that assumes independence between features. It is known for its simplicity and efficiency in text classification tasks. Naive Bayes calculates the probability of a text belonging to a specific sentiment category based on the occurrence of words in the training data.

- **Logistic Regression:** is a widely used algorithm for binary classification tasks. It models the relationship between the independent variables (words or features) and the binary sentiment outcome. Logistic Regression provides interpretable coefficients that represent the influence of each feature on the sentiment prediction.

- **LDA (Latent Dirichlet Allocation):** algorithm that clusters text data into topics. It assumes that each document is a mixture of various topics, and each word is associated with a particular topic. LDA can provide insights into the underlying themes or topics within the text data, which can be useful for sentiment analysis.

- **QDA (Quadratic Discriminant Analysis):** is a classification algorithm that assumes different covariance matrices for each class. It can capture more complex relationships between features compared to thelinear assumptions of Logistic Regression. QDA may be suitable if there are non-linear relationships between features and sentiment.

- **Random Forest:** is an ensemble learning algorithm that combines multiple decision trees to make predictions. Each tree is trained on a different subset of the data, and the final sentiment prediction is determined by aggregating the results of all the trees. Random Forest can handle non-linear relationships and provide robust predictions.

- **Support Vector Classification:** is a powerful algorithm that separates data into different classes by finding an optimal hyperplane. SVC can handle both linear and non-linear relationships between features and is known for its ability to handle high-dimensional data effectively.

## Best Hyperparameters & Accuracies

**Saad LAMJARED:**

The choice of the best hyperparameters for sentiment analysis algorithms depends on the specific algorithm being used and the characteristics of the dataset. Here are some commonly tuned hyperparameters for the mentioned algorithms:

### Naive Bayes:

- 'alpha' : 0.1
- 'fit_prior: True

**Accuracies:**

- Training: 99%
- Validation: 88%
- Test: 92%

### Logistic Regression:

- Regularization parameter (C): 'C': 10
- 'penalty': True

**Accuracies:**

- Training: 99%
- Validation: 92%
- Test: 94%

LDA (Latent Dirichlet Allocation):

- 'solver': 'lsqr'

**Accuracies:**

- Training: 50%
- Validation: 43%
- Test: 44%

QDA (Quadratic Discriminant Analysis):

- 'reg_param': 0.1

**Accuracies:**

- Training: 99%
- Validation: 82%
- Test: 87%

Random Forest:

- 'n_estimators': 200
- Maximum depth: 'max_depth': 20

**Accuracies:**

- Training: 96%
- Validation: 89%
- Test: 93%

Support Vector Classification:

- Regularization parameter (C): 'C': 10
- Kernel type: 'Kernel':'linear'

**Accuracies:**

- Training: 99%
- Validation: 91%
- Test: 94%

**Simo Life:**

Naive Bayes:

No specific hyperparameters to tune.

**Accuracies:**

- Training: 98%
- Validation: 78%
- Test: 82%

Logistic Regression:

- Regularization parameter (C): 'C': 0.5
- 'penalty': '12'

**Accuracies:**

- Training: 98%
- Validation: 80%
- Test: 83%

LDA (Latent Dirichlet Allocation):

- 'reg_param': 0.01
- 'priors': None
- 'store_convariqnce': True
- 'tol': le-06

**Accuracies:**

- Training: 55%
- Validation: 65%
- Test: 63%

QDA (Quadratic Discriminant Analysis):

- 'solver': 'lsqr'
- 'shrinkage': 0.5

**Accuracies:**

- Training: 99%
- Validation: 79%
- Test: 81%

### Random Forest:

- Maximum depth: 'max_depth': 15
- 'min_samples_split': 10
- 'min_samples_leaf': 1
- 'n_estimators': 300

**Accuracies:**

- Training: 96%
- Validation: 89%
- Test: 93%

### Support Vector Classification:

- Regularization parameter (C): 'C': 10
- Kernel type: 'Kernel':'linear'
- Gamma='scale'

**Accuracies:**

- Training: 99%
- Validation: 91%
- Test: 94%

The "Splitting, Training, and Model Evaluation" phase involved dividing the dataset into training, validation, and test sets. Various sentiment analysis algorithms, including Naive Bayes, Logistic Regression, LDA, QDA, Random Forest, and Support Vector Classification, were trained and evaluated.

