

# Prova de Conceito #2

## Descrição do desafio

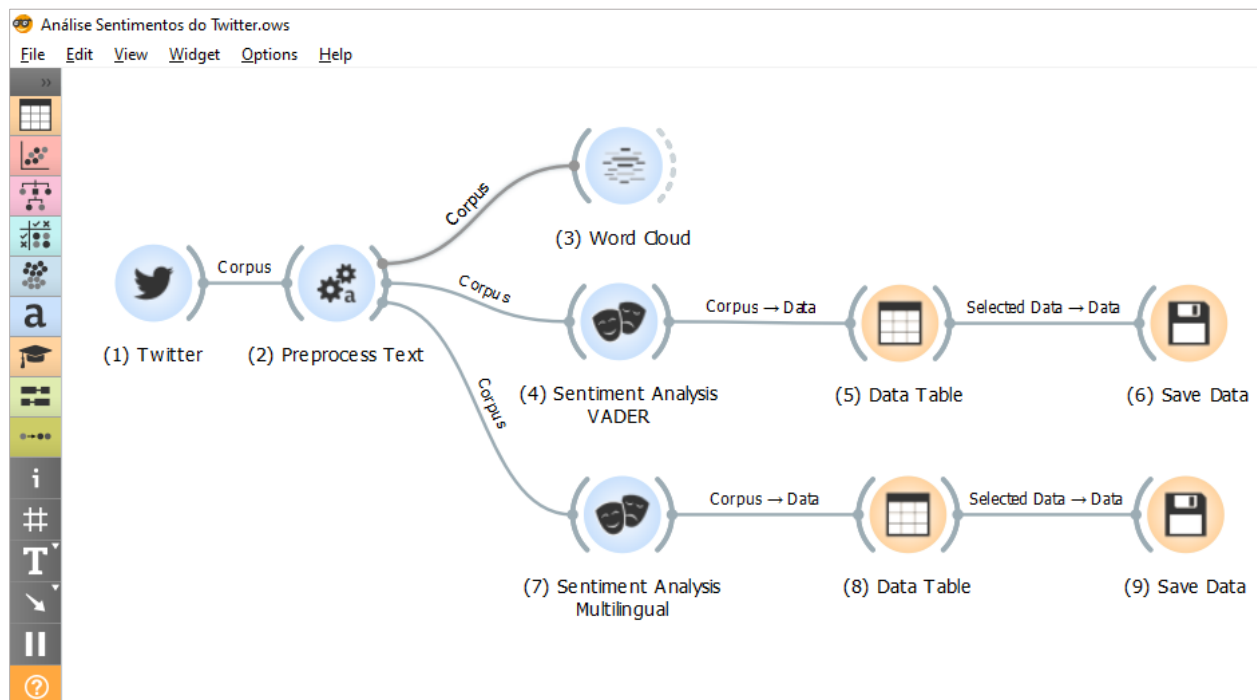
Perante um conjunto de tweets como input, retornar como output uma classificação de sentimento desses tweets.

