CRATA AI

04 March 2023

Sustainability Classifier by Jack Neville

Executive Summary

The Sustainability Classifier is designed to automate the process of classifying companies based on their focus on sustainability. Developed for an investment firm specialising in sustainable and environmentally focused projects, this tool leverages advanced machine learning algorithms to analyse the "About" section of LinkedIn profiles, determining the company's alignment with sustainability goals. Utilising a Python-based technology stack, including Flask for the web application framework and scikit-learn for machine learning, the project successfully delivers an easy-to-use platform for immediate classification results, aiding the investment decision-making process. Code can be viewed on GitHub.

Solution Overview

The Sustainability Classifier Web App addresses this challenge by automating the classification process through a machine learning model that assesses textual data from LinkedIn profiles.

Technology Stack:

- **Python:** Serves as the core programming language for developing both the web application logic and the machine learning model.
- **Flask**: A lightweight and flexible Python web application framework used to create the web interface for submitting LinkedIn URLs and displaying the classification results.
- **scikit-learn**: A Python library for machine learning, used to train the classification model on pre-labeled data, enabling it to predict the sustainability focus of new companies.
- **Beautiful Soup and Selenium (Chromedriver)**: Initially used for scraping LinkedIn profiles for the "About" section text. (Future iterations may transition to using LinkedIn's official API for more reliable data access.)

General Architecture:

The application follows a client-server model, where the user interacts with a web-based frontend, submitting LinkedIn URLs for analysis. The Flask backend receives these requests, processes the URLs, and extracts the "About" section of the LinkedIn profiles. This text is then fed into the pre-

trained machine learning model, which classifies the company as either "Sustainable" or "Not Sustainable" based on learned patterns from the training dataset.

The classification result is sent back to the frontend and displayed to the user, providing immediate feedback on the sustainability focus of the analysed company. This integration of web technologies and machine learning offers a scalable solution to the investment firm's challenge.

Steps

1. Initial Dataset Analysis and Insights

The foundation of the solution is a robust analysis of the provided dataset, which includes LinkedIn profiles labeled for sustainability. Key steps in our dataset analysis included:

- **Key Columns:** Loaded only key columns for analysis (‘about’ and ‘Label’)
- **Preprocessing function**:
  - **Cleaning:** Remove URLs, special characters, and numbers, standardising the text for analysis
  - **Tokenisation**: Broke down the cleaned text into individual words or tokens to facilitate further processing.
  - **Lemmatisation:** applied different process whether language was detected as Spanish or English.
- **Graphs** (appendix.)**:**
  - Number of sustainable to non-sustainable companies in data set.
  - Length of bios of sustainable company compared to non-sustainable.
  - Most common words for sustainable and non-sustainable companies.
  - n-Grams Analysis

Insights Gained:

- **Distribution of Classes**
  - **Observation:** The dataset contains significantly more non-sustainable companies than sustainable ones.
  - **Decision:** Implement weight balancing in the model to account for class imbalance.
- **Text Length and Sustainability**
  - **Observation:** Sustainable companies tend to have longer bios compared to non-sustainable companies.
  - **Decision:** Incorporate the length of the company bio as a feature in the model.
- **Common Words and n-Grams Analysis**
  - **Observation:** Certain phrases and words frequently appear in sustainable companies' bios but not in those of non-sustainable companies.

    - **Decision:** Utilise n-Grams and/or specific keywords as features in the model to capture these distinctions.

These insights directly informed the subsequent model training, emphasising features and text characteristics that are indicative of a company's focus on sustainability.

2. Training of a Classification Algorithm

Building on the dataset analysis, the model training phase involved:

- **Lemmatisation**: Applied to standardise words to their base or dictionary form, improving the model's ability to recognise similar words.
- **Weighting for Balance**: Employed class weighting to address any imbalances in the dataset, ensuring the model does not favor the more prevalent class.
- **Model Selection and Training**: Chose Logistic Regression for its interpretability and suitability for binary classification tasks.
- **Pipeline:** Leveraged a pipeline approach that facilitated streamlined experimentation with various feature extraction techniques and logistic regression parameters. This method evaluated the impact of different configurations on model performance.
- **Regularisation strength:** Through above mentioned pipeline experimentation, found a regularisation strength (C) of 10 significantly outperformed lower values such as 1 or 0.1. This indicated that a slightly less stringent regularisation, allowing for more model complexity, yielded better accuracy and model performance overall.

In the model optimisation phase, a detailed comparison was conducted between the enhanced and base Logistic Regression models. Despite the integration of advanced features such as n-grams, sustainable keywords, and adjustments for class imbalance, the base model—with a configuration that did not include Spanish language inclusion—consistently outperformed its enhanced counterparts, as evidenced by its scores:

- **Accuracy**: 87.31%
- **Precision**: 86.22%
- **Recall**: 87.31% (a critical metric, given our focus on optimising the recall score to prioritise the identification of sustainable companies, even at the risk of increasing false positives)
- **F1 Score**: 85.34%
- **Confusion Matrix**: True Negatives (474), False Positives (11), False Negatives (63), True Positives (35).