For Saad Lamjarred, the highest accuracies were achieved using Logistic Regression (94% on the test set) and Support Vector Classification (94% on the test set). Naive Bayes also performed well with an accuracy of 92% on the test set.

In the case of Simo Life, Naive Bayes and Logistic Regression yielded accuracies of 82% and 83% on the test set, respectively.

While LDA showed lower accuracies for both celebrities (around 44% for Saad Lamjarred and 63% for Simo Life), QDA achieved moderate accuracies (ranging from 79% to 81%). Random Forest exhibited higher accuracies, with test accuracies of 93% for both Saad Lamjarred and Simo Life.

Based on these results, Logistic Regression and Support Vector Classification demonstrated the best overall performance in sentiment analysis for both Saad Lamjarred and Simo Life.

# Testing the model on unlabeled data

After training the sentiment analysis models on labeled data, the next step is to test the models on unlabeled data to evaluate their performance in real-world scenarios. This involves applying the trained models to analyze the sentiment of text or comments related to Saad LAMJARED and Simo Life, which were not included in the training or evaluation stages.

By testing the models on unlabeled data, we can assess their ability to accurately classify sentiment and gain insights into public opinions surrounding these celebrities. The unlabeled data may consist of social media posts, news articles, or other textual content that mentions Saad LAMJARED or Simo Life.

During the testing phase, the sentiment analysis models will classify each text as positive, negative, or neutral based on the learned patterns and features from the training data. The predictions made by the models will be compared to the actual sentiment or manually labeled sentiment, if available, to evaluate their accuracy and performance.

The results of testing on unlabeled data will provide valuable information about how well the models generalize to real-world inputs and whether they can effectively capture the sentiment associated with Saad LAMJARED and Simo Life across various sources. This analysis will contribute to understanding the overall perception and sentiment towards the celebrities beyond the labeled data used for training and evaluation.

## Saad LAMJARED

**Before:**

| | |
|---|---|
| بغض النظر على ان على حق او لا، انتم تنتقدونه ونسيتم انكم فقط تنقصون من ذنوبه باغتيالهم له. | |
| الله يدير السلامة تاني | |
| لا حول ولاقوة قوة إلا بالله | |
| لي راجل راجل اكبر شماتة على وجه الكرة الأرضية هو هدا من تواضع لله رفعه | |
| الله استر نا حتى استرنا التراب | |
| لاحول ولا قوة الا بالله دبا هدا لبس يلبسوه الرجال | |
| اللهم استرنا فوق الارض وتحت الارض ويوم العرض | |
| تيبغي يدير فيها ديفيرون على الناس تيلبس شي دريالة و يشدها ب حزام جداه زعمة عميق وهو داير بحال شي بوهالي | |
| يا ربي الا ماتسمحلي منومكنحملوش ما.....مفهمتش...ممرجلش ناقص بزاف تلف ما عرف شنو يتبع....ولا حاااااااامد | |
| هد سعد ولى مشكوك في امره لبس ديالوا شاد لاحولة ولا قوة إلا بالله | |
| الله يلطف بنا | |
| ذهبت الرجولة الفن رجع عفن | |
| في اي بحث علمي سينفع به الانسانية | |
| السناسل فالعنق بحال شي مرة | |
| هاذا مكروه عند المغاربة ابوه كان محبوبا لدى الجميع اما هاد شبه الرجل في القليل كان يسمي نفسه سعيدة احسن لار | |
| هذا السروال كيفكرني كان عند جدي كيلبسو نهار لكيمشي يحصد | |

