SENTIMENT ANALYSIS IN TWITTER USING ORANGE

A THESIS SUBMITED TO

THE FACULTY OF ARCHITECTURE AND ENGINEERING

OF

EPOKA UNIVERSITY

BY

REI NURIU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

FEBRUARY 2023

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled **“**Sentiment analysis in Twitter using Orange**”** and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

____

Dr. Arban Uka
Head of Department
Date: __________

Examining Committee Members:

Assoc. Prof. Dr. Carlo Ciulla	(Computer Engineering)	________________
Dr. Arban Uka	(Computer Engineering)	________________
Dr. Shkelqim Hajrulla	(Computer Engineering)	________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Rei Nuriu

Signature: _____________

ABSTRACT

SENTIMENT ANALYSIS IN TWITTER USING ORANGE

Nuriu Rei

Master of Science Department of Computer Engineering

Supervisor: Dr. Arban Uka

In this thesis the fundamentals of Natural Language Processing (NLP) will be presented. All the methods that I have tested are performed in a platform called Orange which is going to be explained throughout the thesis.

The main focus of the thesis is analyzing the semantic meaning of a large dataset of tweets. This dataset will be about tweets regarding a luxury brand named "Balenciaga". Anyway, until the reach of our final goal there are a vast number of other methods that provide data/text analytics of the tweets that our dataset contains. Each step that I have followed will be illustrated and explained in the thesis. At the end I have had some results and conclusions.

Keywords: *NLP, Orange, twitter, data analytics, sentiment analysis, dataset*

ABSTRAKT

ANALIZE SENTIMENTI NE TWITTER DUKE PERDORUR ORANGE

Nuriu Rei

Master Shkencor, Departamenti i Inxhinierisë Kompjuterike

Udhëheqësi: Dr. Arban Uka

Kjo tezë do t'i prezantojë lexuesit me Përpunimin e gjuhës natyrale (NLP). Gjatë tezës jam munduar që të përkufizoj dhe të shpjegoj bazat e Përpunimit të gjuhës natyrale. Të gjitha metodat që kam testuar janë realizuar në një platformë që quhet Orange, e cila do të shpjegohet përgjatë tezës.

Fokusi kryesor I tezës eshte analiza e sentimentit të nje bashkësie te madhe me tweet-e. Kjo bashkësi do të jetë rreth tweeteve për nje firme të njohur, luksi që quhet Balenciaga. Gjithsesi, derisa të arrijme në qëllimin tonë përfundimtar do të shohim një numër te gjerë metodash të tjera të cilat performojnë analiza të dhënash dhe teksti të tweet-eve që përmban bashkësia që kemi zgjedhur. Cdo hap që kam ndjekur është i ilustruar dhe i shpjeguar gjatë tezës. Në fund kam arritur në nxjerrjen e disa rezultateve dhe konkluzioneve.

Fjalë kyçe: Përpunimi I gjuhës natyrale, Orange, twitter, analizë të dhënash, analizë sentimenti, bashkësi.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Arban Uka for his guidance and support during all the stages of my thesis.

I would also like to thank Dr. Igli Hakrama who guided me in the Graduation Project subject, which was a big help for making this thesis.

A big thank goes to all the professors of Epoka University who guided me through all this journey.

Also, I would especially thank my family for all their support and help during my all studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

These days Only 21% of the data is in structured form. Other data is generated as we speak, tweet and send messages by using whatsapp, instagram and other platforms. Majority of this data is in textual form which is highly unstrctured in nature. In order to produce significant and actionable insights from this text data it is important to get acquainted with the techniques of NLP. NLP is a part of comp science and AI which deals with human languages.

First of all, we need to understand what is NLP and what is Text Mining?

NLP or Natural Language Processing refers to a branch of AI that deals with human languages. It gives to machines the ability to read, understand an derive meaning from human languages.

When it comes to Text Mining it can be defined as the process of deriving meaningful information from language text.

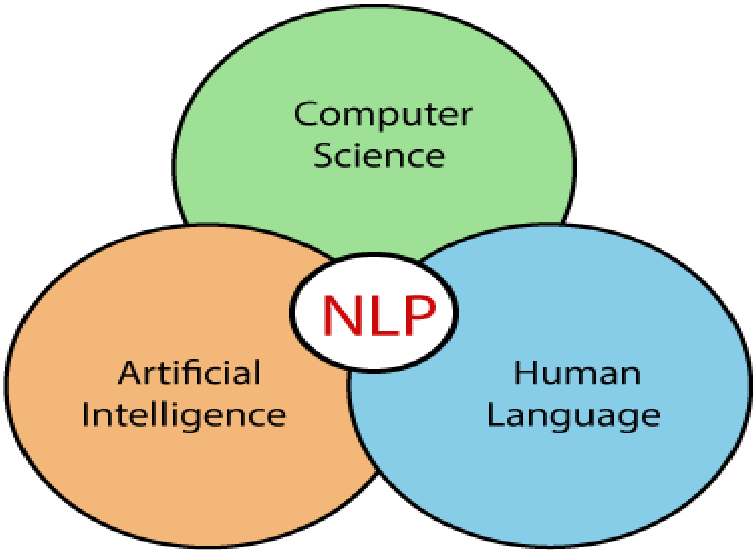


***Figure 1**. NLP scheme*

NLP is a part of Computer Science and Artificial Intelligence which deals with human languages.

Nowadays NLP is being used in a various number of applications.

Some of the most common applications of NLP include: Spell checking, Keyword search, Information exctraction in websites and documents, Advertisement matching (recomandation of ads based on your search), Sentimental Analysis, Speech recognition (voice assistants ex. siri), Chatbot, Machine translation (ex. google translate)

1.1 Thesis Objective

In this thesis, you will be introduced to a very powerful text mining and data analysis platform called Orange. We are going to use this platform to perform a detailed data analysis for a twitter dataset that I am going to choose and treat like a case study. The main focus will be on sentiment analysis of a big number of comments generated from twitter. So, we will be focused on the application of NLP on sentiment analysis but differing from the authors that are going to be mentioned in the Literature Review our experiment is going to be conducted using Orange Platform.

In this thesis I have decided to make a case study about one of the most famous brands that has been under a large number of critics by its clients and public because of a scandal that happened in one of their last campaigns. This luxury brand is called Balenciaga.

A full Data Analytics of the comments found online on Twitter platform was performed about this brand.

All these Data Analytics are shown using Orange. In the methodology and results section are shown every step that I have performed. All widgets used are explained including their function and their results.

There will also be a results chapter which will show to the readers the results of the work performed. Conclusions chapter will be the end of the thesis.

1.2 Motivation (Scope of the Work)

The scope of the work is to conduct some data analytics for a twitter dataset that I am going to use. The main analysis will be about detecting the sentiment of the tweets that I have in my dataset. This kind of analysis can be very useful for the brand to analyse the opinion that people have about their products and can be used by them to improve the brand performance.

1.3 Thesis Structure

The thesis will be organized in 5 chapters. You already had a look on the first chapter which was an overview of what you are going to read in the following chapter and what I want to achieve in the end of the thesis. The second chapter is about literature review, in which are gathered information around our topic from different authors and journals. The third chapter shows the methodology of work and all methods that I have used in this thesis. The fourth chapter is about results and the last chapter brings a conclusion to my thesis.

CHAPTER 2

LITERATURE REVIEW

In this chapter we are going to understand some of the basics of NLP. There will be shown the two main components of Natural Language Processing. Their main functions will be explained and what are they used for. After explaining the two main components there will be shown the main elements that are always taken in consideration when processing different kind of texts. Also, different kind of techniques used by different authors are going to be explained

At these first steps I have tried to understand and present to the readers some of the simplest techniques and applications of NLP and NLU as a main component of NLP.

Driven by the increasing demand for advanced technologies in industries such as finance, healthcare, and e-commerce. The ability to analyze and understand human language has numerous potential applications, ranging from sentiment analysis in social media to automated customer service and natural language generation.

In recent years, there have been significant advancements in NLP, particularly in the areas of deep learning and representation learning. This has led to the development of models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer), which have achieved state-of-the-art results in a wide range of NLP tasks, including named entity recognition, question answering, and sentiment analysis.

Despite these advancements, NLP is still a challenging field, and there remain many open research questions related to understanding the meaning and context in natural language. However, with continued advancements in NLP and semantic analysis, it is likely that the field will continue to grow and have a significant impact on a wide range of industries in the future.

2.1 Components of NLP

When we talk about NLP, we should have in mind two main components:

1.Natural Language Understanding(NLU)**,** which is the part that takes care of mapping input to useful representations and also analyzing different aspects of language.

2.Natural Language Generation which is the process of producing meaningful phrases and sentences in the form of natural language from some internal representation. It includes: Text planning, Sentence planning, Text realization

Before we continue with our thesis first, I have to show how does NLP work. Everything can be seen in this form: At first you start with a document which is going to be analyzed. When you say a document, it means that there can also be analyzed different kind of datasets and a various number of text forms.

To make the algorithm understand what is going on, you need to process it into a form which is easy comprehensible by the machine. (like making a child to read). These are the steps or elements needed to be taken in consideration when processing a text:

- Tokenization(break a complex sentence into words. Understand the importance of each with respect to the sentence. Produce a structural description on an input sentence.)
- Stemming (Normalizing the word into its root form)
- Lemmatization (groups together different inflected forms of a word, called Lemma. Somehow similar to Stemming, as it maps several words into a common root. Output of Lemmatization is a proper word. Ex: a Lemmatiser should map gone, going into go)
- Stop words (not helpful in language processing. A list of words ex: not, and, just, do etc.)
- POS(parts of speech) tags (noun, verb, adverb, singular, plural etc.)
- Named Entity Recognition (move , monetary value, organization, location, quantities, person)
- Syntax (Syntax Tree)
- Chunking(picking up individual pieces of information and grouping them into bigger pieces)

2.2 NLP and Semantic Analysis

Through many years a vast number of studies and papers have been written by different authors which experiment on different applications of NLP. Surveys have been abducted. My main focus in this thesis was on representing different kind of techniques that other authors have used for experimenting on sentiment analysis using NLP.

A literature review for Natural Language Processing (NLP) and semantic analysis is an overview of existing research on the use of computational methods for understanding, interpreting, and generating human language. NLP has been a rapidly growing field in recent years, driven by advances in machine learning, artificial intelligence, and computational linguistics.

Semantic analysis is a crucial part of NLP, as it involves understanding the meaning behind words and phrases in a text. Techniques used in semantic analysis include word sense disambiguation, named entity recognition, part-of-speech tagging, and sentiment analysis.

Some of the key challenges in NLP and semantic analysis include dealing with ambiguity, context-dependency, and the vast variability in language use. These challenges have led to the development of various NLP tools and models, including rule-based systems, machine learning algorithms, and deep learning models such as recurrent neural networks (RNNs) and transformers.

One notable application of NLP and semantic analysis is in the field of information retrieval, where techniques are used to extract relevant information from large collections of text. Other applications include machine translation, text classification, and chatbots.

Recent advancements in NLP and semantic analysis have shown significant improvements in the accuracy and speed of NLP models, particularly with the use of deep learning techniques. However, there is still much room for improvement, and ongoing research continues to explore new methods for making NLP models more robust, scalable, and effective for real-world applications.

2.3 Methods used for Sentiment Analysis

2.3.1 Rule-based Approach

There are several methods used for sentiment analysis. The first one is Rule-based approaches which involves using a set of manually crafted rules to classify the sentiment in a text. Rule-based approaches are a traditional method for hate speech detection in social media. In this approach, a set of predefined rules and patterns are used to identify hateful content in the text. For example, a rule-based approach might flag words or phrases that are commonly used in hate speech, such as racial slurs or insults.

This method can be relatively simple to implement and can achieve decent results for simple cases. However, rule-based approaches can be limited in their accuracy and scalability, as they can miss more complex cases of hate speech that do not match the predefined rules, and they can be difficult to update as new forms of hate speech emerge.

Additionally, rule-based approaches can also suffer from false positives, where non-hateful content is incorrectly flagged as hate speech. For this reason, rule-based approaches are often used in conjunction with other methods, such as machine learning algorithms, to improve the accuracy of hate speech detection.

2.3.2 Lexicon-based Approach

Another method is Lexicon-based approach which involves using a pre-existing dictionary or lexicon of words and their sentiment polarity to classify the sentiment in a text. Lexicon-based approaches are a method for hate speech detection in social media that involves using a pre-compiled list of words and phrases associated with hate speech. The text is analyzed and compared to this lexicon, and any matches are flagged as potential instances of hate speech.

This method can be more accurate than rule-based approaches, as the lexicon can include a wide range of words and phrases used in hate speech. However, lexicon-based approaches can still suffer from false positives, where non-hateful content is incorrectly flagged as hate speech. This can occur because the lexicon may include words that have multiple meanings and can be used in different contexts.

Additionally, lexicon-based approaches may also miss more subtle forms of hate speech that are not included in the lexicon. To address this issue, lexicon-based approaches can be combined with other methods, such as machine learning algorithms, to improve the accuracy of hate speech detection. It's important to note that the quality and coverage of the lexicon used can have a significant impact on the performance of the hate speech detection system, and it's crucial to continually update and evaluate the lexicon to ensure that it remains relevant and effective.

2.3.3 Machine learning-based Approach

Machine learning-based approaches is a method that involves training a model on a labeled dataset to learn the relationship between the text and its sentiment and then using this model to classify the sentiment in new texts. Machine learning-based approaches are a method for hate speech detection in social media that involves training a model to identify hateful content in text. This is done by feeding the model a large amount of annotated training data, which consists of text samples labeled as hate speech or non-hate speech. The model then uses this training data to learn the patterns and features associated with hate speech.