This empirical evidence underscored the base model's efficiency in distinguishing between sustainable and non-sustainable companies. Further analysis revealed that adjusting the regularisation strength to 10 improved recall scores compared to the lower strengths of 1 or 0.1, illustrating the delicate balance required to capture as many sustainable companies as possible while managing model complexity.

3. Model in Production

The final phase involved deploying the trained model into a production environment where it could be easily used by non-tech users:

- **Flask Web Application**: Developed a user-friendly interface where users can submit LinkedIn URLs for analysis. The backend processes these requests, applies the model, and returns the sustainability classification. Instructions to load the application locally can be found in the readMe of the GitHub page.

Future Enhancements:

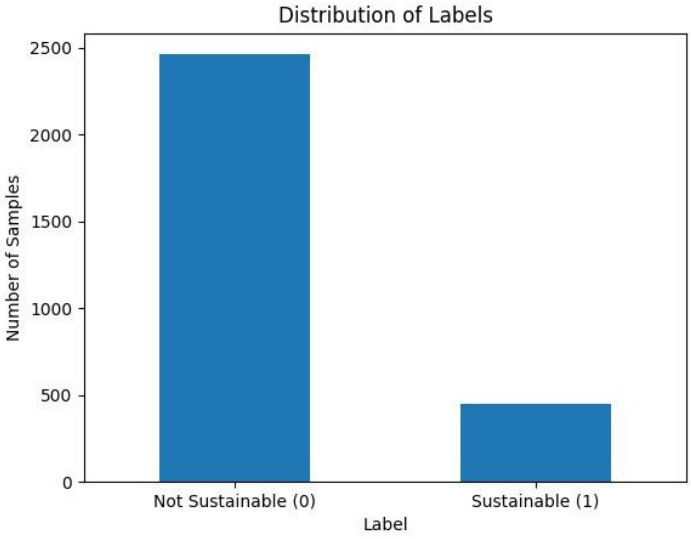
- **LinkedIn API Integration:** Enhance data accuracy and ensure compliance by adopting LinkedIn's official API for data retrieval. This approach promises more reliable and structured access to profile information, facilitating improved analysis.
- **User Feedback for Model Refinement**: Establish a feedback loop enabling users to report on the model's classification accuracy. This valuable input can inform ongoing model training cycles, allowing for iterative refinement and adaptation to emerging data trends.
- **Consider other model:** Exploring alternative machine learning models could yield benefits. For instance, exploring ensemble methods like Random Forests or advanced techniques like Gradient Boosting Machines (GBM) might offer enhanced predictive performance or interpretability. Additionally, experimenting with deep learning architectures, such as BERT or GPT, could provide significant advancements in understanding complex textual data. The initial decision to use Logistic Regression was favoured for its simplicity and interpretability.

