IEC - Instituto de Educação Continuada

Pós-Graduação em Inteligência Artificial

Recuperação da Informação na Web e em Redes Sociais

FALCÃO E O SOLDADO INVERNAL:

Análise de sentimentos dos *tweets* coletados durante a estreia da série

Aluno: Natália da Silva Antunes

Professor: Zilton Cordeiro Jr.

Março

2021

IEC - Instituto de Educação Continuada

Pós-Graduação em Inteligência Artificial

Projeto Final

FALCÃO E O SOLDADO INVERNAL:

Análise de sentimentos dos *tweets* coletados durante a estreia da série

Trabalho apresentado ao Instituto de Educação Con- tinuada (IEC) da pós-graduacão em Inteligência Artifical da PUC Minas, como requisito par- cial para a obtenção de créditos na disciplina de Recuperação da Informação na Web e em Redes So- ciais.

Aluno: Natália da Silva Antunes

Professor: Zilton Cordeiro Jr

Março

2021

Conteúdo

1. Resumo

O presente trabalho tem como objetivo analisar os tweets coletados no dia da estreia do primeiro episódio da série *Falcão e o Soldado Invernal*, do serviço de streaming *Disney+*, do *The Walt Disney Company*, por meio da ferramenta *Knime*. A busca pelos tweets foi baseada nos critérios de 1) publicação com a *hashtag* oficial da série e 2) Busca pelos nomes dos dois personagens título da série, tendo sido recolhidos nos turnos manhã, tarde e noite do dia 19/03/2021. Após essa etapa, um banco de dados foi criado com todos os tweets coletados e, posteriormente, foi feita a análise das informações, fase necessária para investigar a polaridade das informações, ou seja, examinar se os tweets expressavam sentimentos negativos ou positivos. Conclui-se que alguns sentimentos positivos expostos nos tweets estavam relacionados à amizade, ao bem, amigos e amizades, enquanto os sentimentos negativos, ao defeito, ao comum e ao Ultimato, uma clara referência ao filme *Vingadores- Ultimato*, que pertence a mesma franquia da série. Ademais, sugere-se realizar a análise coletando os tweets durante o final de semana para que haja uma quantidade expressiva de informações. Além disso, recomenda-se a coleta de tweets durante toda a temporada da série.

Palavras-chave: Nuvem de palavras; tweets; streaming; análise de sentimento.

2. Introdução

Em novembro de 2019, a Disney lançou nos EUA um serviço de streaming de vídeo chamado *Disney+*, criado tanto para ofertar séries, filmes, documentários, animações da Disney, Marvel, Pixar, Fox, dentre outros, como, também, para competir com diversos outros serviços de streaming, por exemplo, a Netflix, que hoje em dia produz e comanda o cenário audiovisual mundial. Sabendo da importância que as bibliotecas de filmes e séries têm atualmente, a Disney incorpora em seu catalogo produções originais do Universo Cinematográfico Marvel, o MCU, franquia que ajudou a popularizar os super-heróis nos últimos quinze anos.

A série que serve como base para este trabalho nasce desse contexto: Uma produção cinematográfica original de uma franquia popular em um serviço de streaming. A série *Falcão e o Soldado Invernal*, o mais novo lançamento da Disney, teve o seu primeiro episódio lançado mundialmente no dia 19 de março de 2021. Essa é a segunda série que compõe a quarta fase do MCU, sendo o início da quarta fase marcado pela estreia da série *Wandavision*, que se tornou a série mais assistida da atualidade ao alcançar o episódio cinco, de acordo com o site Omelete, o que solidifica a popularidade do serviço de streaming e da franquia do MCU.

Para contextualizar e justificar a importância da série *Falcão e o Soldado Invernal* é preciso relatar um pouco mais as fases do MCU e cada um de seus objetivos. A primeira fase, de 2008 até 2012, foi iniciada pelo filme Homem de Ferro e encerrada pelo filme Os Vingadores. Esse momento sinaliza a ascensão dessas figuras. Após a primeira fase, o MCU se refaz, entre 2013 e 2015, a partir dos efeitos causados pelo grupo no filme de 2012. Por fim, a terceira fase, de 2016 até 2019, ocupa-se com a queda e a reestruturação dos heróis e do mundo, uma vez que a grande ameaça, cuja todas as fases estavam preparando-os, torna-se inevitável e invencível.