Machine learning-based approaches can be more accurate and scalable than rule-based and lexicon-based approaches, as they can identify more complex and nuanced forms of hate speech that are difficult to detect with other methods. Popular machine learning techniques used for hate speech detection include text classification, sentiment analysis, and representation learning.

One advantage of machine learning-based approaches is that they can adapt to new forms of hate speech over time, as long as the training data is updated to reflect these changes. However, these models can also suffer from biases, if the training data is biased or if the model has been trained on a limited set of examples.

It's crucial to evaluate the performance of machine learning-based hate speech detection models to ensure that they are both effective and ethical, and to correct any biases that might be present. Additionally, it's important to balance the need for accuracy with the need for privacy and freedom of expression, to ensure that hateful speech is detected without infringing on the rights of individuals to express their opinions.

2.3.4 Deep learning-based Approach

Also, Deep learning-based approaches is another interesting method that involves using deep neural networks, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to analyze the sentiment in a text. Deep learning-based approaches are a method for hate speech detection in social media that involve using neural networks, a type of machine learning algorithm, to identify hateful content in text. These models are called "deep" because they consist of many interconnected layers that process the input data and learn complex representations of the data.

Deep learning-based approaches have shown promising results for hate speech detection, as they can capture subtle patterns and relationships in the text that other methods may miss. For example, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been used to analyze the sequence of words in a sentence and identify features relevant to hate speech.

One advantage of deep learning-based approaches is that they can learn directly from the raw text data, without the need for manual feature engineering or lexicon-based methods. However, they can also be computationally expensive and require a large amount of labeled training data to achieve good results.

It's important to evaluate the performance of deep learning-based hate speech detection models to ensure that they are both accurate and ethical, and to correct any biases that might be present in the training data or the model itself. Additionally, it's crucial to consider the trade-off between accuracy and privacy, to ensure that hateful speech is detected without infringing on the rights of individuals to express their opinions.

Different techniques are suitable for different types of data, and some techniques may be better suited for certain tasks than others.

For example, rule-based methods are well-suited for simple tasks where the patterns in the data are well-defined, but they can be less effective for more complex tasks where the patterns are less well-defined. Machine learning-based methods can be more effective for complex tasks, but they require annotated training data, which can be time-consuming and expensive to collect. Deep learning-based methods can be very effective for complex tasks, but they require large amounts of data and computational resources.

Additionally, the choice of technique will depend on the desired level of accuracy, the timeline for the project, and the available resources, such as computational resources and expertise. The researcher will need to carefully consider these factors and choose the technique that best meets their needs and constraints.

Each of these methods has its own advantages and disadvantages and the choice of method depends on the specific requirements of the task and the available resources.

2.4 Advantages and Disadvantages in methods used for Sentiment Analysis

2.4.1 Advantages and Disadvantages of rule-based approaches

Advantages of rule-based approaches for sentiment analysis include the Simplicity that Rule-based approaches are easy to understand and implement, making them a good starting point for beginners in sentiment analysis. Another advantage is the Transparency. Rule-based approaches are transparent in their decision-making process, as the rules used to classify the sentiment in a text are clearly defined and can be easily modified. Also, Customizability is another advantage of Rule-based approaches. They can be customized to suit the specific needs of a task by adding or modifying the rules used for classification.

Disadvantages of rule-based approaches for sentiment analysis include the Limited coverage that Rule-based approaches can only classify sentiment for texts that follow the patterns defined by the rules, and may not be able to handle more complex or idiomatic expressions. Brittle performance is another disadvantage for Rule-based approaches. They can be easily thrown off by even small variations in the text, and may not be robust to different writing styles or languages. Also, Lack of generalization is a disadvantage for Rule-based approaches because they are not capable of generalizing to new, unseen texts, as they are only able to classify sentiment based on the specific patterns defined by the rules. In addition to all these, Maintenance is another disadvantage for Rule-based approaches because they require manual maintenance and updating of the rules to keep up with changing language usage and sentiment expressions.

2.4.2 Advantages and Disadvantages of lexicon-based approaches

Advantages of lexicon-based approaches for sentiment analysis include the Ease of use because Lexicon-based approaches are simple to implement and do not require extensive training data. Speed is another advantage for Lexicon-based approaches because they can process large amounts of text quickly, making them suitable for real-time sentiment analysis. Also, High coverage can be mentioned as an advantage for Lexicon-based approaches because they can handle a wide range of sentiments, including positive, negative, and neutral, by leveraging a large pre-existing sentiment lexicon.

Disadvantages of lexicon-based approaches for sentiment analysis include the Limited accuracy that Lexicon-based approaches have because they may not always accurately reflect the sentiment in a text, as the sentiment of a word may change based on the context in which it is used. Another disadvantage is the Lack of context that Lexicon-based approaches have. They do not take into account the context in which a word is used, and may not accurately reflect the sentiment in a text as a result. Dependence on lexicon qualityis another disadvantage. The quality of the sentiment lexicon used can greatly impact the accuracy of the sentiment analysis, and poor-quality lexicons may lead to incorrect classifications. In the end, Lack of customization can also be mentioned as a disadvantage for Lexicon-based approaches. They are limited by the coverage of the sentiment lexicon used and may not be suitable for tasks with specific requirements, such as analyzing sentiment in a specific domain or language.

2.4.3. Advantages and Disadvantages of Machine Learning-based approaches

Advantages of Machine Learning-based approaches include Automation because Machine learning algorithms can automate repetitive tasks and make decisions without human intervention. Scalability is another advantage that Machine learning algorithms have because can handle large amounts of data and can be easily scaled to accommodate future growth. Also, Improved accuracy is an advantage of Machine learning algorithms because they can learn from data and identify patterns that might not be immediately apparent to humans, leading to improved accuracy and decision making. Another advantage is Time efficiency. Machine learning algorithms can process large amounts of data in a short amount of time, allowing organizations to quickly make informed decisions.

Disadvantages of Machine Learning-based approaches include Bias. Machine learning algorithms can perpetuate existing biases in the data used for training, leading to unfair or inaccurate outcomes. Another disadvantage is Overfitting because Machine learning algorithms can sometimes over-learn from the training data and not generalize well to new, unseen data. Lack of transparency is another disadvantage for Machine learning algorithms which can be difficult to understand and interpret, making it hard to explain their decisions and reasoning. Another aspect that can be seen as a disadvantage is that they require large amounts of data. Machine learning algorithms require large amounts of data to train effectively, which can be a challenge for organizations with limited data resources. Dependency on quality of datais another disadvantage. The accuracy of machine learning algorithms depends on the quality of the data used for training, and bad data can lead to poor results.

2.4.4 Advantages and Disadvantages of Deep learning-based approaches

Advantages of Deep learning based approaches include High accuracy. Deep learning models can achieve very high accuracy levels in various tasks, such as image and speech recognition, natural language processing, etc. Handling complex patterns is another advantage for Deep learning models because they are capable of handling complex patterns and relationships within the data. Automated feature extraction is also an advantage for Deep learning models because they can automatically extract features from the input data, reducing the need for manual feature engineering. Another thing that can be mentioned as an advantage is the large-scale deployment. Deep learning models can be deployed at scale, making it possible to handle large amounts of data and make predictions in real-time.

Disadvantages of Deep learning based approaches include Computational cost. Deep learning models require a lot of computational resources, which can make them difficult to train and deploy on low-end hardware. Another aspect that can be mentioned as a disadvantage is Data requirements because Deep learning models require large amounts of labeled data to achieve good results, which can be difficult and time-consuming to obtain. Overfitting is another disadvantage for Deep learning models because they are prone to overfitting, particularly when trained on small amounts of data, which can lead to reduced performance on unseen data. Also, Interpretability is another disadvantage for Deep learning models because they can be difficult to interpret, making it challenging to understand how they are making predictions and identify potential biases in the model.

2.5 Sentiment Analysis in social media

Sentiment Analysis in Social Media refers to the process of using Natural Language Processing (NLP) and Machine Learning techniques to classify the sentiment expressed in texts from social media platforms, such as Twitter, Facebook, and Instagram, into positive, negative, or neutral. The goal of sentiment analysis is to determine the overall attitude of the author of the post toward a particular topic, product, or event.

Advantages of sentiment analysis in social media include:

1. Brand monitoring: Companies can use sentiment analysis to monitor and analyze public perception of their brand, products, or services.
2. Customer feedback: Sentiment analysis can be used to gather valuable customer feedback, helping organizations to make data-driven decisions.
3. Market research: Sentiment analysis can be used to gather insights into public opinions and preferences in real-time, providing valuable market research information.

However, there are also some challenges and limitations associated with sentiment analysis in social media. Ambiguous language: Social media texts often contain informal and ambiguous language, which can make it difficult for sentiment analysis algorithms to accurately classify the sentiment. Irony and sarcasm: Social media users often use irony and sarcasm, which can lead to incorrect sentiment classification by the algorithms. Bias and fairness: Sentiment analysis algorithms can be biased if the training data contains biased information, leading to inaccurate and unfair predictions.

Overall, sentiment analysis in social media is a valuable tool for organizations, but it should be used with caution, considering its limitations and potential biases.

2.6 Text Analytics using NLP

Text Analytics using NLP (Natural Language Processing) refers to the process of analyzing and understanding text data to extract meaningful insights and information. This can include a range of tasks, such as sentiment analysis, named entity recognition, topic modeling, summarization, and more.

Text analytics using NLP techniques can be applied in various domains. These domains include social media analysis: Analyzing public opinions, sentiments, and emotions expressed on social media platforms. It includes Customer feedback analysis: Analyzing customer feedback in surveys, reviews, and other forms of customer feedback to gain insights into customer satisfaction and opinions. Market research is also a part of this domain: Analyzing news articles, press releases, and other forms of public discourse to understand public opinions and preferences. Cybersecurity: Analyzing large volumes of text data from security logs to detect and prevent cyber attacks. Healthcare: Analyzing electronic health records to extract important information, such as diagnosis, treatment plans, and patient outcomes.

Text analytics using NLP is a rapidly growing field that has the potential to revolutionize various industries by providing valuable insights and information from large amounts of text data.

Text Analytics using NLP (Natural Language Processing) is also important for several reasons. Reasons such as understanding customer opinions: NLP can be used to analyze large volumes of customer feedback and identify patterns and trends in customer opinions, which can help organizations to improve their products and services. For Market research: NLP can be used to analyze social media posts and news articles to gain insights into consumer preferences, market trends, and competitor activity. Sentiment analysis: NLP can be used to automatically categorize the emotions and opinions expressed in text, which can be useful for customer feedback analysis, brand monitoring, and market research. Also Text classification: NLP can be used to automatically categorize and classify text into different categories, such as spam filtering, document categorization, and topic modeling. Furthermore Text summarization: NLP can be used to automatically summarize long pieces of text, making it easier to consume large amounts of information quickly. Lastly Natural language generation: NLP can be used to generate new text that is coherent and semantically meaningful, which can be useful for applications such as customer service chatbots and content generation.

In conclusion, NLP plays a critical role in text analytics, enabling organizations to make informed decisions based on large volumes of unstructured text data.

2.7 Hate speech detection in social media using NLP

Hate speech detection in social media can be performed using NLP (Natural Language Processing) techniques. It involves using various algorithms and models to identify and classify hateful content in text, such as racist or sexist slurs, insults, and other forms of hate speech.

This task can be challenging as hate speech often involves the use of slang, sarcasm, and other linguistic nuances that can be difficult to detect. However, techniques such as sentiment analysis, text classification, and representation learning can be used to build models that can accurately identify hate speech in social media.

It's worth mentioning that hate speech detection is an ongoing area of research, and current methods may have limitations and biases. Thus, it's important to continually evaluate and improve these models to ensure that they are ethical and effective in detecting hate speech.

CHAPTER 3

METHODOLOGY

3.1 Orange Platform

Orange is a powerful open-source machine learning and data visualization platform which allows its users to perform data analysis, see data flow and also become much more productive. It was first released in 1996 by the University of Ljubljana but once it launched widgets in 2010 it started being more and more useful.