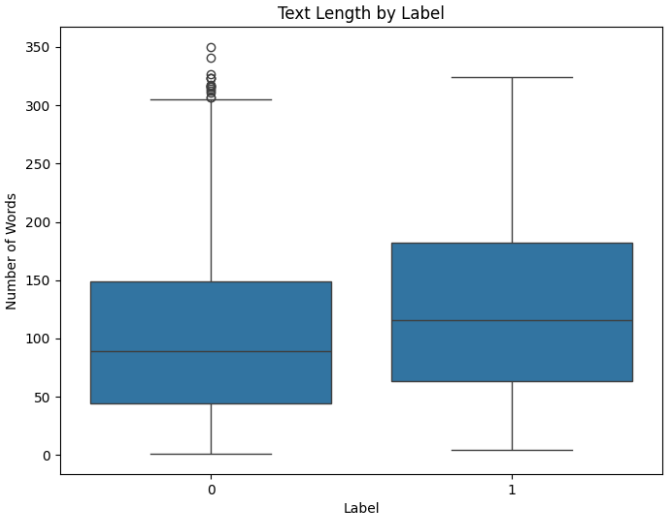
Conclusion

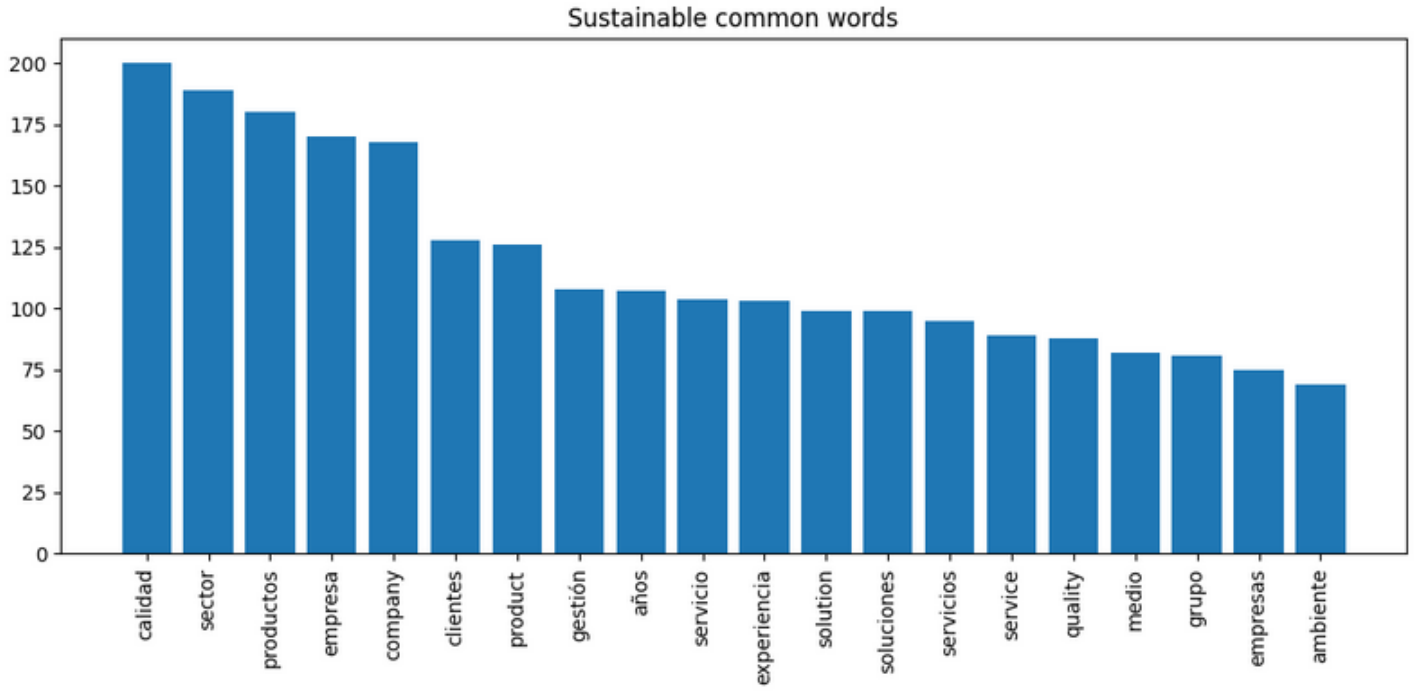
The development of the Sustainability Classifier has given me an insightful journey into the capabilities and limitations of machine learning for sustainability-focused classification. Despite extensive efforts to refine the model through advanced preprocessing and feature engineering, the base