## Ferramentas utilizadas

- Orange Data Mining
- API do Twitter
- Microsoft Excel

## Estrutura do projeto na aplicação Orange

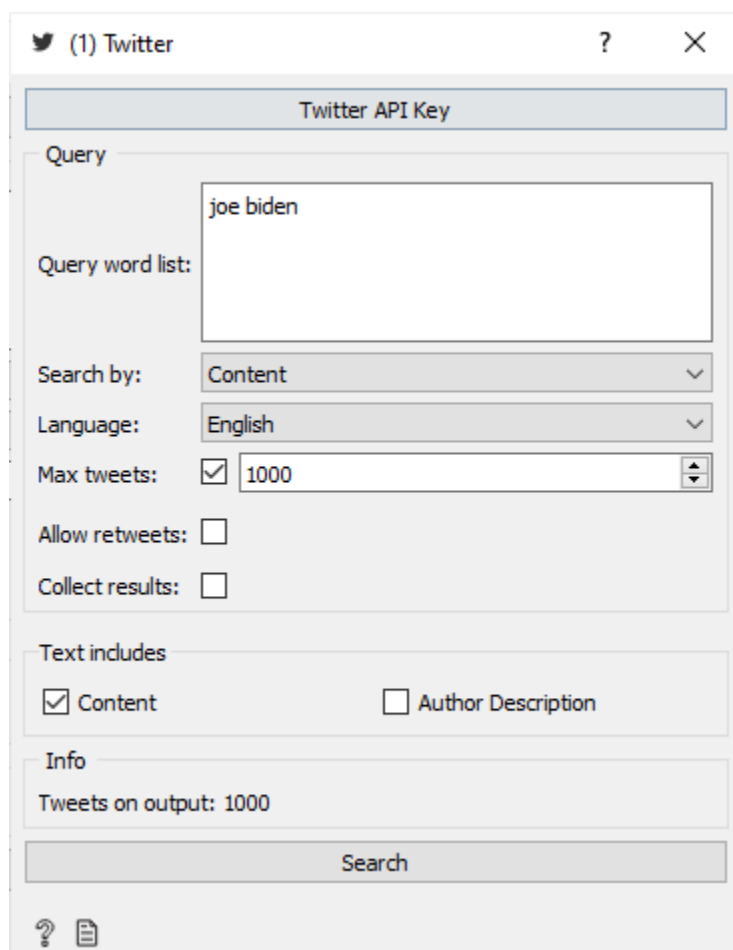
A aplicação Orange Data Mining foi utilizada neste desafio pois disponibiliza funcionalidades que permitem obter tweets do Twitter e fazer análise de sentimentos.