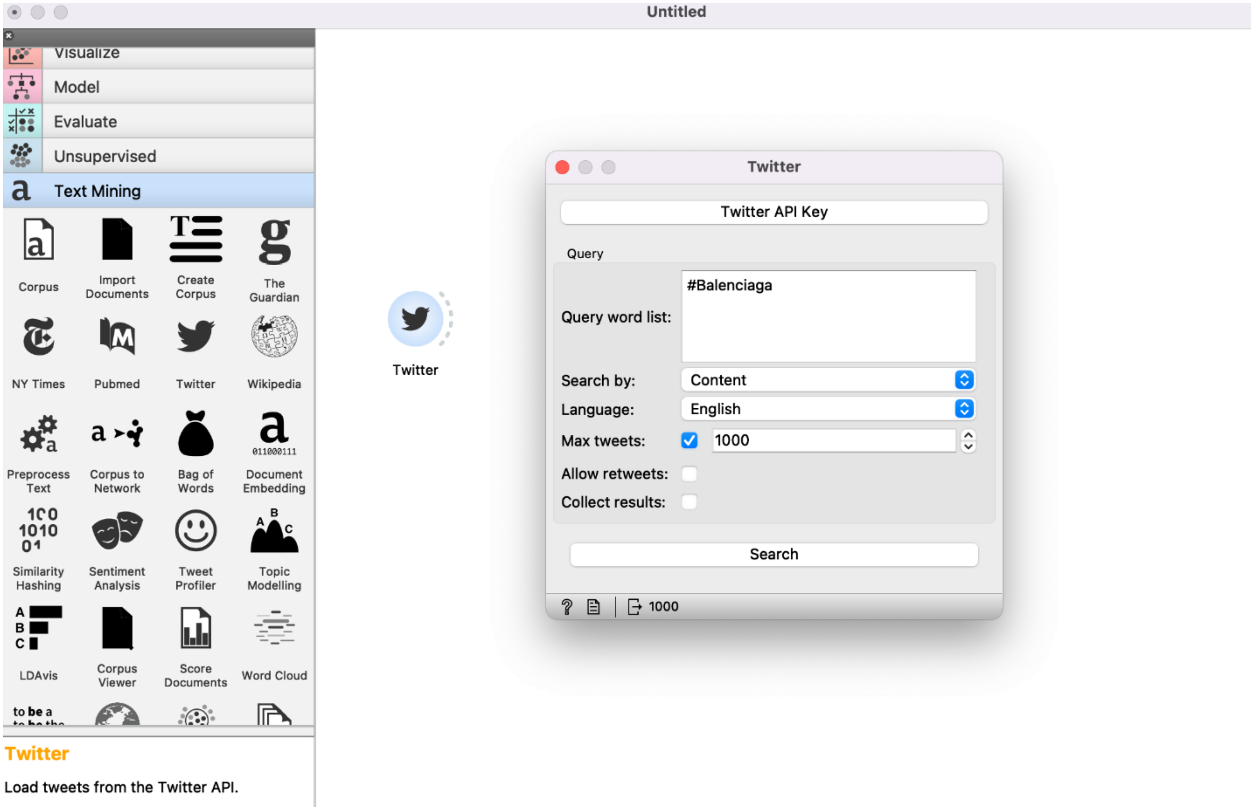
It is used nowadays in biomedicine, genomic research, bioinformatics because it provides a platform for recommendation systems, experiment selection and predictive modelling. Orange is also supported on Windows, macOS and Linux and can be easily installed from the Python Packages.

Orange data mining works with widgets. These widgets are computational units of Orange. Widgets read, process and visualize the data. You can also do clustering by using them or build predictive models. In other words, widgets help us to explore the data.

Orange Canvas is a visual programming environment for Orange. In there all the widgets are integrated and all the relationships between widgets take form.

3.2 Dataset

As we said in the introduction in this thesis it is presented a sentiment analysis for a Twitter Dataset. My dataset is composed of 1000 tweets which have “Balenciaga” brand as their keyword. Twitter widget in Orange is going to be very helpful for us in obtaining the datasets with the exact number of comments or mentions that we want to analyze and the language that we want these comments to be in.



Twitter

Load tweets from the Twitter API.

Query word list: #Balenciaga

Search by: Content

Language: English

Max tweets: ☑ 1000

Allow retweets: □

Collect results: □

Figure 2. *Twitter widget*

In the figure above there are shown the dataset characteristics that I wanted my twitter dataset to have. All text will be in English language in form of comments. There will be a total of 1000 tweets and they will have “Balenciaga as a keyword”.

3.3 Tweets’ content

To see the content of each tweet one by one we have to use another widget that Orange provides us. This widget is called the Data Table widget and it helps us see the content of the tweets associated with their respective author. You can also see the exact time that it was tweeted, the language of the tweet, number of likes that it has and a few other tweets characteristics and information that we may take in consideration when making different kind of data analysis.

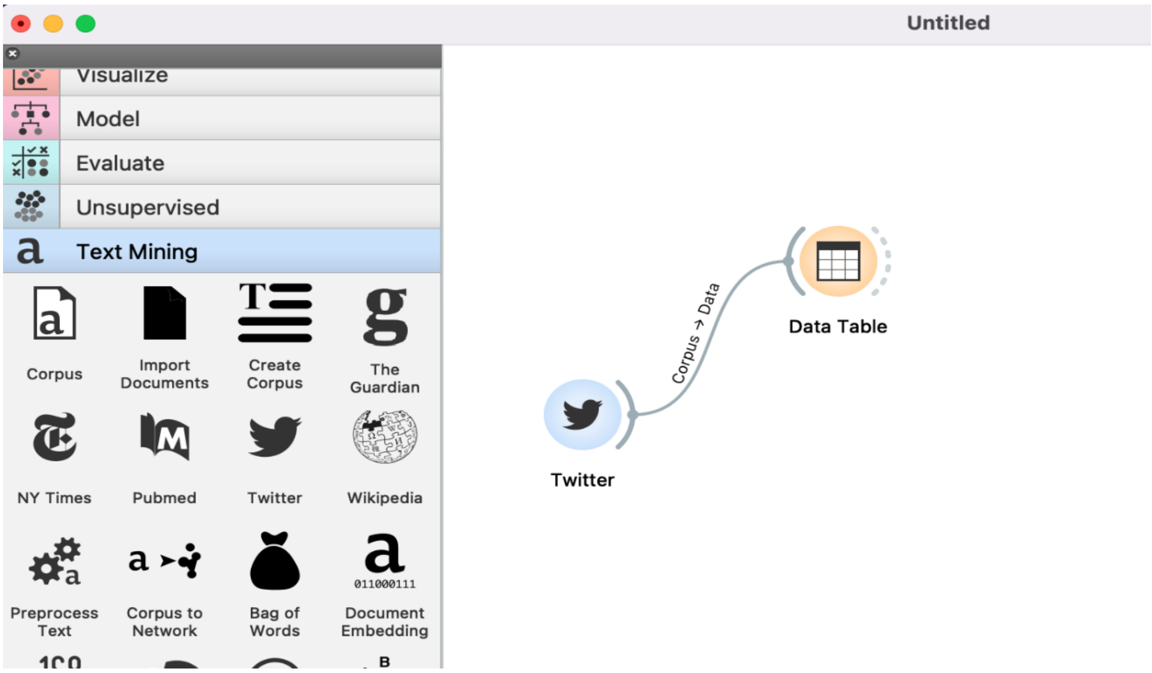
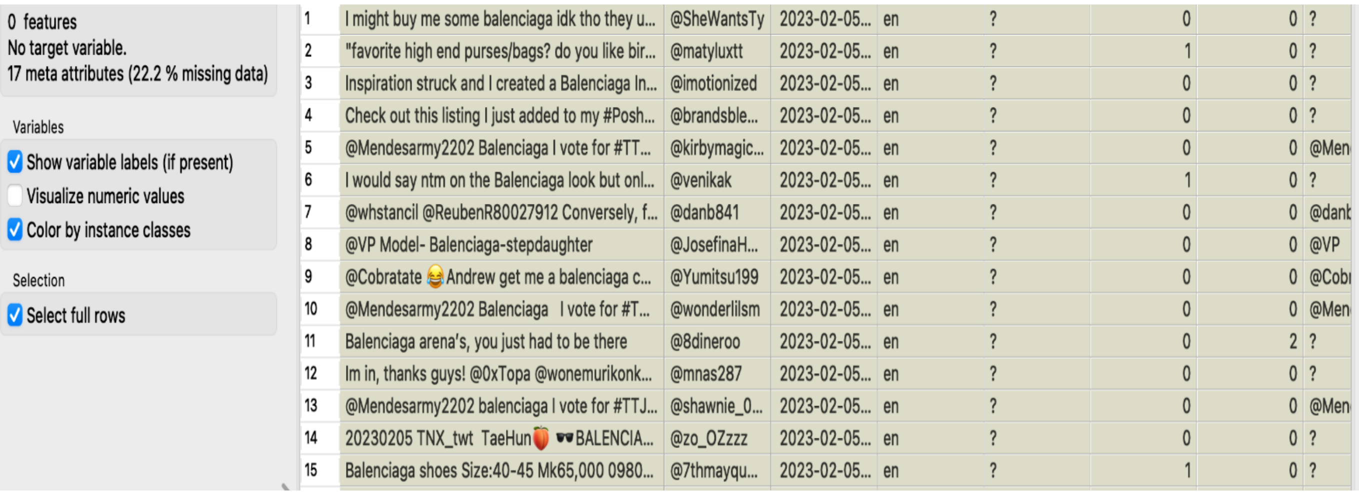


Figure 3. *Data Table widget*

In the figure above it is shown the next step that connects twitter widget with the Data Table widget.

Table 1. *Data Table*



0 features
No target variable.
17 meta attributes (22.2 % missing data)

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

1	I might buy me some balenciaga idk tho they u...	@SheWantsTy	2023-02-05...	en	?	0	0	?
2	"favorite high end purses/bags? do you like bir...	@matyluxtt	2023-02-05...	en	?	1	0	?
3	Inspiration struck and I created a Balenciaga In...	@imotionized	2023-02-05...	en	?	0	0	?
4	Check out this listing I just added to my #Posh...	@brandsble...	2023-02-05...	en	?	0	0	?
5	@Mendesarmy2202 Balenciaga I vote for #TT...	@kirbymagic...	2023-02-05...	en	?	0	0	@Men
6	I would say ntm on the Balenciaga look but onl...	@venikak	2023-02-05...	en	?	1	0	?
7	@whstancil @ReubenR80027912 Conversely, f...	@danb841	2023-02-05...	en	?	0	0	@danb
8	@VP Model- Balenciaga-stepdaughter	@JosefinaH...	2023-02-05...	en	?	0	0	@VP
9	@Cobratate 😂Andrew get me a balenciaga c...	@Yumitsu199	2023-02-05...	en	?	0	0	@Cob
10	@Mendesarmy2202 Balenciaga I vote for #T...	@wonderlilsm	2023-02-05...	en	?	0	0	@Men
11	Balenciaga arena's, you just had to be there	@8dineroo	2023-02-05...	en	?	0	2	?
12	Im in, thanks guys! @0xTopa @wonemurikonk...	@mnas287	2023-02-05...	en	?	0	0	?
13	@Mendesarmy2202 balenciaga I vote for #TTJ...	@shawnie_0...	2023-02-05...	en	?	0	0	@Men
14	20230205 TNX_twt TaeHun🍑 🕶BALENCIA...	@zo_OZzzz	2023-02-05...	en	?	0	0	?
15	Balenciaga shoes Size:40-45 Mk65,000 0980...	@7thmayqu...	2023-02-05...	en	?	1	0	?

In this table there is a preview of the results obtained in the Data table widget. There is shown the content of the tweets associated with authors, exact time it was tweeted and some other characteristics of tweets.

Another widget that displays text and enables us to browse it is the Corpus Viewer.

The two figures below show respectively the Corpus Viewer Widget and the result obtained by using it.

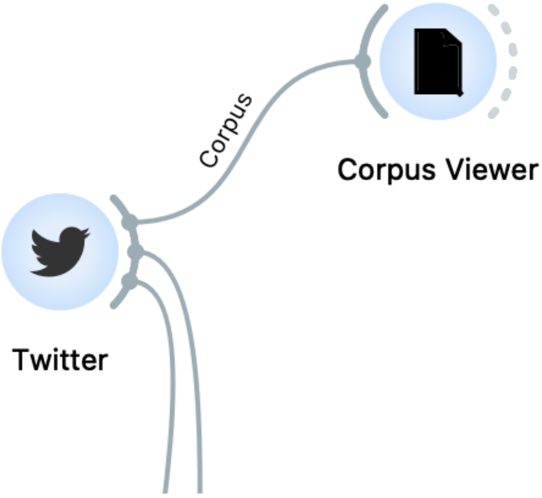
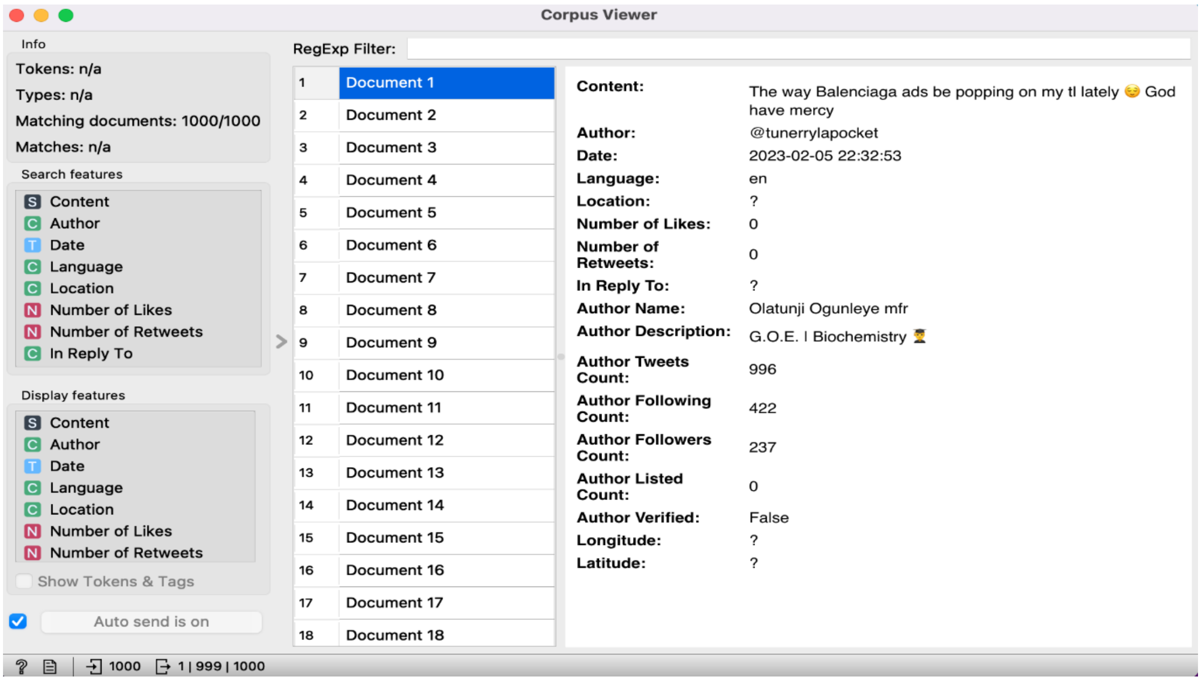


Figure 4. *Corpus Viewer Widget*



Info

Tokens: n/a

Types: n/a

Matching documents: 1000/1000

Matches: n/a

Search features

- S Content
- C Author
- T Date
- C Language
- C Location
- N Number of Likes
- N Number of Retweets
- C In Reply To

Display features

- S Content
- C Author
- T Date
- C Language
- C Location
- N Number of Likes
- N Number of Retweets

□ Show Tokens & Tags

☑ Auto send is on

1000 | 1 | 999 | 1000

RegExp Filter:

1	Document 1
2	Document 2
3	Document 3
4	Document 4
5	Document 5
6	Document 6
7	Document 7
8	Document 8
9	Document 9
10	Document 10
11	Document 11
12	Document 12
13	Document 13
14	Document 14
15	Document 15
16	Document 16
17	Document 17
18	Document 18

Content:	The way Balenciaga ads be popping on my tl lately 😌 God have mercy
Author:	@tunerrylapocket
Date:	2023-02-05 22:32:53
Language:	en
Location:	?
Number of Likes:	0
Number of Retweets:	0
In Reply To:	?
Author Name:	Olatunji Ogunleye mfr
Author Description:	G.O.E. l Biochemistry 👨‍🎓
Author Tweets Count:	996
Author Following Count:	422
Author Followers Count:	237
Author Listed Count:	0
Author Verified:	False
Longitude:	?
Latitude:	?