Logistic Regression model emerged as the most effective.

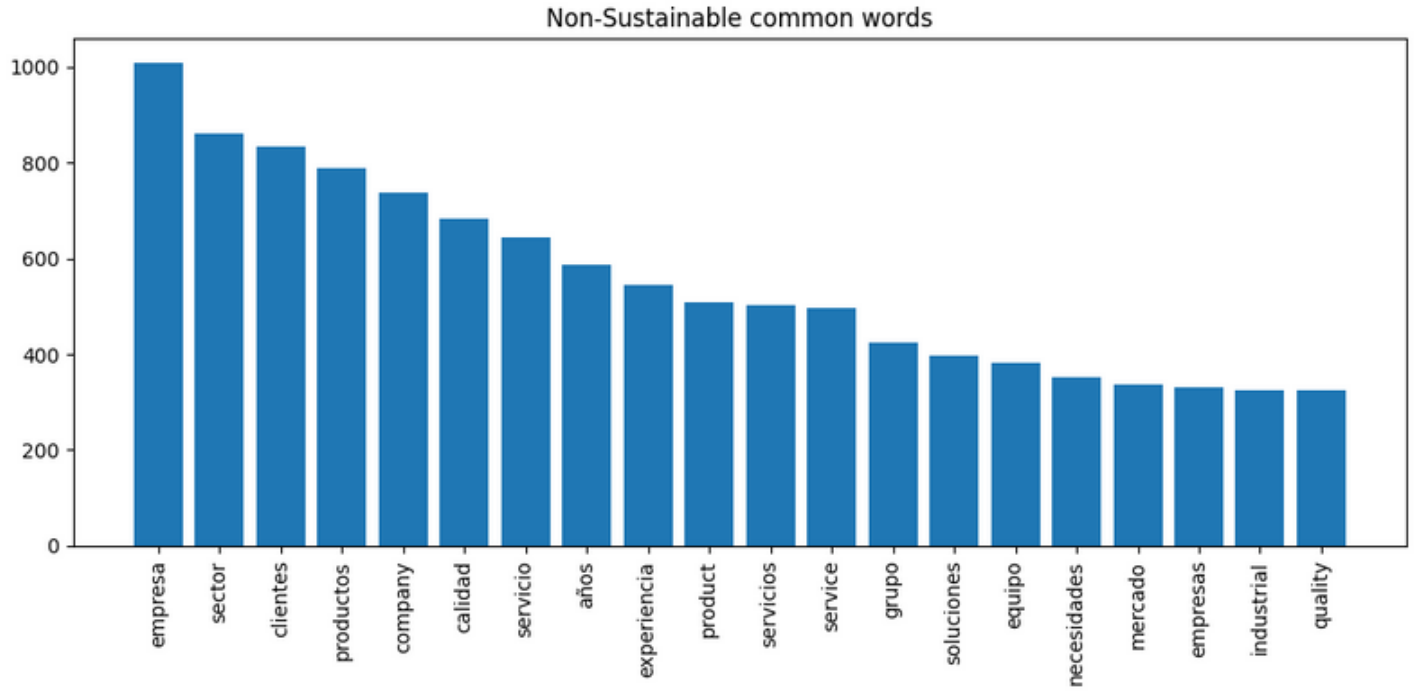
This experience highlights the nuanced balance between model complexity and performance. Moving forward, data accuracy can be enhanced with LinkedIn API integration and to refine the model further based on user feedback, continuing to improve investment decisions with AI-driven insights.