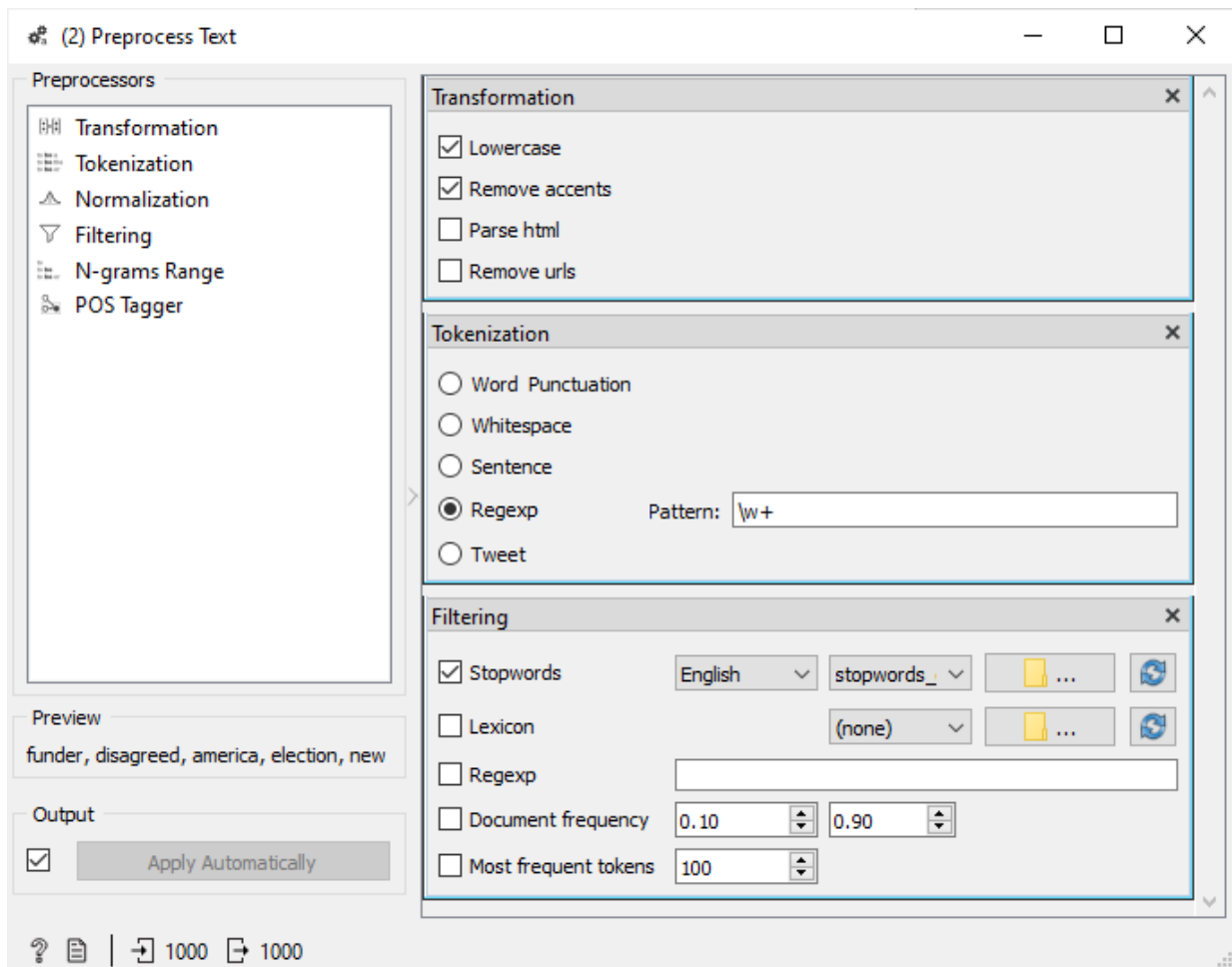
## Descrição da estrutura do projeto

Para estruturar o projeto no Orange, primeiramente foi obtida a chave da API do Twitter, usando o email da minha conta pessoal. Também foi instalado o add-on de Text Mining (que inclui o Widget do Twitter e o Widget da Análise de Sentimentos), através dos menus Options – Add-ons da aplicação Orange. Foram então adicionados os seguintes widgets:

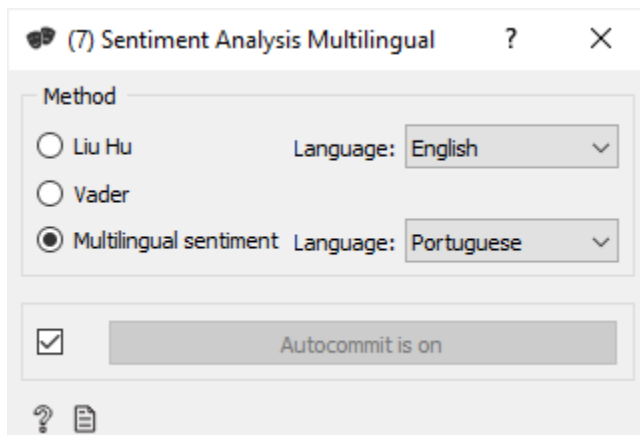
- (1) Twitter – Após inserir as credenciais obtidas para aceder à API do Twitter, esta funcionalidade apresenta um campo para definir as palavras que queremos ver mencionadas nos tweets que queremos obter. Também é possível escolher o tipo de pesquisa, por conteúdo ou autor, o idioma e o número máximo de tweets obtidos.



- (2) Preprocess Text – Antes de analisar os tweets, foi feito um pré-processamento do output obtido. Com esta funcionalidade, é possível transformar o texto, colocando-o, por exemplo, com letras minúsculas e removendo os acentos. O conteúdo dos tweets foi também filtrado de forma a excluir stop words, palavras que são consideradas irrelevantes para análise.