Figure 5. *Corpus Viewer*

3.4 Key words

In all that huge content some of the readers may be interested about the main keywords that all 1000 tweets have in common. By the help of a widget called Word Cloud which is another widget that helps for visualizing the text you can see the word frequencies of all our tweets in the form of a cloud where the biggest words in cloud are the words which are more frequently mentioned in the tweets.

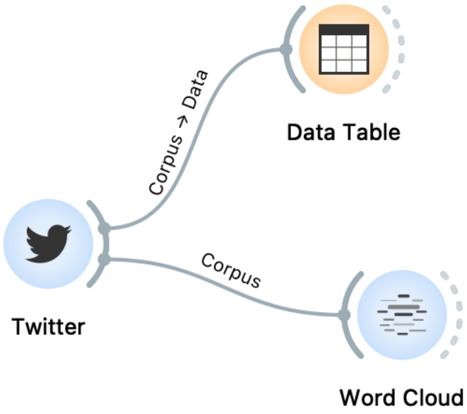


Figure 6. *Word Cloud Widget*

In this figure it is shown the connection of twitter widget with the world cloud widget.

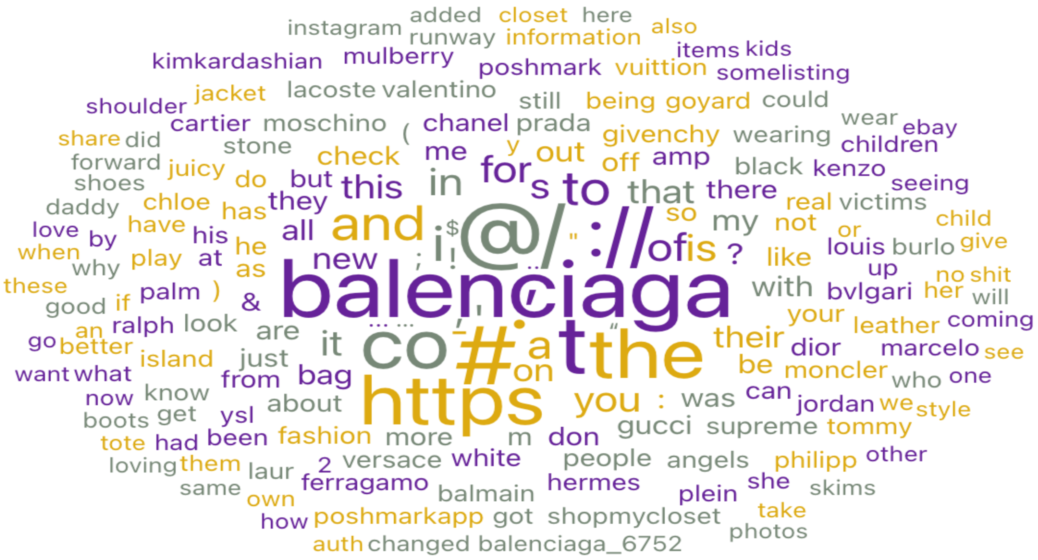


Figure 7. *Word cloud before Preprocessing*

Here it is shown the world cloud obtained from these tweet contents. We can clearly see that the words written in bigger letters are the words that are repeated the most. But we can also notice that there are a lot of "garbage" considered as key words such as : @, https, //, # etc. and we should find a way to get rid of those.

3.5 Preprocessing Text

The Preprocess Text widget performs some of the basic elements of NLP including: splitting of text into smaller units or tokens, stemming, lemmatization, creating n-grams and tags tokens with part-of-speech labels.

When we use preprocessing before the word cloud widget we will obtain better results in the cloud. The word cloud visualization will look much better. The punctuations will be removed, only meaningful words will be retained and also we can remove words that we want to from the cloud.

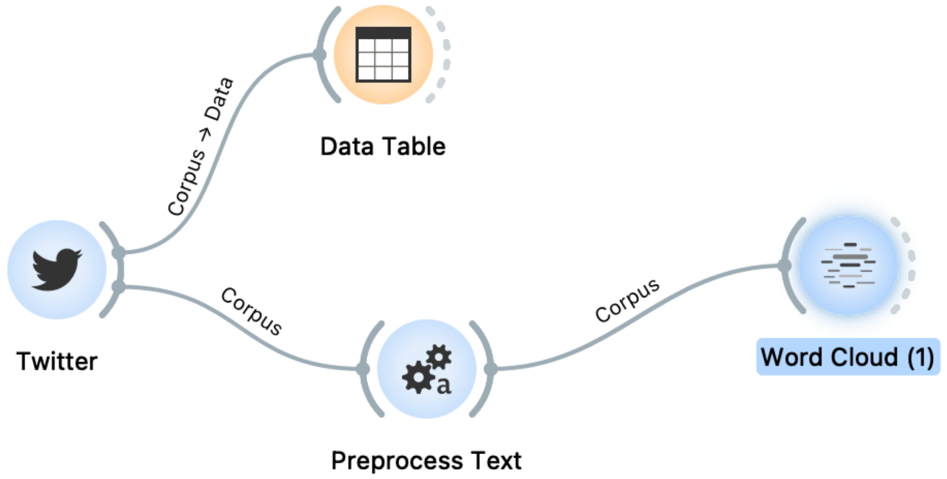


Figure 8. *Preprocess Text Widget*

This figure shows the addition of the preprocess text widget in the canvas right before the word cloud widget.

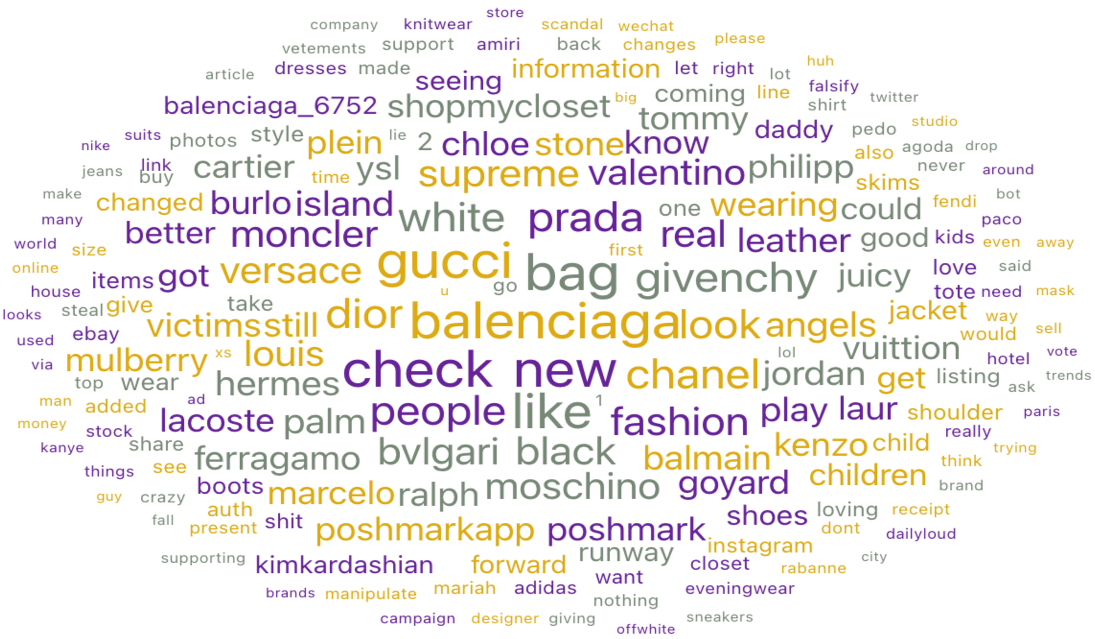


Figure 9. *Word Cloud after preprocessing*

This figure shows the new word cloud without any punctuations and only meaningful word included. We can clearly see that our key word is the brand name “Balenciaga” and some other key words include other brand names.

3.6 Bag of Words

Bag of Words widget is used to create a corpus with word counts for each tweet. This count can be in three forms: absolute, binary or sub linear.

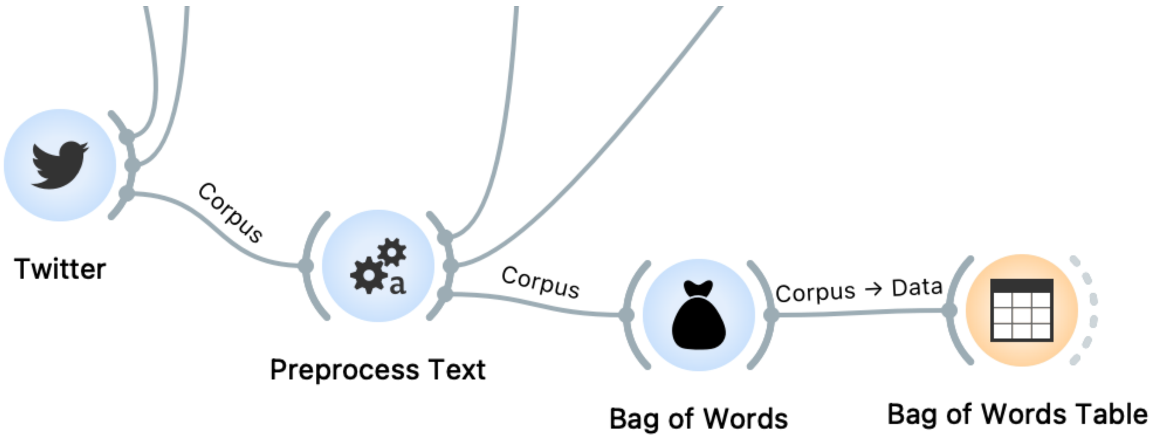
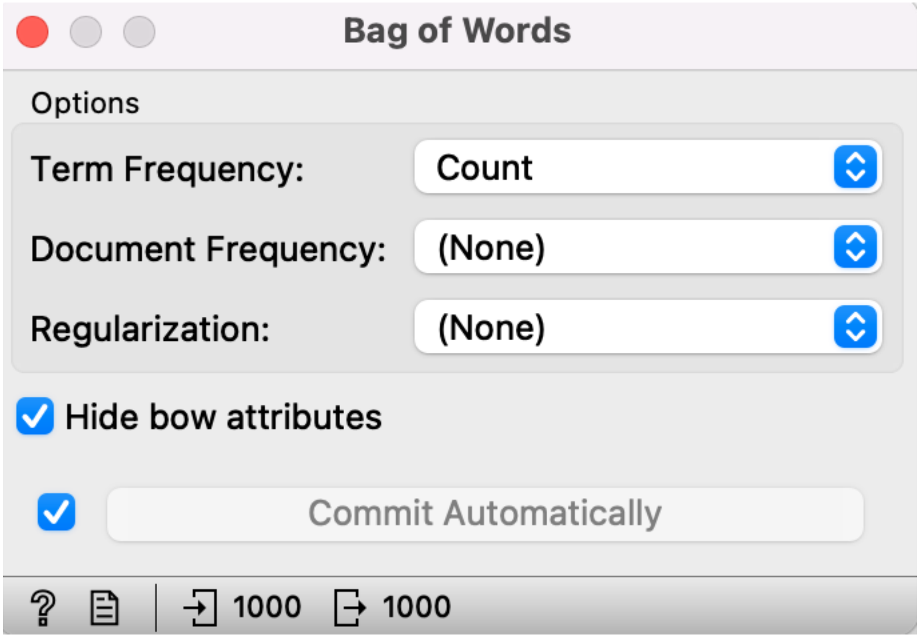


Figure 10. *Bag of Words Widget*

In this figure it is shown the addition of the Bag of Words widget in the canvas associated with a Bag of Words table where we will see the results.