A quarta fase, que está decorrendo no momento, com dito anteriormente, inicia-se em *Wandavison*, atravessa *Falcão e o Soldado Invernal* e posteriormente outras séries ainda não lançadas pela Disney, tendo como objetivo fusionar a experiência da televisão e do cinema, proporcionando ao telespectador uma nova forma de consumir as histórias tão conhecidas e tão aclamadas. Por esse motivo, a série *Falcão e o Soldado Invernal* não era apenas muito esperada como também recaía sobre ela o peso de dar uma boa continuidade para a quarta fase do MCU. Assim, neste trabalho será abordada a análise de tweets coletados no dia da estreia do primeiro episódio da série *Falcão e o Soldado Invernal*.

3. Descrição das Atividades

3.1. Coleta de Dados

A coleta de dados foi realizada no dia da estreia da série, utilizando a ferramenta Knime. Optou-se por buscar os tweets mais recentes e/ ou populares seguindo os seguintes critérios:

- Busca pelos tweets publicados com a *hashtag* oficial da série (#FalcaoeoSoldadoInvernal).
- Busca pelos nomes dos dois personagens título da série, desde que estivessem na mesma sentença.

Após a disponibilização do episódio na plataforma de streaming, o espectador pode assisti-lo a qualquer momento, não sendo possível prever um horário em que a maioria dos espectadores estão assistindo ao episódio. No intuito de recuperar a maior informação possível além de tentar minimizar um possível viés na análise, optou-se pela coleta de tweets em três etapas ao longo do dia, distribuídas da seguinte maneira:

- 1ª coleta realizada no turno da manhã
- 2ª coleta realizada no turno da tarde
- 3ª coleta realizada no turno da noite

A figura 1 apresenta o fluxo de coleta de dados. Uma vez que os tweets foram coletados em horários diferentes, cada coleta foi armazenada em uma base distinta e posteriormente foram agregadas, criando-se assim uma única base de dados.

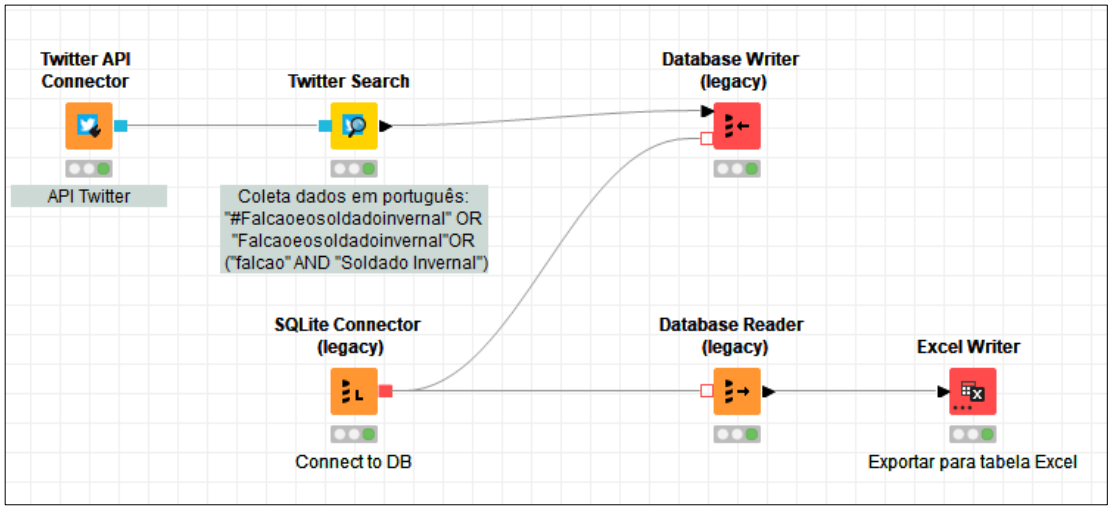

Figura 1 Fluxo de coleta de tweets

3.2. Banco de dados e Pré-Processamento da base

Após as etapas de coletas de dados, as três bases foram unificadas, criando-se um único arquivo. Foram acrescentadas duas novas colunas, agregando as informações do turno da coleta (manhã, tarde ou noite) e o grupo da coleta (coleta 1 se turno manhã, coleta 2 se turno tarde e coleta 3 se turno noite).