- (3) Word Cloud – representação visual das palavras dos Tweets obtidos, onde o tamanho de cada palavra indica a frequência com que aparece no output.
- (4) Sentiment Analysis VADER e (7) Sentiment Analysis Multilingual – a aplicação Orange disponibiliza 3 métodos de análise de sentimentos: Liu Hu, Vader e Multilingual. Neste desafio, vamos testar os métodos Vader, para tweets de língua inglesa, e o Multilingual, para tweets de língua portuguesa.



- (5) e (8) Data Table – Apresenta, numa tabela, os dados obtidos dos tweets e a respetiva análise de sentimentos.
- (6) e (9) Save Data – Funcionalidade que guarda os dados da tabela gerada num ficheiro Excel.

## Apresentação e análise dos resultados obtidos

Para a realização deste desafio, foram feitas 3 pesquisas diferentes, onde foram obtidos 3000 tweets que mencionassem as seguintes Strings:

- “happy” (em inglês)
- “joe biden” (em inglês)
- “europa” (em português)

Após a obtenção dos tweets, cada pesquisa passou por um pré-processamento para filtrar a informação a ser tratada, de forma a facilitar a análise.

O texto de cada tweet foi transformado para ficar com letras minúsculas e os acentos e os urls foram removidos. O texto também foi filtrado de forma a serem analisadas apenas as palavras do conteúdo de cada tweet.

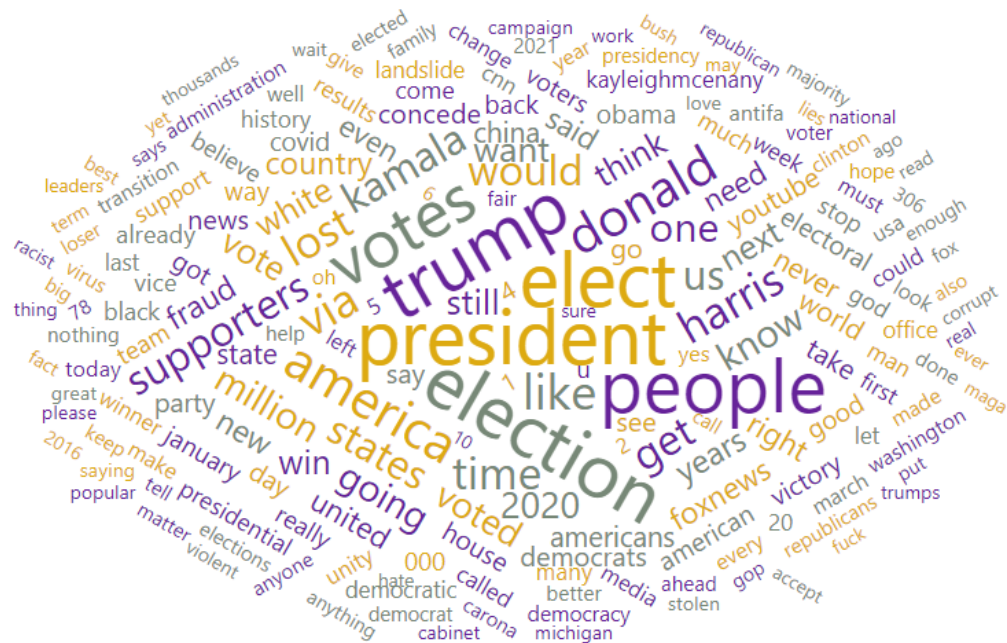
Para finalizar esta fase, foram definidos ficheiros de stop words personalizados para cada pesquisa (*stopwords\_en\_happy.txt*, *stopwords\_en\_joebiden.txt*, *stopwords\_pt\_europa.txt*). De acordo com o idioma do conteúdo dos tweets pesquisados, foram adicionadas as stop words mais usadas em inglês e português. Também fez sentido adicionar a palavra pesquisada à lista (por exemplo, se estamos a pesquisar “europa”, então essa palavra deve estar também na lista de stop words pois vai aparecer com frequência nos resultados).