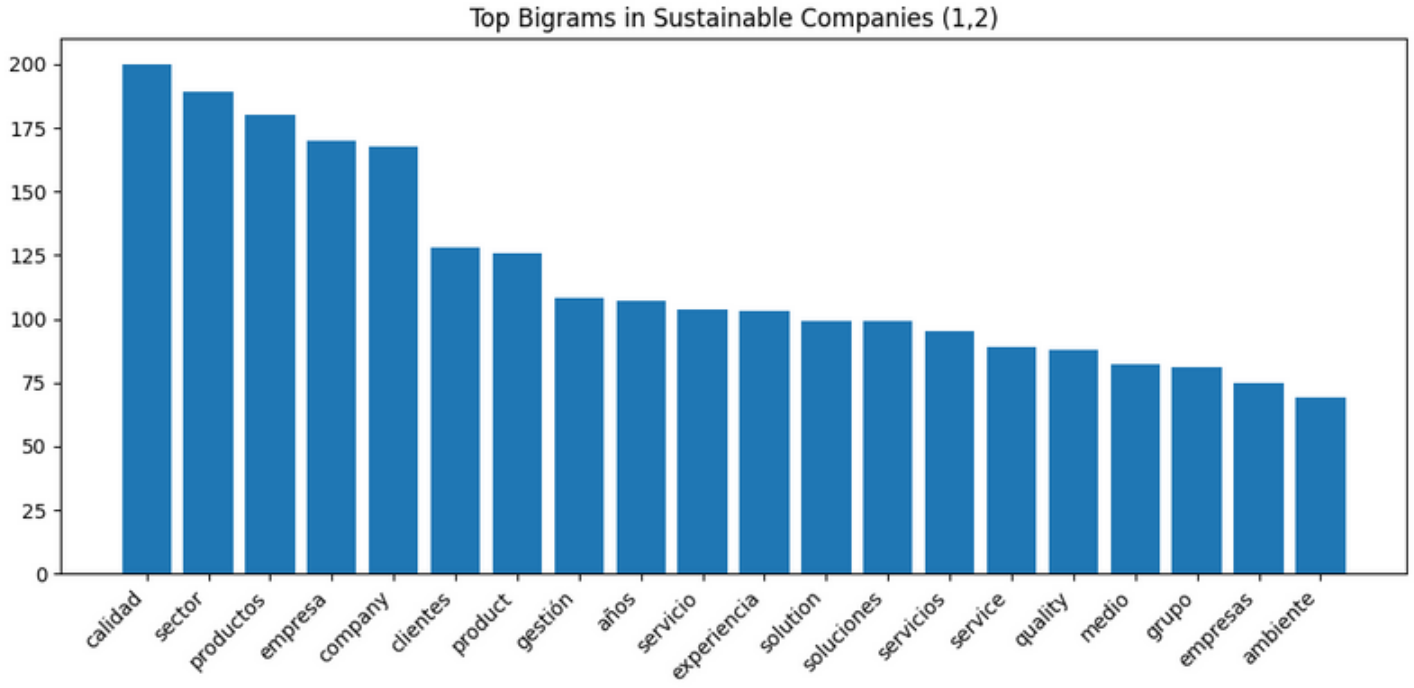
Appendix

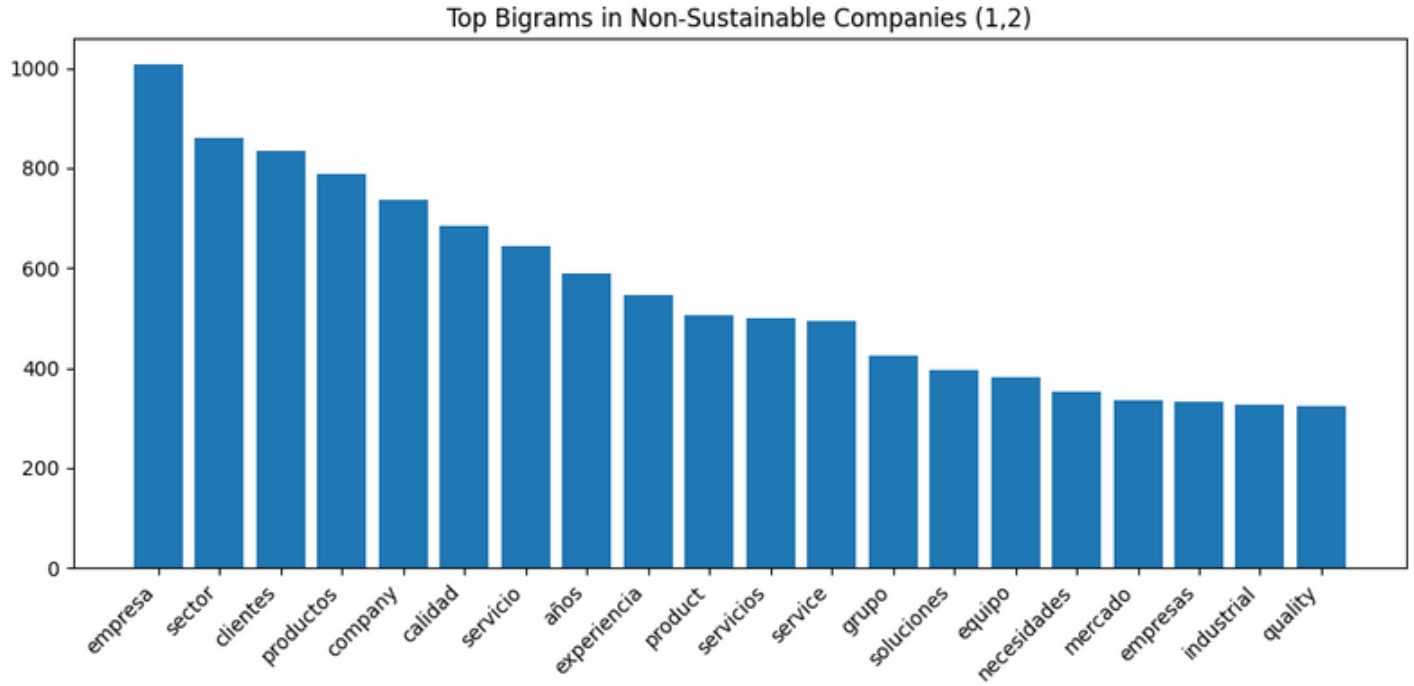

Distribution of Labels
Number of Samples
2500
2000
1500
1000
500
0
Not Sustainable (0)
Sustainable (1)
Label