Ao todo foram coletados 19.500 tweets, sendo que 12.373 (63%) são retweets, o que representa a maior parte dos textos publicados. De acordo com o site oficial do Twitter, entende-se retweet como "*O Tweet que você compartilha publicamente com seus seguidores é conhecido como um Retweet. Essa é uma ótima maneira de transmitir notícias e descobertas interessantes no Twitter*.". Por esse motivo, optou-se por mantê-los nas análises uma vez que eles expressam, de uma certa maneira, a opinião do usuário.

As variáveis geográficas do tweet e o idioma do usuário não foram carregadas durante a coleta dos dados. A figura 2 descreve o banco de dados que será utilizado nas análises seguintes.

```
base.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19500 entries, 0 to 19499
Data columns (total 28 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Grupo                 19500 non-null  object
 1   turno_coleta          19500 non-null  object
 2   Tweet                 19500 non-null  object
 3   Tweet ID              19500 non-null  int64
 4   Time                  19500 non-null  object
 5   Favorited             19500 non-null  int64
 6   Retweeted             19500 non-null  int64
 7   Is Favourited         19500 non-null  bool
 8   Is Retweeted          19500 non-null  bool
 9   Is Retweet            19500 non-null  bool
 10  Retweet from          12373 non-null  object
 11  Latitude              0 non-null      float64
 12  Longitude             0 non-null      float64
 13  Country               160 non-null    object
 14  User                  19500 non-null  object
 15  User - Profile image  19422 non-null  object
 16  User - Name           19500 non-null  object
 17  User - ID             19500 non-null  int64
 18  User - Description    17801 non-null  object
 19  User - URL            7173 non-null   object
 20  User - Creation time  19500 non-null  object
 21  User - Language       0 non-null      float64
 22  User - Location       13648 non-null  object
 23  User - Time Zone      0 non-null      float64
 24  User - Statuses       19500 non-null  int64
 25  User - Followers      19500 non-null  int64
 26  User - Friends        19500 non-null  int64
 27  User - Favourites     19500 non-null  int64
dtypes: bool(3), float64(4), int64(8), object(13)
memory usage: 3.8+ MB
```

Figura 2 Descrição banco de dados

Foi realizada também alguns refinamentos na base de dados. Utilizando -se o nodo "Java Snippet", foram removidas do documento os links, hashtags dos tweets originais.

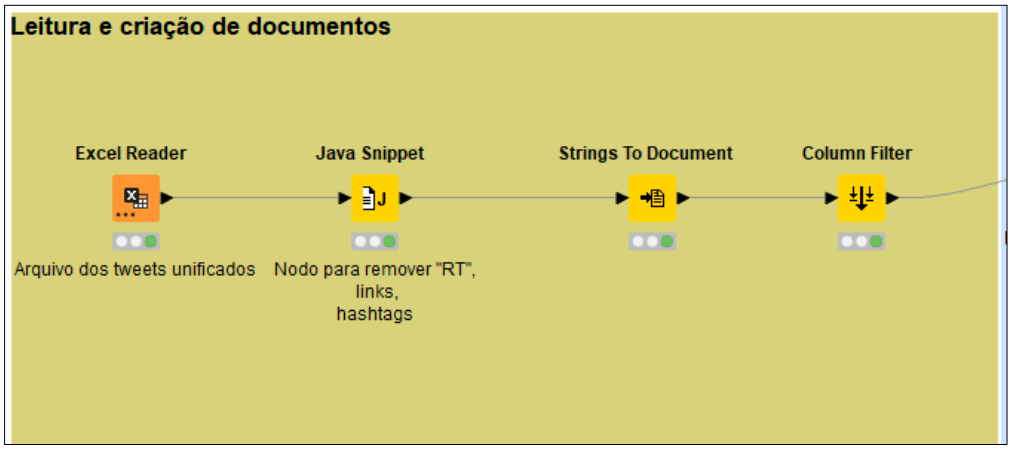

Figura 3 Pipeline Leitura de criação dos documentos

4. Análise dos Resultados

4.1. Análise Exploratória dos Dados

A análise exploratória dos dados foi realizada utilizando com o auxílio da ferramenta Jupyter.