De seguida, são apresentadas, em Word Clouds, as palavras que aparecem com mais frequência no conteúdo dos tweets obtidos.

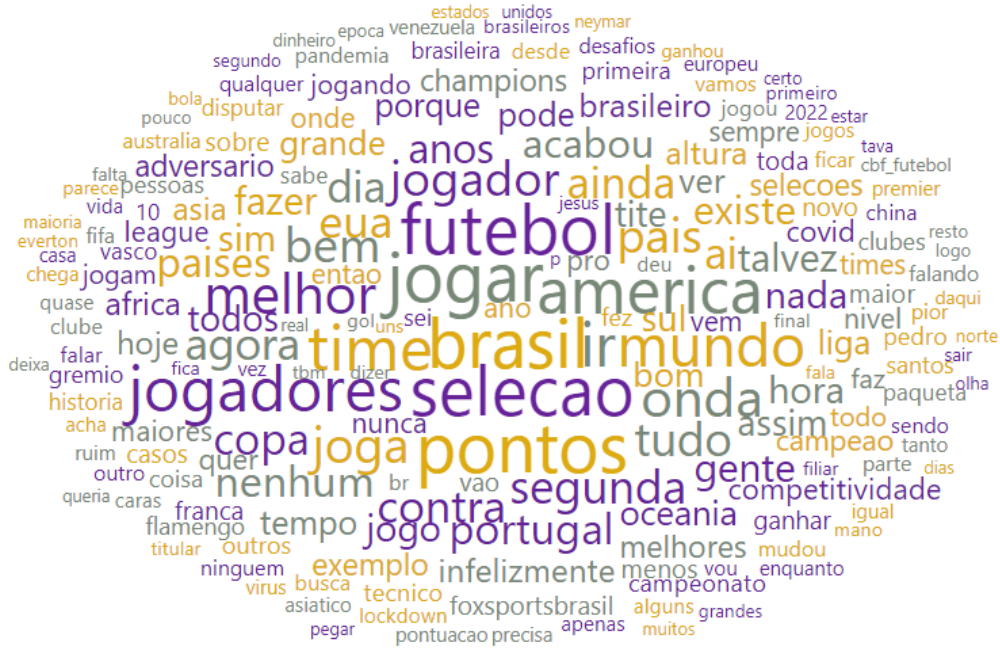
### Word Cloud – Tweets com texto “happy” mencionado



Word Cloud – Tweets com texto “joe biden” mencionado



## Word Cloud – Tweets com texto “europa” mencionado



Fazendo uma análise rápida, conseguimos perceber que a pesquisa de tweets que mencionam a palavra “happy” contêm palavras que transmitem sentimentos mais positivos (exemplo: hope, love, good). As Word Clouds das outras duas pesquisas apresentam palavras relacionadas com o texto pesquisado e os sentimentos transmitidos são mais variados.

## Análise de sentimentos

Foram usados 2 algoritmos de análise de sentimentos: o algoritmo Vader, para as pesquisas de tweets ingleses e o algoritmo Multilingual, para a pesquisa de tweets portugueses.