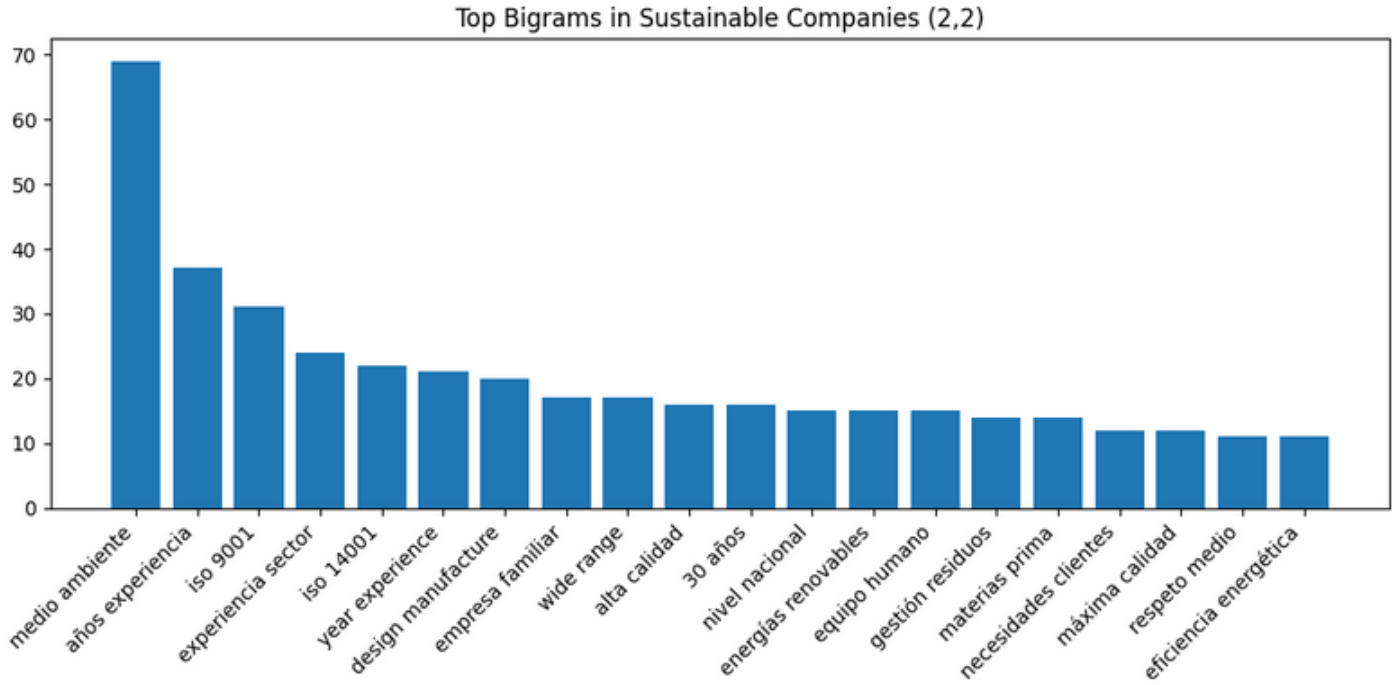

Text Length by Label
Number of Words
350
300
250
200
150
100
50
0
0
1
Label

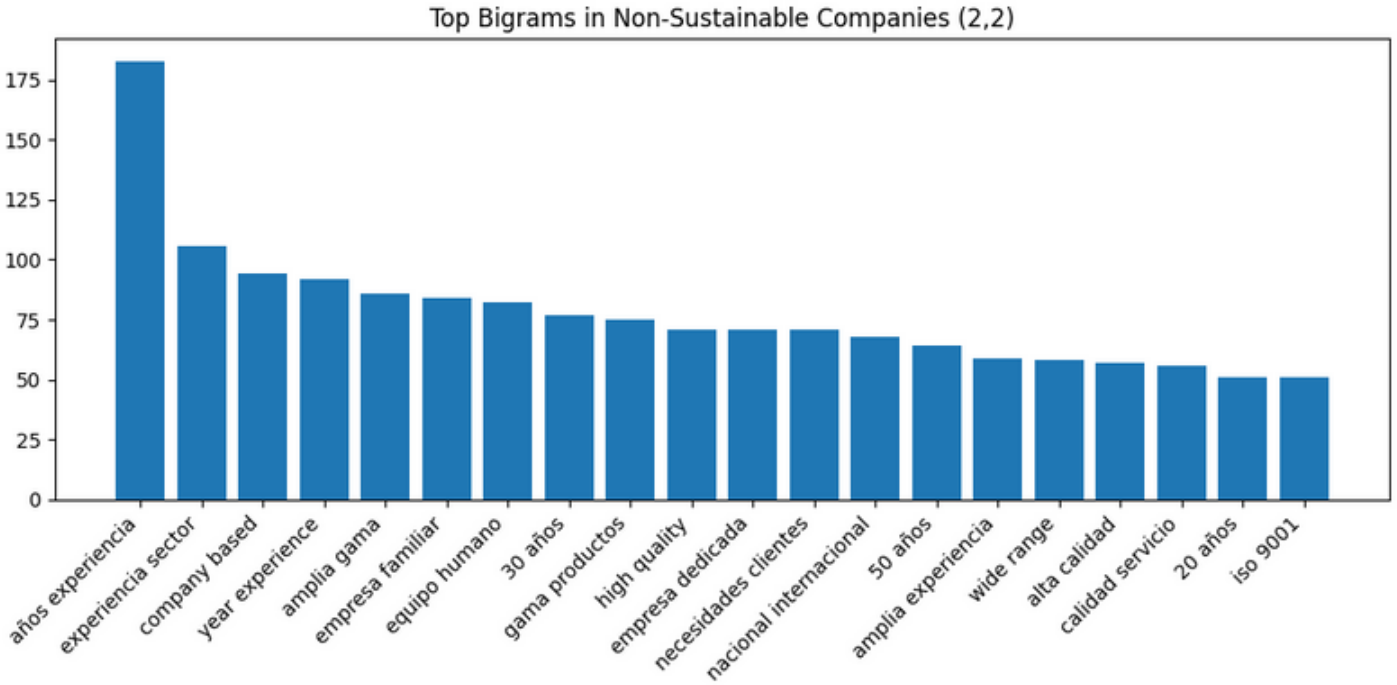

Sustainable common words
200
175
150
125
100
75
50
25
0
calidad
sector
productos
empresa
company
clientes
product
gestión
años
servicio
experiencia
solution
soluciones
servicios
service
quality
medio
grupo
empresas
ambiente