*Figure 11: Sadd Lamjared unlabeled data*

## After using Logistic Regression:

| | |
|---|---|
| 1 | بغض النظر على ان على حق او لا، انتم تنتقدونه ونسيتم انكم فقط تنقصون من ذنوبه باغتيالهم له. |
| 1 | الله يدير السلامة تاني |
| 1 | لا حول ولا قوة الا بالله |
| 1 | لا حول ولاقوة قوة إلا بالله |
| -1 | لي راجل راجل اكبر شماتة على وجه الكرة الأرضية هو هدا من تواضع لله رفعه |
| 1 | الله استر نا حتى استرنا التراب |
| 1 | لاحول ولا قوة الا بالله دبا هدا لبس يلبسوه الرجال |
| 1 | اللهم استرنا فوق الارض وتحت الارض ويوم العرض |
| -1 | تيبغي يدير فيها ديفيرون على الناس تيلبس شي دريالة و يشدها ب حزام جداه زعمة عميق وهو داير بحال شي بوها |
| -1 | يا ربي الا ماتسمحلي منومكنحملوش ما.....مفهمتش...ممرجلش ناقص بزاف تلف ما عرف شنو يتبع....ولا حاااااااامد |
| -1 | هد سعد ولى مشكوك في امره لبس ديالوا شاد لاحولة ولا قوة إلا بالله |
| 1 | الله يلطف بنا |
| 1 | ذهبت الرجولة الفن رجع عفن |
| 1 | في اي بحث علمي سينفع به الانسانية |
| -1 | السناسل فالعنق بحال شي مرة |
| -1 | هاذا مكروه عند المغاربة ابوه كان محبوبا لدى الجميع اما هاد شبه الرجل في القليل كان يسمي نفسه سعيدة احسن |
| 1 | هذا السروال كيفكرني كان عند جدي كيلبسو نهار لكيمشي يحصد |

*Figure 12: Saad Lamjared testing on unlabeled data*

## Simo Life

**Before:**

| | |
|---|---|
| 175 | بنادم وأصلا فيه للعضم وهدا سائر كيطرطق علينا |
| 176 | يقول المثل من جد وجد ومن زرع حصد يا اخي العزيز |
| 177 | صراحة كنحتارم هاد السيد الله يعطيك الصحة خاااي سيمو ريسبيكت ليك برووو |
| 178 | ميكروب |
| 179 | وفقكم الله خويا سيمو |
| 180 | بني وعلي وسير وخلي |
| 181 | كسيوق صورة ديالو |
| | Fouad Palo Mista Wadye Chriki Radi Mohamed Ali |
| 182 | كنت واخد وحد النظرة غالطة على هاد السيد ، تبارك الله عليه يستحق ما وصل اليه من نجاح |
| 183 | الله يعونوا او يعاون جميع المسلمين |
| 184 | هذا هو معنى القتال من اجل المستقبل |
| 185 | عاد فهمت مليء كيغول بغي يدير فيرم و يرعي فيها الغنام |
| 186 | دير دورة فتبياع لجعل |
| 187 | سيد هارب عليك بزاف وتبتقن لغات اجنبية وعندو عقلية رجل اعمال عالمي تحكرتي قدامو الصلعاني تقبك ومزقك باش ما بغ |
| 188 | بصدق كانحتارم هاد السيد الله يسهل على شبابنا |