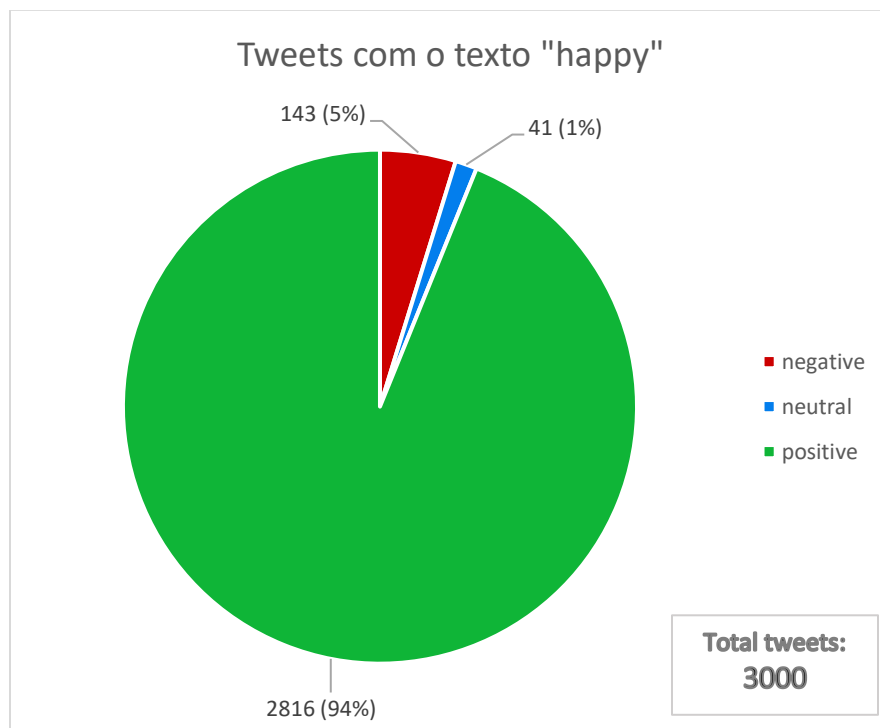
O algoritmo Vader (Valence Aware Dictionary and sEntiment Reasoner) é uma ferramenta de análise de sentimentos baseada em regras e foi desenvolvida especificamente para analisar sentimentos transmitidos nas redes sociais. Para cada análise feita, foram o algoritmo cria 4 colunas com número decimais para representar os sentimentos de cada comentário: pos, neg, neu, compound. A coluna compound faz uma média das outras três colunas e foi baseado nos valores desta coluna que a análise de sentimentos foi feita.

Na análise onde é usado o algoritmo Multilingual, este algoritmo foi escolhido porque apresenta uma lista de idiomas que está apto para analisar, incluindo o português. Este tipo de análise de sentimentos cria 1 coluna (sentiment) com números decimais para representar o sentimento transmitido em cada comentário.