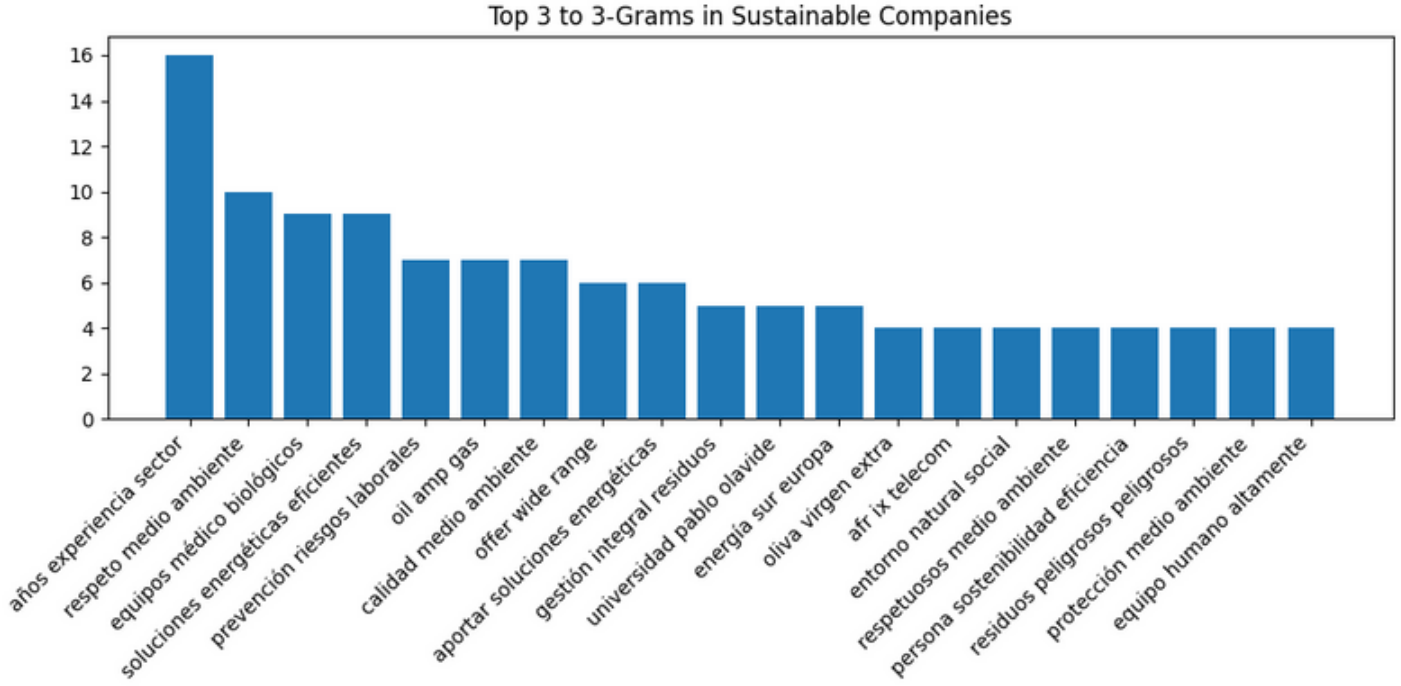

Non-Sustainable common words
1000
800
600
400
200
0
empresa
sector
clientes
productos
company
calidad
servicio
años
experiencia
product
servicios
service
grupo
soluciones
equipo
necesidades
mercado
empresas
industrial
quality