Bag of Words

Options

Term Frequency: Count

Document Frequency: (None)

Regularization: (None)

☑ Hide bow attributes

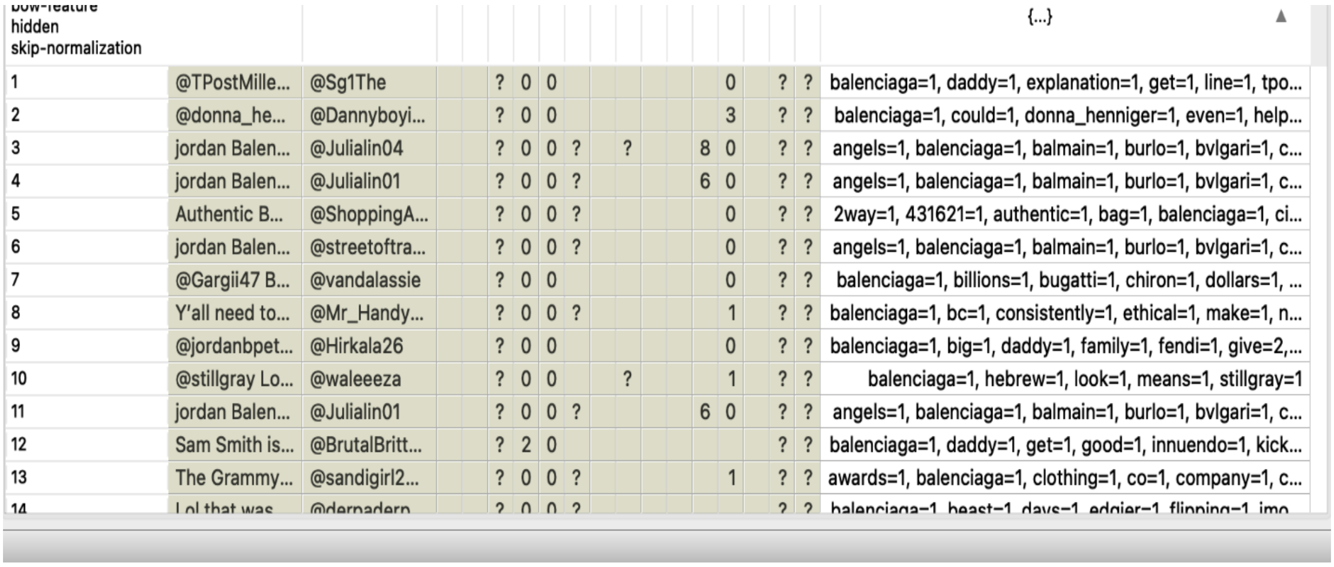
☑ Commit Automatically

1000 1000

Figure 11. *Bag of Words options*

In this figure we can see the parameters for the Bag of Words widget. In the Term frequency option I have chosen "count" which will output to us the number of occurrences of a word in the respective tweet. While for the Document frequency and Regularization options I do not want any parameter as an output.

__Table 2__. Bag of Words Table



bow-feature hidden skip-normalization																		{...}
1	@TPostMille...	@Sg1The			?	0	0							0		?	?	balenciaga=1, daddy=1, explanation=1, get=1, line=1, tpo...
2	@donna_he...	@Dannyboyi...			?	0	0							3		?	?	balenciaga=1, could=1, donna_henniger=1, even=1, help...
3	jordan Balen...	@Julialin04			?	0	0	?		?			8	0		?	?	angels=1, balenciaga=1, balmain=1, burlo=1, bvlgari=1, c...
4	jordan Balen...	@Julialin01			?	0	0	?					6	0		?	?	angels=1, balenciaga=1, balmain=1, burlo=1, bvlgari=1, c...
5	Authentic B...	@ShoppingA...			?	0	0	?						0		?	?	2way=1, 431621=1, authentic=1, bag=1, balenciaga=1, ci...
6	jordan Balen...	@streetoftra...			?	0	0	?						0		?	?	angels=1, balenciaga=1, balmain=1, burlo=1, bvlgari=1, c...
7	@Gargii47 B...	@vandalassie			?	0	0							0		?	?	balenciaga=1, billions=1, bugatti=1, chiron=1, dollars=1, ...
8	Y'all need to...	@Mr_Handy...			?	0	0	?						1		?	?	balenciaga=1, bc=1, consistently=1, ethical=1, make=1, n...
9	@jordanbpet...	@Hirkala26			?	0	0							0		?	?	balenciaga=1, big=1, daddy=1, family=1, fendi=1, give=2,...
10	@stillgray Lo...	@waleeeza			?	0	0			?				1		?	?	balenciaga=1, hebrew=1, look=1, means=1, stillgray=1
11	jordan Balen...	@Julialin01			?	0	0	?					6	0		?	?	angels=1, balenciaga=1, balmain=1, burlo=1, bvlgari=1, c...
12	Sam Smith is...	@BrutalBritt...			?	2	0									?	?	balenciaga=1, daddy=1, get=1, good=1, innuendo=1, kick...
13	The Grammy...	@sandigirl2...			?	0	0	?						1		?	?	awards=1, balenciaga=1, clothing=1, co=1, company=1, c...
14	Lol that was	@dernadern			?	0	0	?								?	?	balenciaga=1, beast=1, days=1, edgier=1, flipping=1, imo

At the last column of this Bag of words table we can the term frequencies for each tweet in our dataset.

3.7 Twitter Profiler

The Twitter Profiler widget takes as an input the dataset of 1000 tweets that we have chosen and outputs a corpus with information on the sentiment of each document. It retrieves the information for each tweet and after sending data to the server, a model computes emotion scores.

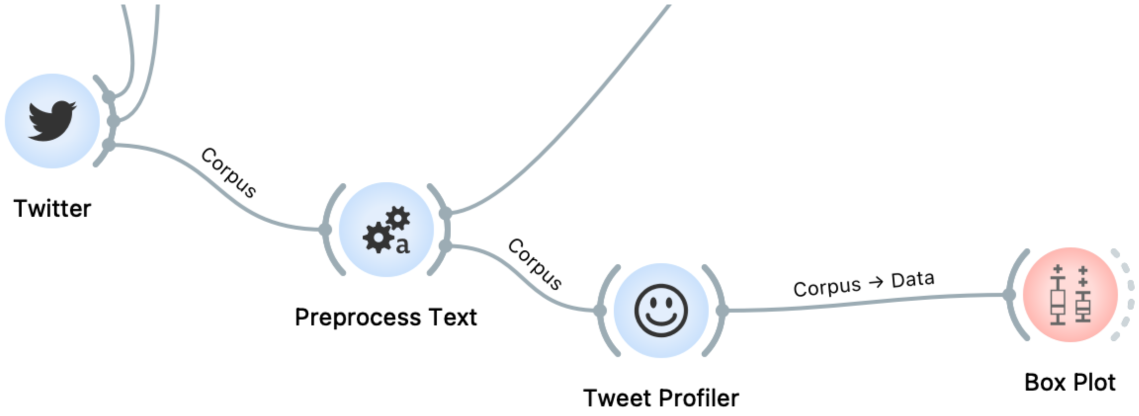
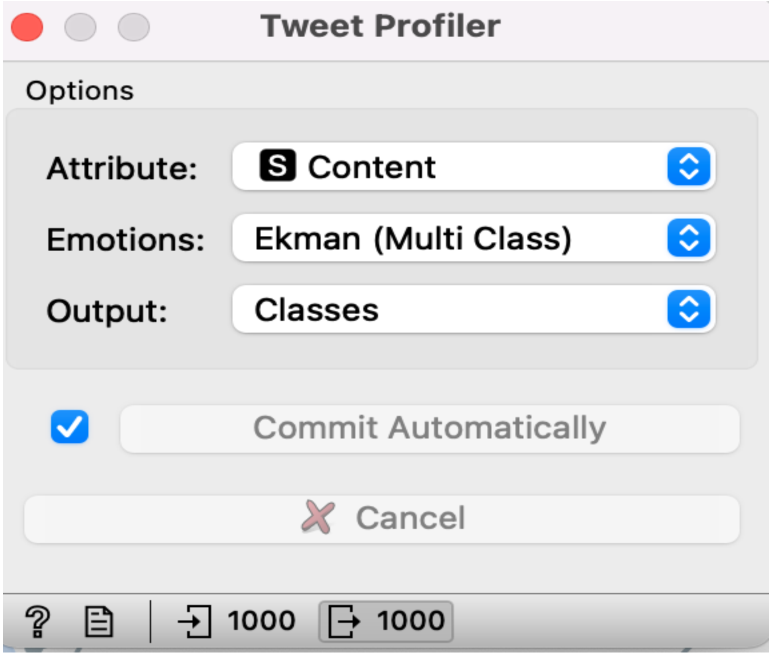


***Figure 12.** Tweeter Profiler and Box Plot Widgets*

This figure shows the addition of the twitter profiler widget in the canvas. After twitter profiler there is a Box Plot widget where we will observe the results depending on the variables and the subgroups in which we are interested in.



Tweet Profiler

Options

Attribute: S Content

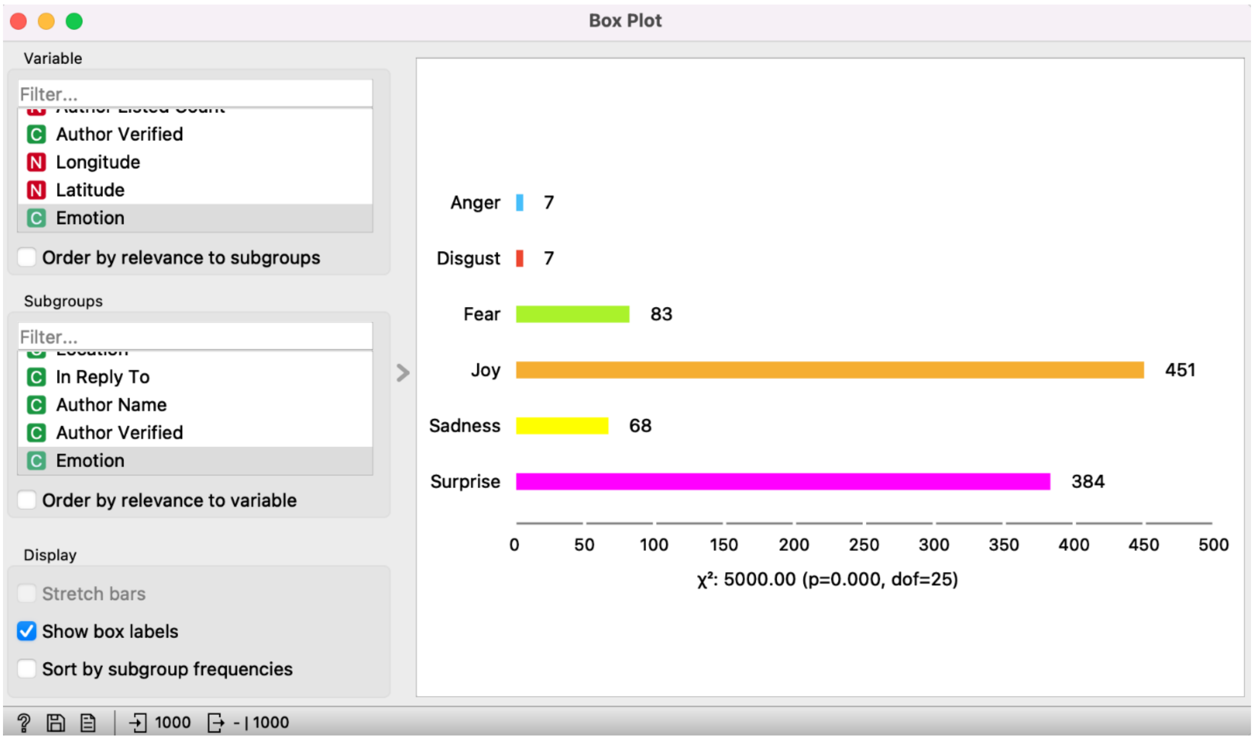
Emotions: Ekman (Multi Class)

Output: Classes

☑ Commit Automatically

Figure 13. *Tweeter Profiler Options*

In this figure are shown the options of Twitter profiler widget. I have decided to use content attribute for the analysis, Ekman’s classification of emotions (with the multi class option) and we will output the results as class.



☐ Order by relevance to subgroups

☐ Order by relevance to variable

☐ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ^2: 5000.00 (p=0.000, dof=25)

Figure 14. *Box Plot*

In this figure are shown the Box Plot results. In this widget I have selected emotion as the variable in which the classification will be based. Also the subgroups will be based by the emotion. This is why we can see the number of tweets that represents each emotion.

3.8 Sentiment Analysis

Sentiment analysis will be the fundamental part of this thesis. We are going to use the Sentiment Analysis widget which predicts sentiment for each tweet in the corpus. The two models that we are going to use for the sentiment analysis: Liu & Hu and Vader sentiment models from the NLTK.

NLTK or Natural Language Toolkit is used for NLP and text analytics. Always we should have it installed in the system when we want to use its library.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 LIU & HU model sentiment analysis

The Liu & Hu sentiment analysis model computes and provide to us a single normalized score of the sentiment in a tweet. We will have a negative score for a negative sentiment, a positive score for a positive sentiment and 0 score for a neutral sentiment.