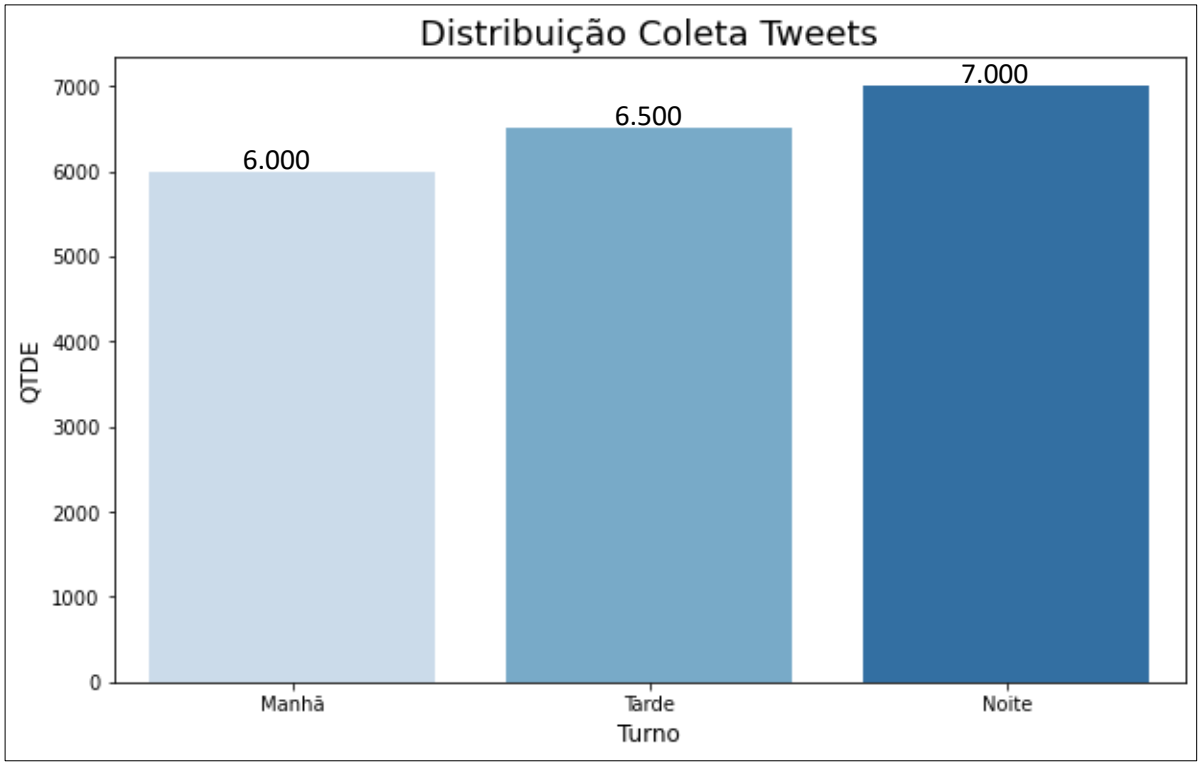

Figura 4 Gráfico Distribuição coleta

Dos coletados 19.500 tweets, 12.373 (63%) são retweets. A coleta do turno da tarde compreendeu o maior número de retweets. Neste turno, 71% (4.600) dos textos coletados foram retuitados pelos usuários.

A coleta realizada no turno da manhã foi a mais equilibrada dentre os turnos. Dos 6.000 tweets coletados, 46% (2.780) era originais e 54% (3.220) retweets.