*Figure 13: Simo Life unlabeled data*

## After using Logistic Regression (grid search):

| | | |
|---|---|---|
| 175 | بنادم واصلا فيه للعضم وهدا سائر كيطرطق علينا | -1 |
| 176 | يقول المثل من جد وجد ومن زرع حصد يا اخي العزيز | 1 |
| 177 | صراحة كنحتارم هاد السيد الله يعطيك الصحة خاااي سيمو ريسبيكت ليك برووو | 1 |
| 178 | ميكروب | -1 |
| 179 | وفقكم الله خويا سيمو | 1 |
| 180 | بني وعلي وسير وخلي | -1 |
| 181 | كسيوق صورة ديالو | -1 |
| 182 | Fouad Palo Mista Wadye Chriki Radi Mohamed Aliلى هاد السيد ، تبارك الله عليه يستحق ما وصل اليه من نجاح | 1 |
| 183 | الله يعونوا او يعاون جميع المسلمين | 1 |
| 184 | هذا هو معنى القتال من اجل المستقبل | 1 |
| 185 | عاد فهمت مليء كيغول بغي يدير فيرم و يرعي فيها الغنام | -1 |
| 186 | دير دورة فتبياع لجعل | -1 |
| 187 | سيد هارب عليك بزاف وتبتقن لغات اجنبية وعندو عقلية رجل اعمال عالمي تحكرتي قدامو الصلعاني تقبك ومزقك باش ما بغيت | -1 |
| 188 | بصدق كانحتارم هاد السيد الله يسهل على شبابنا | 1 |

*Figure 14: Simo Life testing on unlabeled data*

# Word Cloud

A word cloud is a visual representation of text data where the size of each word is proportional to its frequency or importance within the given context. In the context of sentiment analysis for Saad Lamjarred and Simo Life, creating word clouds can help to identify the most frequently occurring words or themes associated with their public perception.

To generate a word cloud, the textual data related to Saad Lamjarred and Simo Life, such as social media comments or news articles, is processed and analyzed. Commonly used words and stopwords (e.g., "the," "and," "is") are removed to focus on more meaningful and relevant terms.

The remaining words are then visualized in a cloud-like format, with larger and bolder words representing higher frequency or importance. This visual representation allows for quick and intuitive identification of the prominent themes or sentiments associated with Saad Lamjarred and Simo Life.

By examining the word cloud, one can gain insights into the public's perception of these celebrities. Positive or negative sentiment can be inferred based on the frequency and prominence of certain words or phrases. Additionally, recurring themes or topics that are closely associated with Saad Lamjarred and Simo Life can be identified.