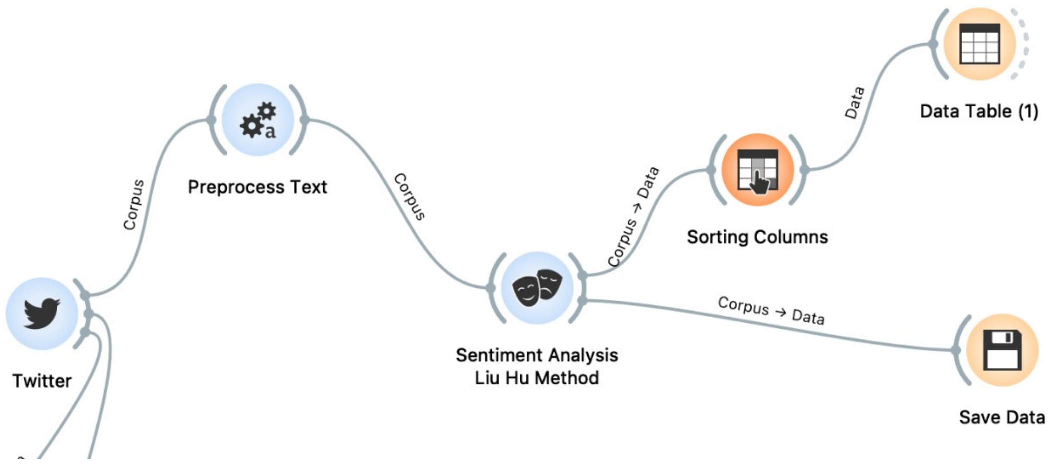
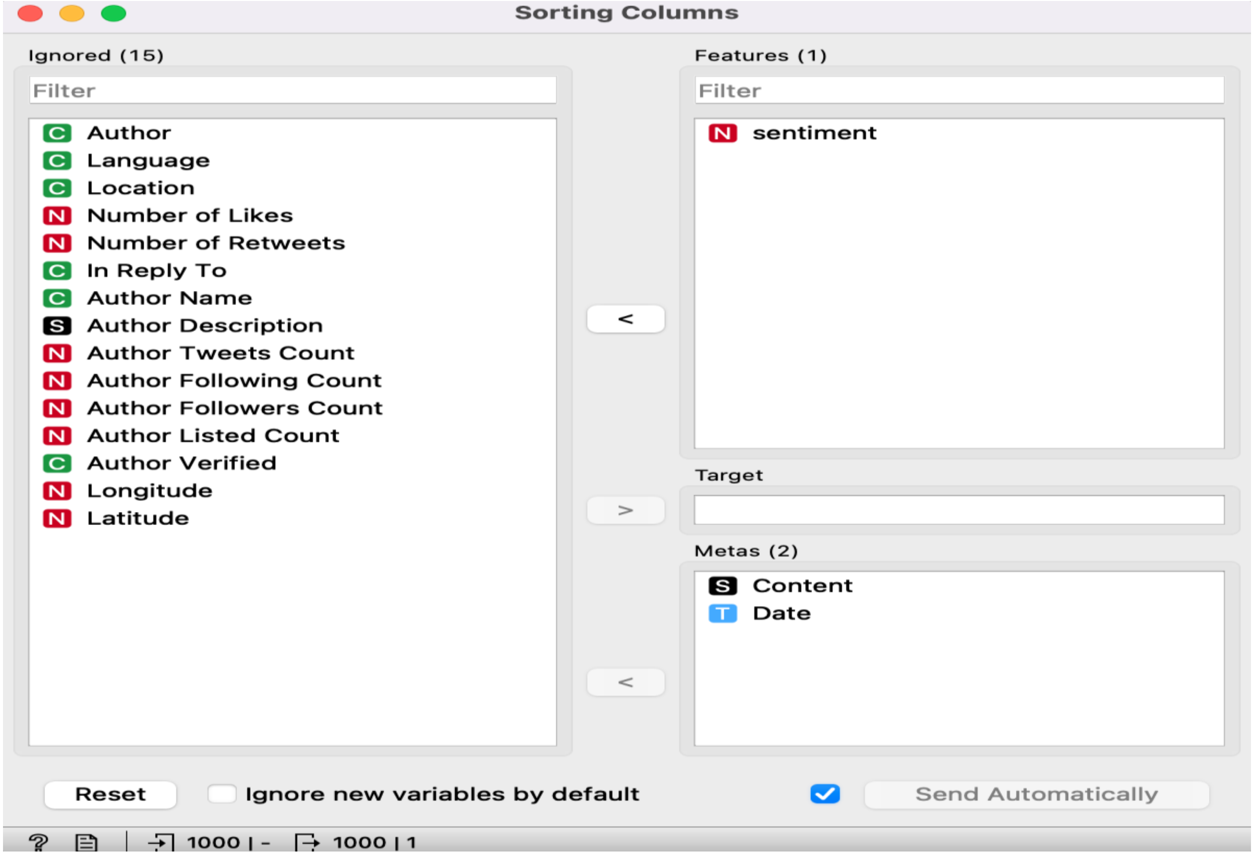


Figure 15. *Sentiment Analysis Widget (Liu Hu)*

This figure shows addition of the Sentiment analysis widget in the canvas. As an option for the method I have chosen the Liu Hu model. Also we have the Sorting Column widget which will allow us to see only the columns that we want in the data table by filtering the attributes. The Save Data widget will allow us to save the values of the table in the form that we want.



Sorting Columns

Ignored (15)

- C Author
- C Language
- C Location
- N Number of Likes
- N Number of Retweets
- C In Reply To
- C Author Name
- S Author Description
- N Author Tweets Count
- N Author Following Count
- N Author Followers Count
- N Author Listed Count
- C Author Verified
- N Longitude
- N Latitude

Features (1)

- N sentiment

Target

Metas (2)

- S Content
- T Date

□ Ignore new variables by default

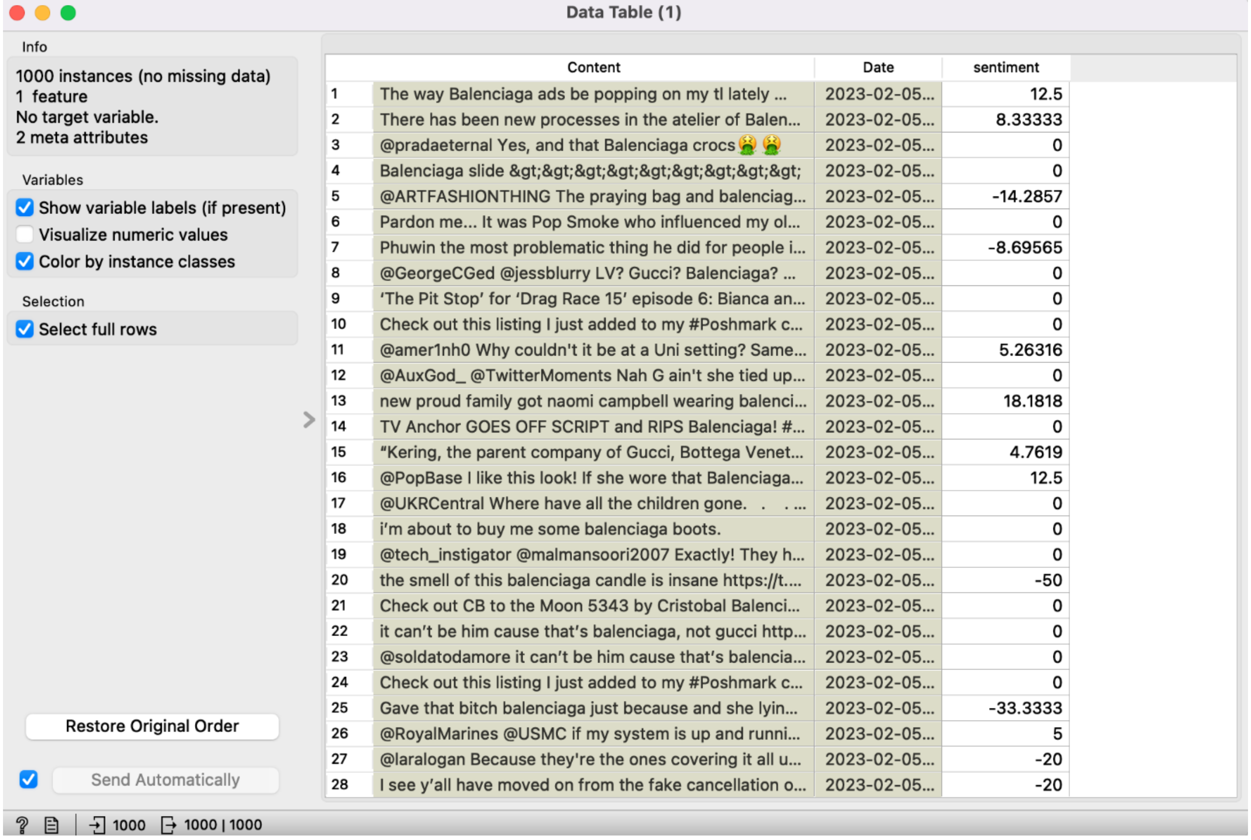
☑ Send Automatically

1000 | - 1000 | 1

Figure 16. *Sorting Column Options*

This figure shows the options of the Sorting Column widget. I have selected content and the date of the widget to be shown together with the respective sentiment value.

Table 3. *Liu Hu Sentiment Table*



Data Table (1)

Info

1000 instances (no missing data)
1 feature
No target variable.
2 meta attributes

Variables

☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

	Content	Date	sentiment
1	The way Balenciaga ads be popping on my tl lately ...	2023-02-05...	12.5
2	There has been new processes in the atelier of Balen...	2023-02-05...	8.33333
3	@pradaeternal Yes, and that Balenciaga crocs 🤮 🤮	2023-02-05...	0
4	Balenciaga slide &gt;&gt;&gt;&gt;&gt;&gt;&gt;&gt;&gt;	2023-02-05...	0
5	@ARTFASHIONTHING The praying bag and balenciag...	2023-02-05...	-14.2857
6	Pardon me... It was Pop Smoke who influenced my ol...	2023-02-05...	0
7	Phuwin the most problematic thing he did for people i...	2023-02-05...	-8.69565
8	@GeorgeCGed @jessblurry LV? Gucci? Balenciaga? ...	2023-02-05...	0
9	'The Pit Stop' for 'Drag Race 15' episode 6: Bianca an...	2023-02-05...	0
10	Check out this listing I just added to my #Poshmark c...	2023-02-05...	0
11	@amer1nh0 Why couldn't it be at a Uni setting? Same...	2023-02-05...	5.26316
12	@AuxGod_ @TwitterMoments Nah G ain't she tied up...	2023-02-05...	0
13	new proud family got naomi campbell wearing balenci...	2023-02-05...	18.1818
14	TV Anchor GOES OFF SCRIPT and RIPS Balenciaga! #...	2023-02-05...	0
15	"Kering, the parent company of Gucci, Bottega Venet...	2023-02-05...	4.7619
16	@PopBase I like this look! If she wore that Balenciaga...	2023-02-05...	12.5
17	@UKRCentral Where have all the children gone.	2023-02-05...	0
18	i'm about to buy me some balenciaga boots.	2023-02-05...	0
19	@tech_instigator @malmansoori2007 Exactly! They h...	2023-02-05...	0
20	the smell of this balenciaga candle is insane https://t....	2023-02-05...	-50
21	Check out CB to the Moon 5343 by Cristobal Balenci...	2023-02-05...	0
22	it can't be him cause that's balenciaga, not gucci http...	2023-02-05...	0
23	@soldatodamore it can't be him cause that's balencia...	2023-02-05...	0
24	Check out this listing I just added to my #Poshmark c...	2023-02-05...	0
25	Gave that bitch balenciaga just because and she lyin...	2023-02-05...	-33.3333
26	@RoyalMarines @USMC if my system is up and runni...	2023-02-05...	5
27	@laralogan Because they're the ones covering it all u...	2023-02-05...	-20
28	I see y'all have moved on from the fake cancellation o...	2023-02-05...	-20

This are the results that we obtained for 1000 tweets. We can notice on the last column that we have the numerical values representing the sentiment value. As we said before, for the Liu & Hu model, the negative values represent negative sentiment, positive values represent positive sentiment, while 0 value represent neutral sentiment.

4.2 Vader model sentiment analysis

Vader is an NLTK model that also provides the sentiments of our tweets based on the words used. But differing from Liu Hu which provides us a single normalized value for the sentiment, Vader model outputs a score for each category (positive, negative and neutral) and together with these three it outputs us a total sentiment score called a compound.