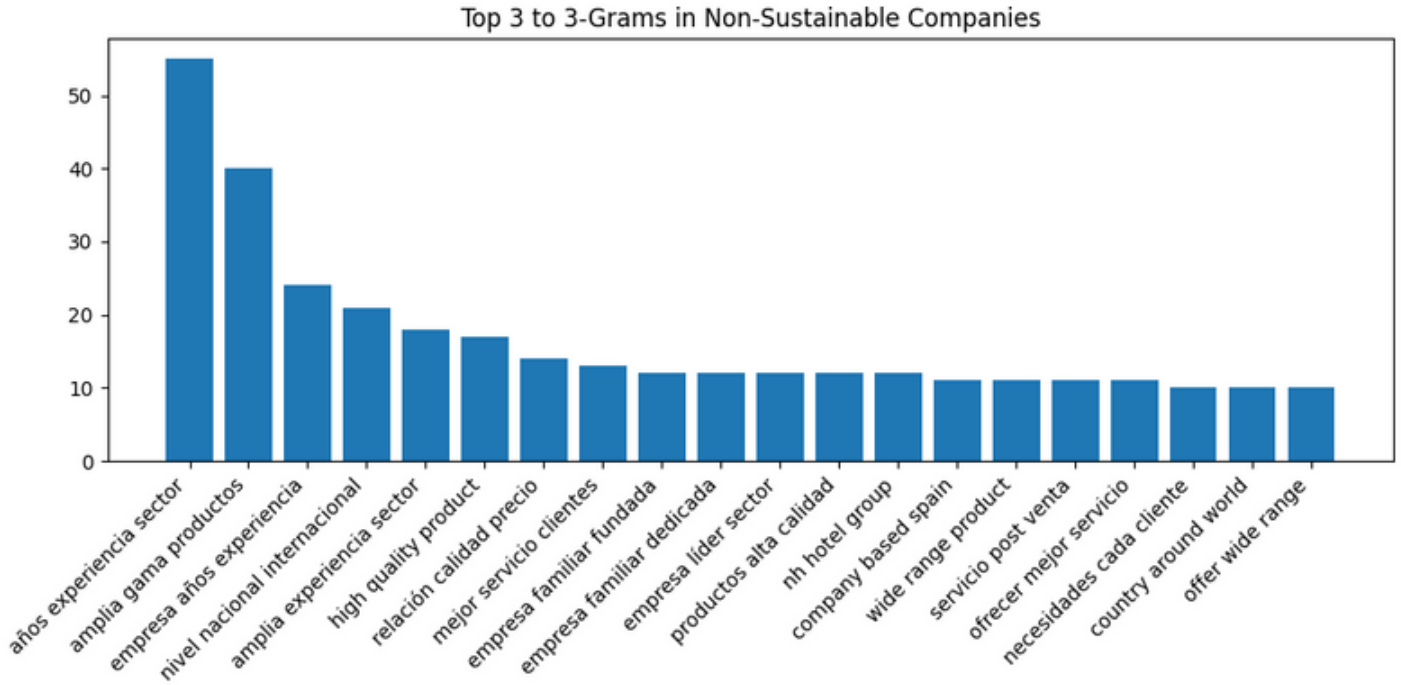

Top Bigrams in Sustainable Companies (1,2)
200
175
150
125
100
75
50
25
0
calidad
sector
productos
empresa
company
clientes
product
gestión
años
servicio
experiencia
solution
soluciones
servicios
service
quality
medio
grupo
empresas
ambiente


Top Bigrams in Non-Sustainable Companies (1,2)
1000
800
600
400
200
0
empresa
sector
clientes
productos
company
calidad
servicio
años
experiencia
product
servicios
service
grupo
soluciones
equipo
necesidades
mercado
empresas
industrial
quality


Top Bigrams in Sustainable Companies (2,2)
70
60
50
40
30
20
10
0
medio ambiente
años experiencia
iso 9001
experiencia sector
iso 14001
year experience
design manufacture
empresa familiar
wide range
alta calidad
30 años
nivel nacional
energías renovables
equipo humano
gestión residuos
materias prima
necesidades clientes
máxima calidad
respeto medio
eficiencia energética


Top Bigrams in Non-Sustainable Companies (2,2)
175
150
125
100
75
50
25
0
años experiencia
experiencia sector
company based
year experience
amplia gama
empresa familiar
equipo humano
30 años
gama productos
high quality
empresa dedicada
necesidades clientes
nacional internacional
50 años
amplia experiencia
wide range
alta calidad
calidad servicio
20 años
iso 9001


Top 3 to 3-Grams in Sustainable Companies
0
2
4
6
8
10
12
14
16
años experiencia sector
respeto medio ambiente
equipos médico biológicos
soluciones energéticas eficientes
prevención riesgos laborales
oil amp gas
calidad medio ambiente
offer wide range
aportar soluciones energéticas
gestión integral residuos
universidad pablo olavide
energía sur europa
oliva virgen extra
afr ix telecom
entorno natural social
respetuosos medio ambiente
persona sostenibilidad eficiencia
residuos peligrosos peligrosos
protección medio ambiente
equipo humano altamente


Top 3 to 3-Grams in Non-Sustainable Companies
0
10
20
30
40
50
años experiencia sector
amplia gama productos
empresa años experiencia
nivel nacional internacional
amplia experiencia sector
high quality product
relación calidad precio
mejor servicio clientes
empresa familiar fundada
empresa familiar dedicada
empresa líder sector
productos alta calidad
nh hotel group
company based spain
wide range product
servicio post venta
ofrecer mejor servicio
necesidades cada cliente
country around world
offer wide range