# Analysis and visualization

**Saad LAMJARED**

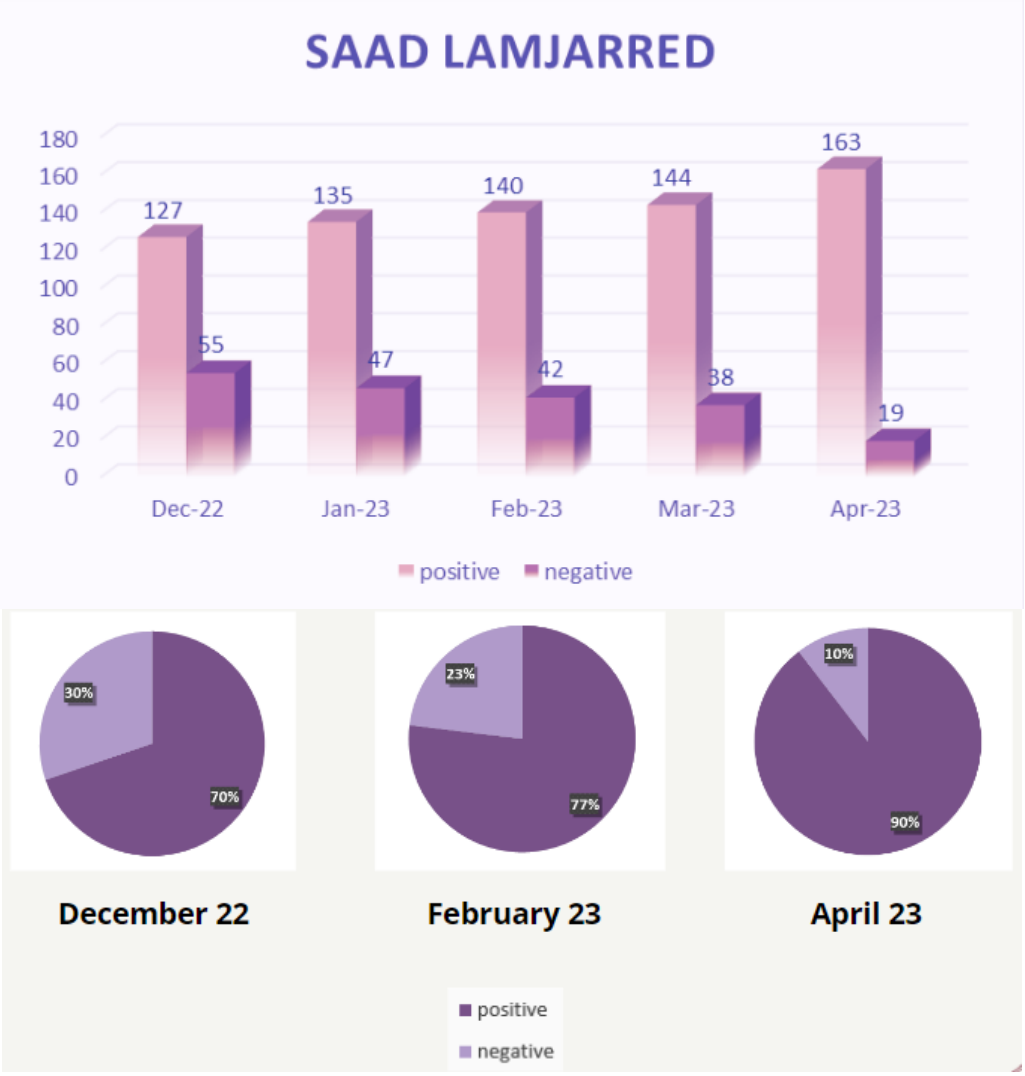


*Figure 15: Chart of the sentiments towards Saad LAMJARED*

The chart displaying the sentiments towards Saad Lamjarred from December 2022 to April 2023 reveals interesting patterns and trends. Here's an analysis of the sentiment data:

**Positive Sentiments:**

- The number of positive sentiments towards Saad Lamjarred shows an upward trend throughout the given period, starting from 127 positives in December 2022 and peaking at 168 positives in April 2023.
- The steady increase in positive sentiments suggests a growing appreciation, support, and admiration for Saad Lamjarred during this time.
- The significant jump from February to March and a further increase in April indicate a surge in positive sentiment towards Saad Lamjarred during those months.

**Negative Sentiments:**

- The number of negative sentiments shows a consistent decrease over time, starting from 55 negatives in December 2022 and reaching a low of 19 negatives in April 2023.
- The declining trend in negative sentiments implies a reduction in criticism or unfavorable opinions about Saad Lamjarred during the analyzed period.
- The relatively low number of negative sentiments compared to positives suggests a generally positive perception of Saad Lamjarred among the public during the specified timeframe.