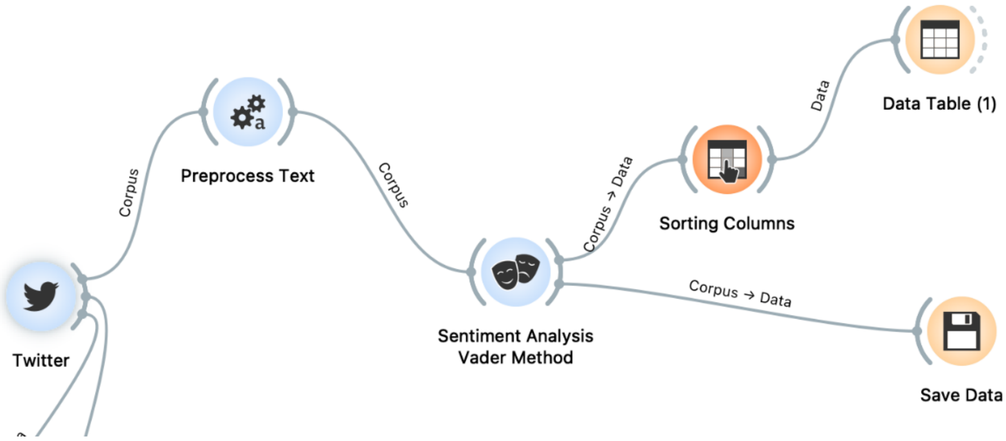
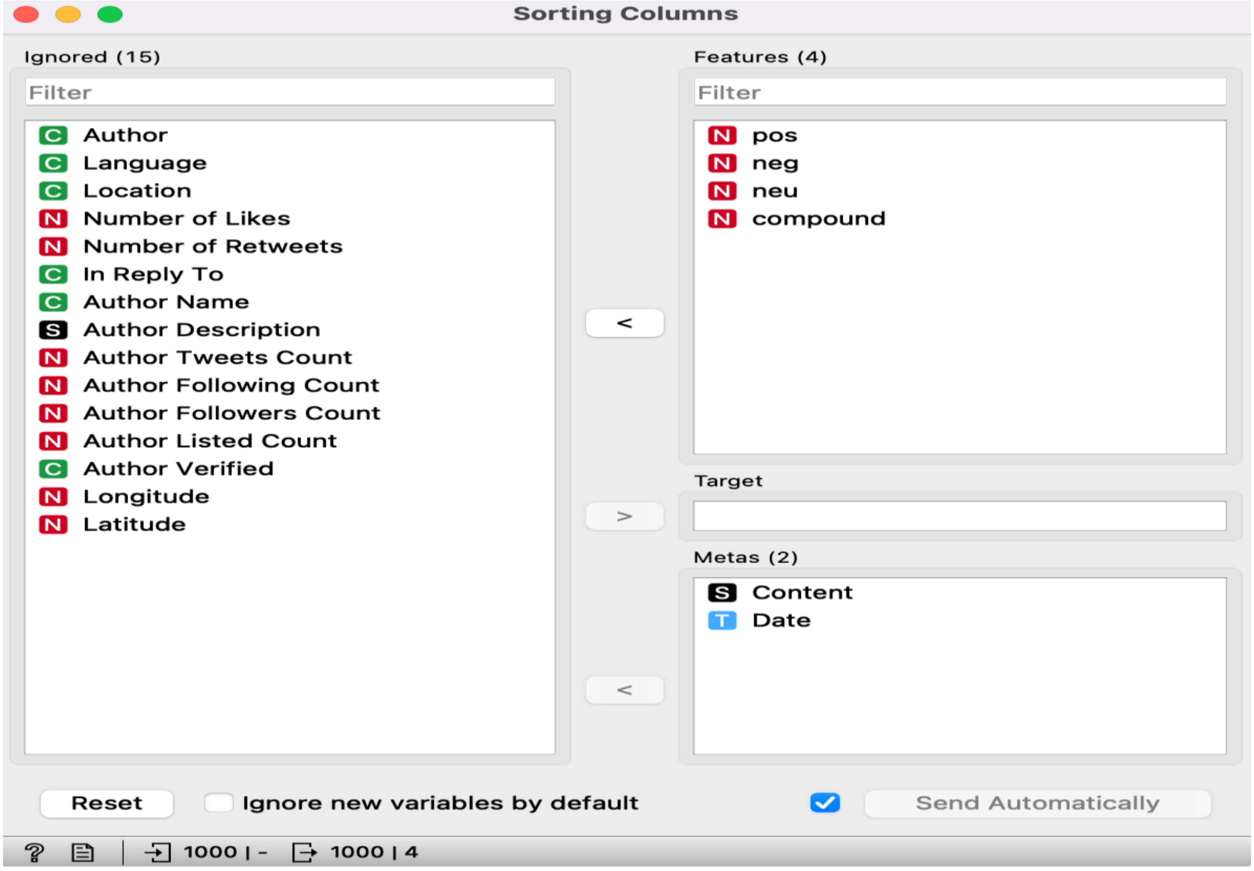


Figure 17. *Sentiment Analysis Widget (Vader)*

This figure shows the addition of the Sentiment Analysis widget (this time with the Vader model option selected). After the Sentiment widget there comes the Sorting Column widget associated with the Data Table widget that will show us the output. Also we have the Save Data Widget.



Sorting Columns

Ignored (15)

Filter

- C Author
- C Language
- C Location
- N Number of Likes
- N Number of Retweets
- C In Reply To
- C Author Name
- S Author Description
- N Author Tweets Count
- N Author Following Count
- N Author Followers Count
- N Author Listed Count
- C Author Verified
- N Longitude
- N Latitude

Features (4)

Filter

- N pos
- N neg
- N neu
- N compound

Target

Metas (2)

- S Content
- T Date

Reset ☐ Ignore new variables by default ☑ Send Automatically

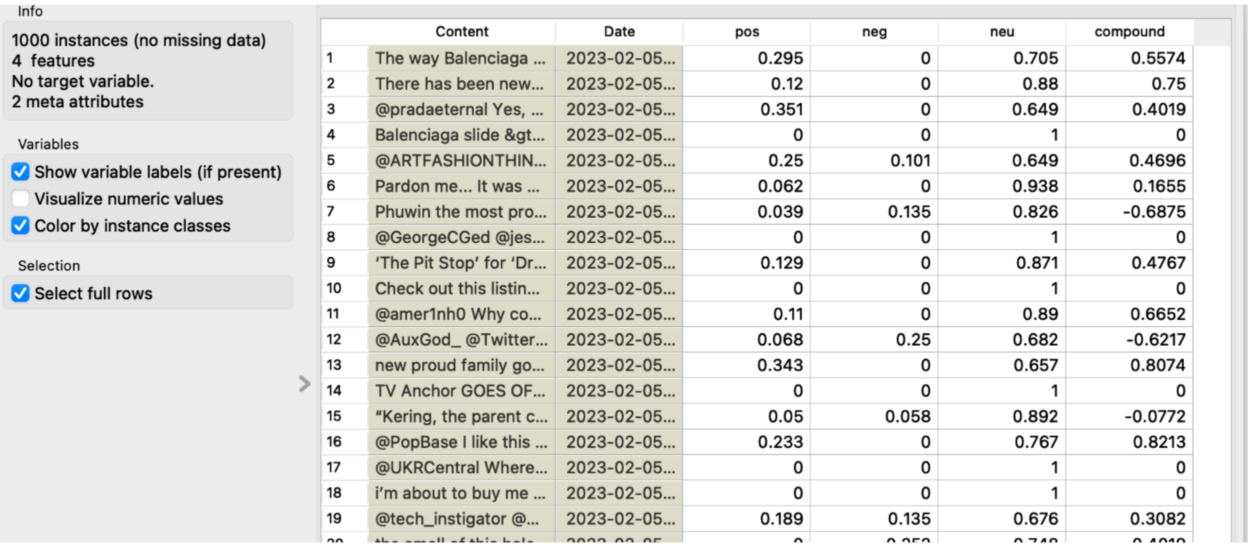
1000 | - 1000 | 4

Figure 18. *Sorting Column Options*

In the sorting column widget I have selected again the content and the date of the tweet as the attributes that will be shown as outputs in the data table together with the numerical values of the 4 kinds of sentiment values.

In this table we can see the values of each category for 1000 tweets and a total sentiment score called the compound.

Table 4. *Vader Sentiment Table*



Info

1000 instances (no missing data)
4 features
No target variable.
2 meta attributes

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

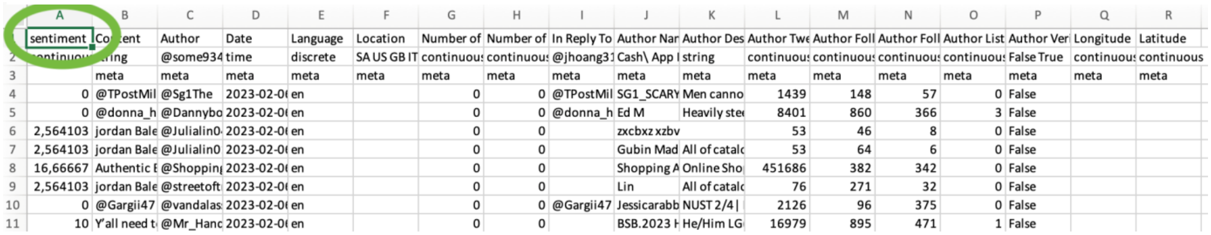
	Content	Date	pos	neg	neu	compound
1	The way Balenciaga ...	2023-02-05...	0.295	0	0.705	0.5574
2	There has been new...	2023-02-05...	0.12	0	0.88	0.75
3	@pradaeternal Yes, ...	2023-02-05...	0.351	0	0.649	0.4019
4	Balenciaga slide &gt...	2023-02-05...	0	0	1	0
5	@ARTFASHIONTHIN...	2023-02-05...	0.25	0.101	0.649	0.4696
6	Pardon me... It was ...	2023-02-05...	0.062	0	0.938	0.1655
7	Phuwin the most pro...	2023-02-05...	0.039	0.135	0.826	-0.6875
8	@GeorgeCGed @jes...	2023-02-05...	0	0	1	0
9	'The Pit Stop' for 'Dr...	2023-02-05...	0.129	0	0.871	0.4767
10	Check out this listin...	2023-02-05...	0	0	1	0
11	@amer1nh0 Why co...	2023-02-05...	0.11	0	0.89	0.6652
12	@AuxGod_ @Twitter...	2023-02-05...	0.068	0.25	0.682	-0.6217
13	new proud family go...	2023-02-05...	0.343	0	0.657	0.8074
14	TV Anchor GOES OF...	2023-02-05...	0	0	1	0
15	"Kering, the parent c...	2023-02-05...	0.05	0.058	0.892	-0.0772
16	@PopBase I like this ...	2023-02-05...	0.233	0	0.767	0.8213
17	@UKRCentral Where...	2023-02-05...	0	0	1	0
18	i'm about to buy me ...	2023-02-05...	0	0	1	0
19	@tech_instigator @...	2023-02-05...	0.189	0.135	0.676	0.3082

4.3 Analyzing Results

After the Sentiment Analysis widget provided us with sentiment values for each tweet we are going to go a little bit further. With the Save Data widget I am going to save my Table values in an excel sheet and then I will calculate the average sentiment for all 1000 tweets in both models.

4.3.1 Liu & Hu Model Results

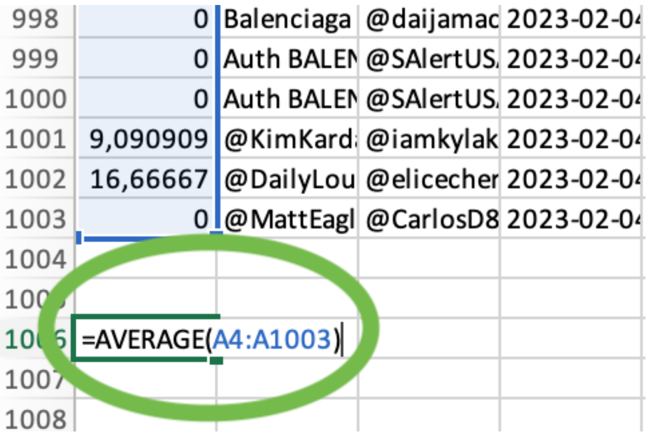
Table 5. *Liu Hu excel sheet*



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sentiment	Content	Author	Date	Language	Location	Number of	Number of	In Reply To	Author Nar	Author Des	Author Twe	Author Foll	Author Foll	Author List	Author Veri	Longitude	Latitude
2	continuou	string	@some934	time	discrete	SA US GB IT	continuou	continuou	@jhoang31	Cash\ App I	string	continuou	continuou	continuou	continuou	False True	continuou	continuous
3		meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta	meta
4	0	@TPostMil	@Sg1The	2023-02-0	en		0	0	@TPostMil	SG1_SCARY	Men canno	1439	148	57	0	False		
5	0	@donna_h	@Dannybo	2023-02-0	en		0	0	@donna_h	Ed M	Heavily ste	8401	860	366	3	False		
6	2,564103	jordan Bale	@Julialin0	2023-02-0	en		0	0		zxcbxz xzbv		53	46	8	0	False		
7	2,564103	jordan Bale	@Julialin0	2023-02-0	en		0	0		Gubin Mad	All of catalo	53	64	6	0	False		
8	16,66667	Authentic E	@Shopping	2023-02-0	en		0	0		Shopping A	Online Sho	451686	382	342	0	False		
9	2,564103	jordan Bale	@streetoft	2023-02-0	en		0	0		Lin	All of catalo	76	271	32	0	False		
10	0	@Gargii47	@vandalas	2023-02-0	en		0	0	@Gargii47	Jessicarabb	NUST 2/4 |	2126	96	375	0	False		
11	10	Y'all need t	@Mr_Hanc	2023-02-0	en		0	0		BSB.2023 H	He/Him LG	16979	895	471	1	False		

In the excel sheet above I can see that the first column is the sentiment column which has our 1000 sentiment values.