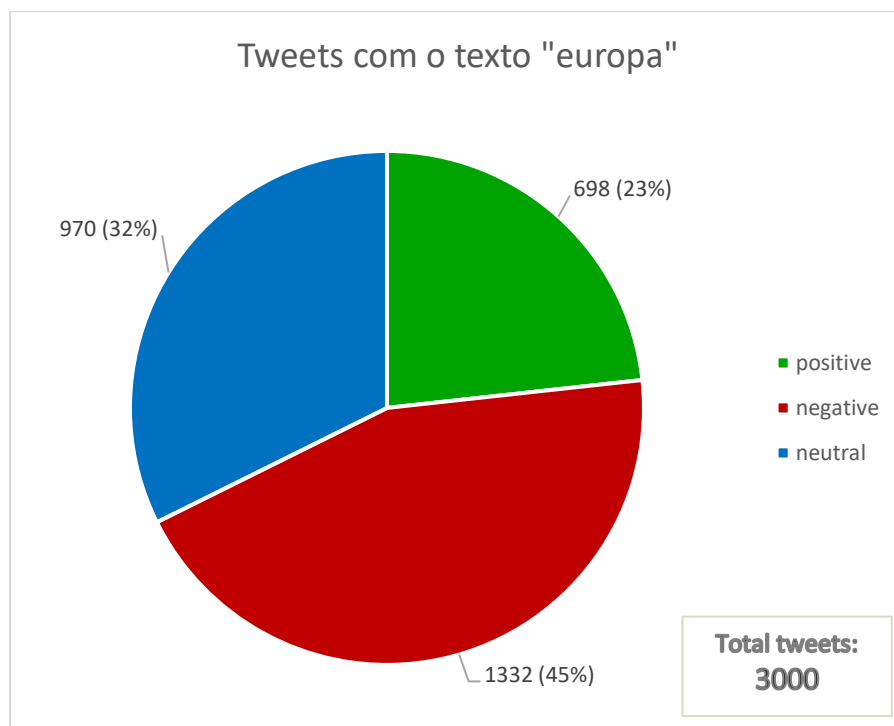
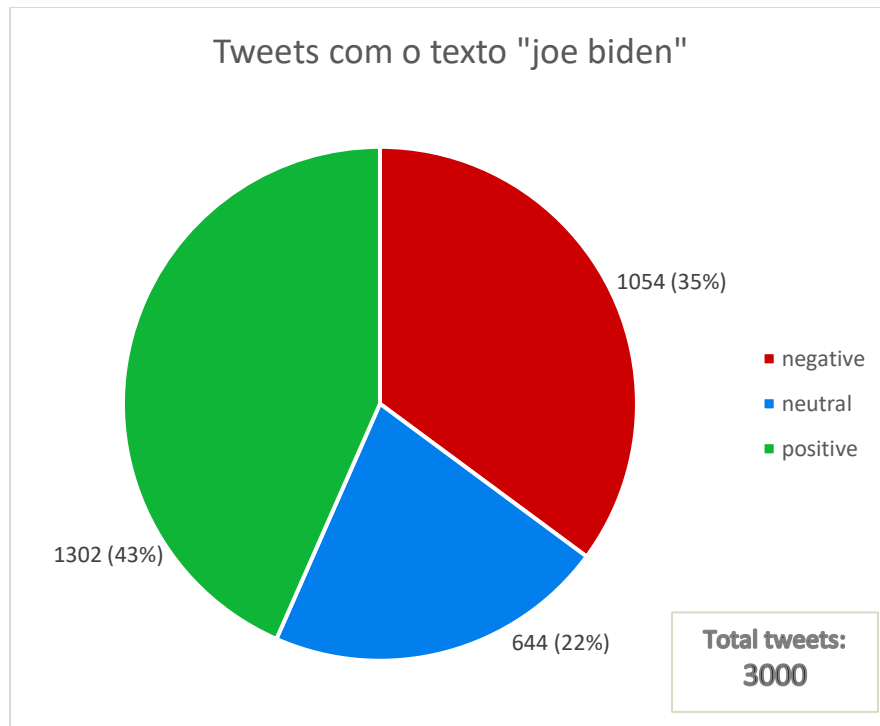
Para cada uma das 3 análises, foi extraído um Excel (através das funcionalidades Data Table -> Save Data) com as informações referentes a cada tweet e com a análise feita pelo algoritmo utilizado.

A fim de gerar um gráfico para facilitar a leitura proveniente da análise de sentimentos, foi adicionada uma coluna, no fim de cada tabela, com o nome “Sentiments”. Esta coluna pode ter os valores “positive”, se o valor da análise feita for superior a 0, “negative” se o valor da análise feita for inferior a 0 e “neutral”, se o valor da análise feita for igual a 0.

A partir destes dados, foram produzidos os seguintes gráficos:



Como seria de esperar, o sentimento geral transmitido pelos tweets, onde o texto “happy” é mencionado, é positivo.



Os sentimentos analisados na pesquisa feita, onde os textos “joe biden” e “europa” são mencionados, são mais variados. Analisando o conteúdo dos comentários é possível compreender os sentimentos resultantes.



## Conclusão

Nesta Prova de Conceito, foi pedido fazer a análise de sentimentos a partir de um conjunto de comentários obtidos através da API do Twitter. A implementação deste desafio foi feita utilizando a ferramenta de Data Mining Orange, onde foi obtida uma amostra de tweets reais e analisados os sentimentos transmitidos em cada um.

Como melhorias, a análise de sentimentos poderia ser feita utilizando o algoritmo Liu Hu.