## Simo Life

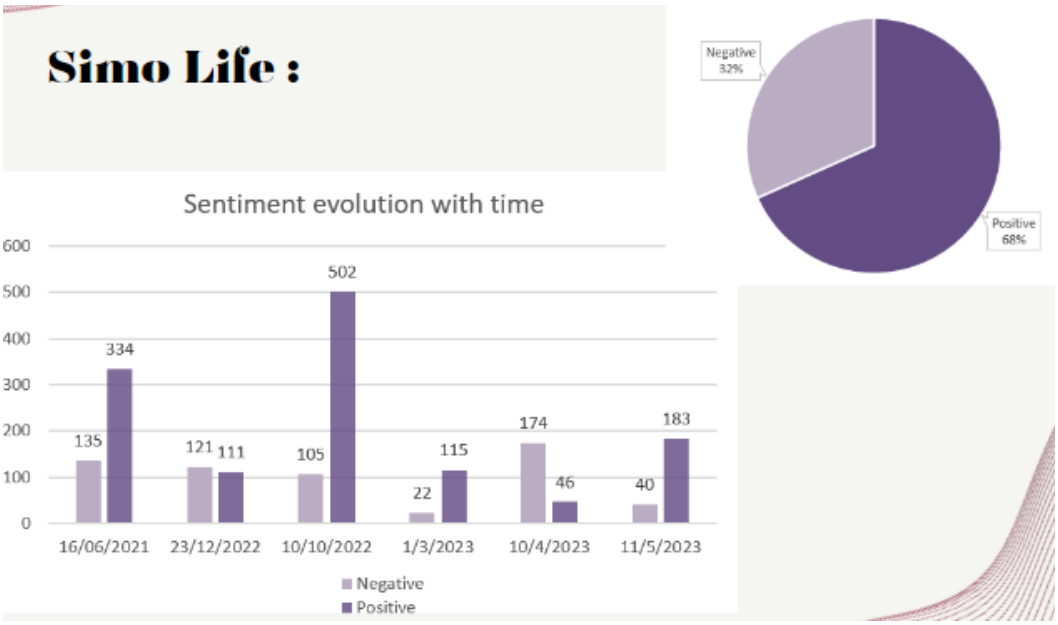


*Figure 16: Chart of the sentiments towards Simo Life*

The chart displaying the sentiments towards Simo Life from June 2021 to May 2023 provides insights into the evolving perception of the public. Here's an analysis of the sentiment data:

**Positive Sentiments:**

- Positive sentiments towards Simo Life show fluctuating trends over the analyzed period.
- Notable spikes in positive sentiments occurred in October 2022 with 502 positives and May 2023 with 183 positives.
- These peaks indicate instances when the public expressed higher levels of appreciation, support, and positive opinions towards Simo Life.

**Negative Sentiments:**

- The chart reveals varying levels of negative sentiments towards Simo Life throughout the analyzed period.
- The highest number of negative sentiments occurred in December 2022 with 121 negatives, followed by April 2023 with 174 negatives.
- These peaks indicate periods when the public expressed more criticism or unfavorable opinions towards Simo Life.

## Results and Analysis

In the sentiment analysis of Saad Lamjarred and Simo Life, we examined the public sentiments towards these two celebrities from June 2021 to May 2023. The sentiment chart provided valuable insights into the shifting perceptions of both individuals. For Saad Lamjarred, the chart indicated a positive trend with increasing positive sentiments and decreasing negative sentiments over time. Particularly noteworthy were the significant spikes in positive sentiments in April 2023 and the decline in negative sentiments throughout the analyzed period. This suggests a growing appreciation and support for Saad Lamjarred among the public. On the other hand, sentiments towards Simo Life exhibited more fluctuations and variations. While there were notable peaks in positive sentiments in October 2022 and May 2023, there were also spikes in negative sentiments in December 2022 and April 2023. These fluctuations indicate a mixed perception of Simo Life among the public, with both positive and negative sentiments expressed. Overall, the sentiment analysis provides valuable insights into the evolving public perception of Saad Lamjarred and Simo Life and can be used to understand the dynamics of their popularity and reputation over the analyzed period.

# Conclusion

This report presents the findings of a sentiment analysis project that aimed to understand public sentiment towards Saad Lamjarred and SimoLife, two popular celebrities. By gathering data from various online sources and applying preprocessing techniques, the project obtained accurate and consistent sentiment data. Multiple sentiment analysis algorithms were trained and evaluated, with varying performance for each celebrity. Testing the models on unlabeled data provided insights into their real-world applicability. Word cloud visualizations highlighted significant themes and sentiments associated with the celebrities. The analysis revealed increasing positive sentiment towards Saad Lamjarred over time, while sentiment towards SimoLife remained mixed. Overall, the project offered valuable insights into the public perception of these celebrities.