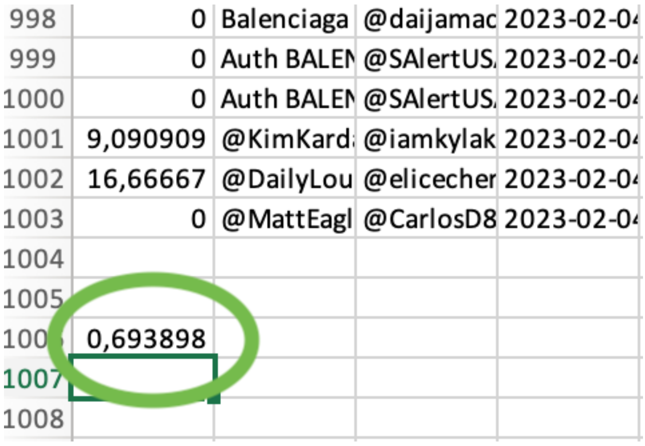
Table 6. *Average Calcuation*



998	0	Balenciaga	@daijamac	2023-02-04
999	0	Auth BALEN	@SAlertUS.	2023-02-04
1000	0	Auth BALEN	@SAlertUS.	2023-02-04
1001	9,090909	@KimKard:	@iamkylak	2023-02-04
1002	16,66667	@DailyLou	@elicecher	2023-02-04
1003	0	@MattEagl	@CarlosD8	2023-02-04
1004				
1005				
1006	=AVERAGE(A4:A1003)			
1007				
1008				

With the Help of the Average formula in Excel I can easily calculate the average of all rows which have sentiment values.

Table 7. *Average Result*

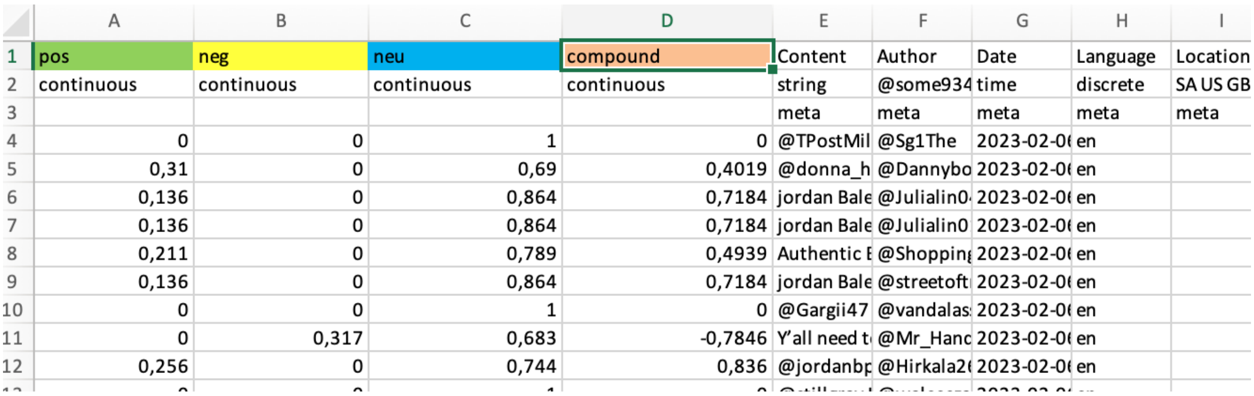


998	0	Balenciaga	@daijamac	2023-02-04
999	0	Auth BALEN	@SAlertUS.	2023-02-04
1000	0	Auth BALEN	@SAlertUS.	2023-02-04
1001	9,090909	@KimKard:	@iamkylak	2023-02-04
1002	16,66667	@DailyLou	@elicecher	2023-02-04
1003	0	@MattEagl	@CarlosD8	2023-02-04
1004				
1005				
1006	0,693898			
1007				
1008				

Here we can see the computed result of the overall sentiment. The result gives us a positive value so the overall sentiment for the Balenciaga brand appears to be positive.

4.3.2 Vader Model Results

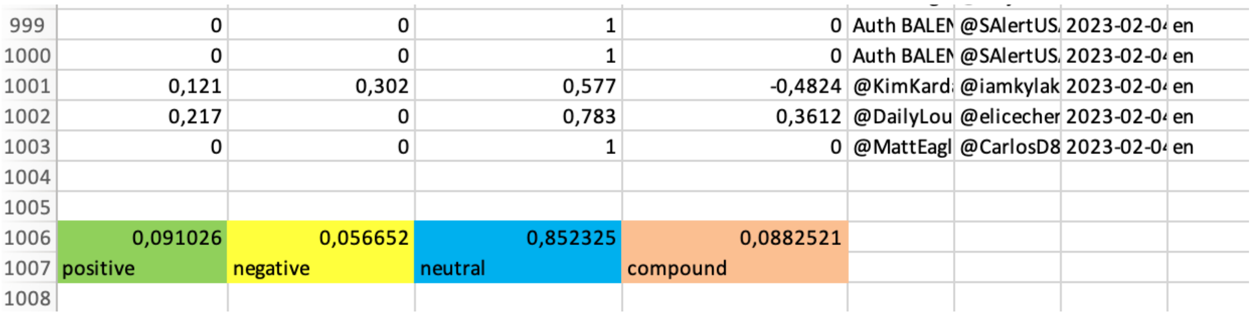
Table 8. *Vader Excel sheet*



	A	B	C	D	E	F	G	H	I
1	pos	neg	neu	compound	Content	Author	Date	Language	Location
2	continuous	continuous	continuous	continuous	string	@some934	time	discrete	SA US GB
3					meta	meta	meta	meta	meta
4	0	0	1	0	@TPostMil	@Sg1The	2023-02-0(	en	
5	0,31	0	0,69	0,4019	@donna_h	@Dannybo	2023-02-0(	en	
6	0,136	0	0,864	0,7184	jordan Bale	@Julialin0	2023-02-0(	en	
7	0,136	0	0,864	0,7184	jordan Bale	@Julialin0	2023-02-0(	en	
8	0,211	0	0,789	0,4939	Authentic E	@Shopping	2023-02-0(	en	
9	0,136	0	0,864	0,7184	jordan Bale	@streetoft	2023-02-0(	en	
10	0	0	1	0	@Gargii47	@vandalas	2023-02-0(	en	
11	0	0,317	0,683	-0,7846	Y'all need t	@Mr_Hanc	2023-02-0(	en	
12	0,256	0	0,744	0,836	@jordanbp	@Hirkala2(	2023-02-0(	en	

Here we can see the excel sheet that we obtained from the Vader Model sentiment analysis. The first four columns show the values of the sentiments for each tweet. As we said before for the Vader model we have 4 values for each tweet (positive, negative, neutral and compound).

Table 9. *Vader averages results*



999	0	0	1	0	Auth BALEN	@SAlertUS.	2023-02-0	en
1000	0	0	1	0	Auth BALEN	@SAlertUS.	2023-02-0	en
1001	0,121	0,302	0,577	-0,4824	@KimKard	@iamkylak	2023-02-0	en
1002	0,217	0	0,783	0,3612	@DailyLou	@elicecher	2023-02-0	en
1003	0	0	1	0	@MattEagl	@CarlosD8	2023-02-0	en
1004								
1005								
1006	0,091026	0,056652	0,852325	0,0882521				
1007	positive	negative	neutral	compound				
1008								

At the end of each column I have calculated the average for each column and those are the results that I obtained.

In the Vader model we interpret the results like in the following way. When we have a positive tweet, the compound score should be >=0.05. When we have a negative tweet, the compound score should be <=-0.05. While when have a neutral tweet the compound score should be >-0.05 and <0.05.

In our dataset the average of the Compound score is around 0.088 which is >0.05. So, in overall the sentiment about the Balenciaga brand according to Vader Sentiment analysis model is again positive.

4.4 Discussions

In the following picture we can see my whole work through the thesis. All the widgets that are used are shown in the canvas together with all the relations between them. We can see that the Preprocess text widget, which performs all the elements of NLP is the initialization of every other widget that perform functions.

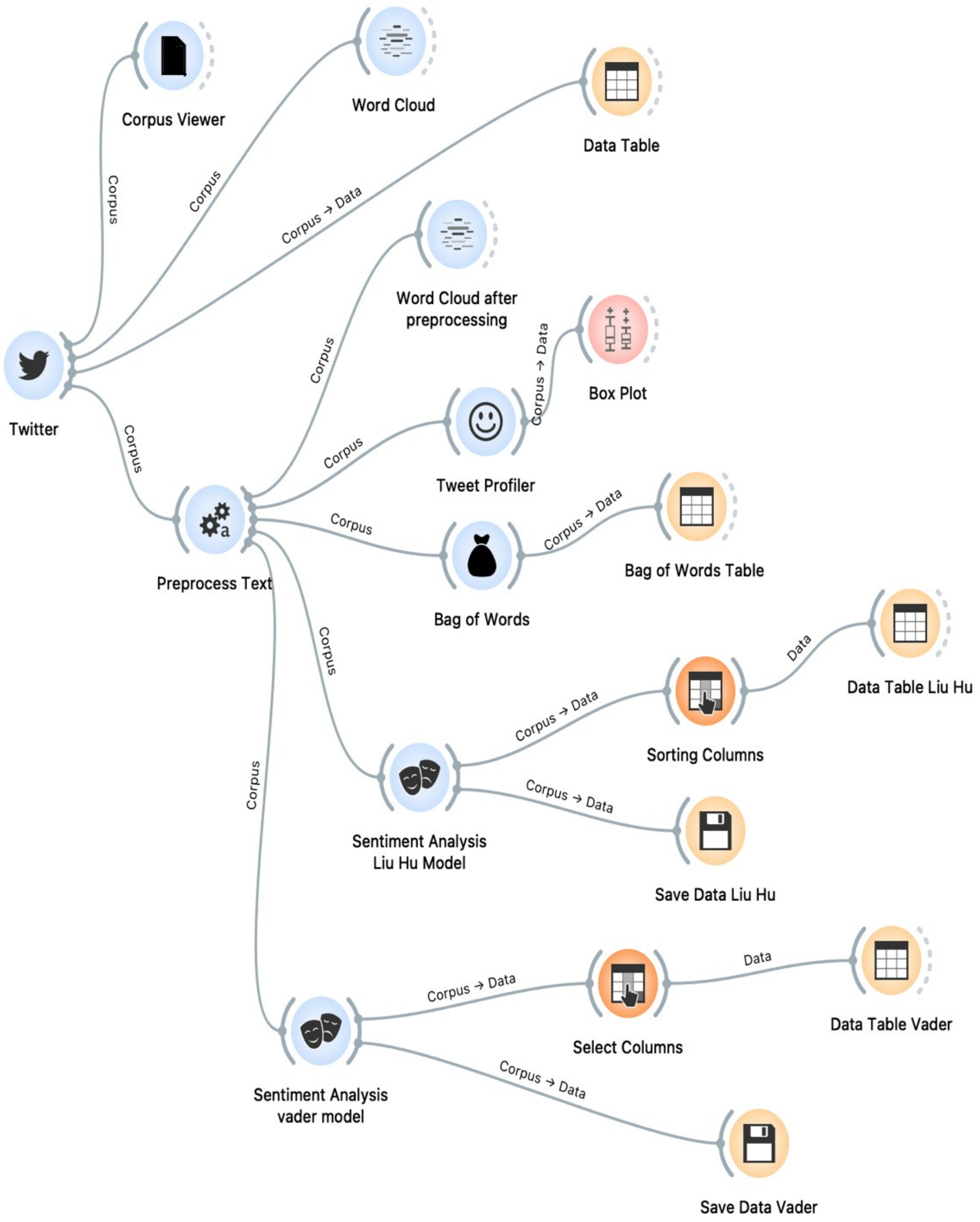


Figure 19. *Full Canvas*

CHAPTER 5

CONCLUSIONS

5.1 Conclusions

The aim of this thesis was to introduce readers with NLP as a branch of AI which gives machines ability to read human language, understand it and also derive meanings from it. I used Orange platform to perform some of the NLP elements. Understanding semantic meaning and performing sentiment analysis is one of the best applications of NLP these days. When it comes to social media, we can say that it there are some huge platforms which are part of it. Twitter is one platform of social media where people are always expressing comments about their everyday life, experiences, products they use or buy and a multiple number of other elements. My thesis fundamental part was to present to different companies, how they can understand people's opinions about the products they are launching or the services they are offering. Overall, I can conclude that using Orange platform can be very useful in analyzing sentiment of comments in Twitter platform.

5.2 Future Work

There are also some very other social media platforms existing these days such as Instagram, Facebook, Tik Tok, etc. Analyzing data and comments from these platforms can be a field of big interest that can help companies a lot in their campaigns and product reviews. Finding suitable ways to import their datasets in Orange or some other platforms can lead to very interesting results.

REFERENCES

1. Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. *Evaluating Natural Language Understanding Services for Conversational Question Answering Systems.* In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.
2. Han van der Aa, Josep Carmona, Henrik Leopold, Jan Mendling, and Lluís Padró. 2018. Challenges and Opportunities of Applying Natural Language Processing in Business Process Management. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2791–2801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
3. Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
4. Sara Rosenthal, Kathy McKeown, and Apoorv Agarwal. 2014. *Columbia NLP: Sentiment Detection of Sentences and Subjective Phrases in Social Media.* In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 198–202, Dublin, Ireland. Association for Computational Linguistics.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
6. Rajput, G., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021, December). *Hate speech detection using static BERT embeddings.* In *International Conference on Big Data Analytics* (pp. 67-77). Springer, Cham.
7. Ravi Arunachalam and Sandipan Sarkar. 2013. *The New Eye of Government: Citizen Sentiment Analysis in Social Media.* In *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 23–28, Nagoya, Japan. Asian Federation of Natural Language Processing.

8. Arunachalam, R., & Sarkar, S. (2013, October). The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP)* (pp. 23-28).
9. Mosca, E., Wich, M., & Groh, G. (2021, June). Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media* (pp. 91-102).
10. Anna Schmidt and Michael Wiegand. 2017. *A Survey on Hate Speech Detection using Natural Language Processing.* In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.