Turno Coleta	**Retweet**				**Total**	
	Não		Sim			
	Qtde	%	Qtde	%	Qtde	%
Manhã	2.780	46%	3.220	54%	6.000	100%
Tarde	1.900	29%	4.600	71%	6.500	100%
Noite	2.447	35%	4.553	65%	7.000	100%
Total	**7.127**	**37%**	**12.373**	**63%**	**19.500**	**100%**

Tabela 1 Tabela cruzada retweet por turno

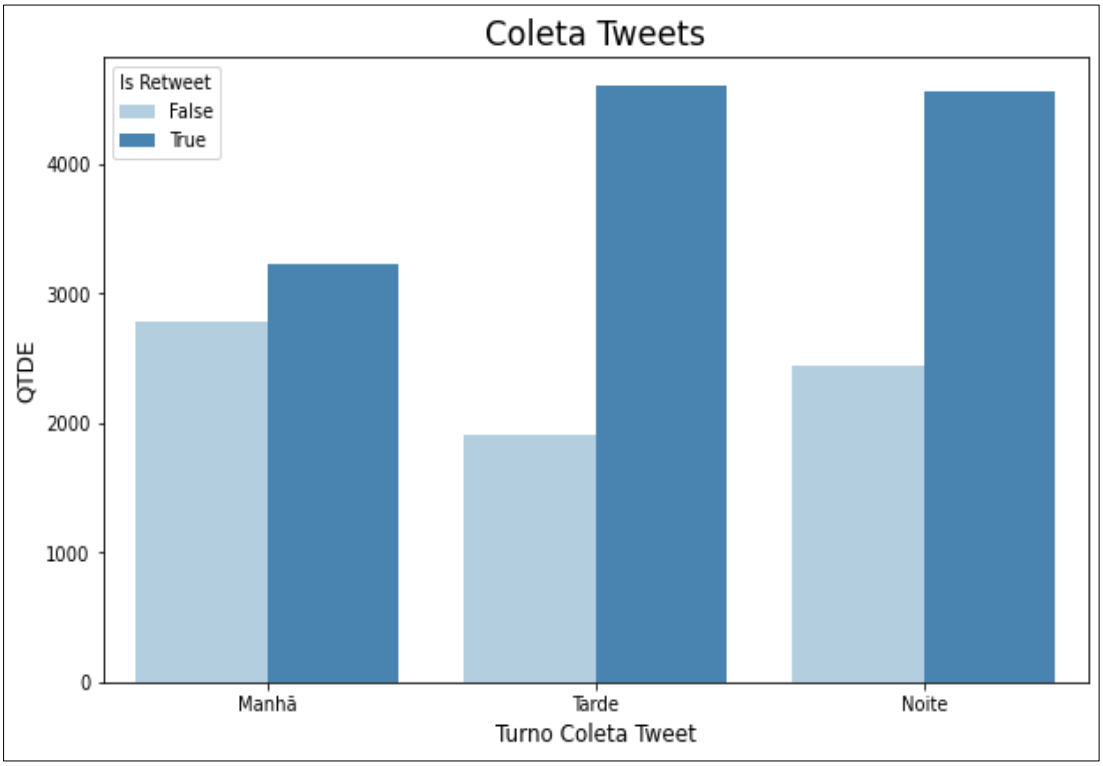

Figura 5 Distribuição da coleta de tweets

O tweet mais retuitado, compartilhado 1.056 vezes pelos usuários da rede no momento da coleta de dados, foi publicado originalmente pela conta oficial da Disney+ no Brasil. Seu conteúdo original é:

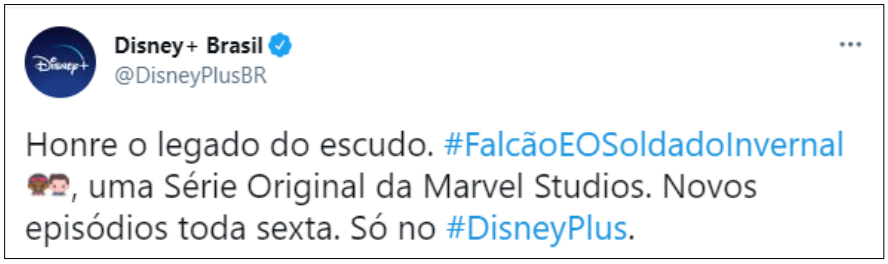

Figura 6 Tweet mais retuitado

4.2. Mineração de Texto

As etapas de pré-processamento da base, enriquecimento, transformação foram realizadas utilizando o software Knime. Por se tratar de tweets coletados em português, toda etapa de enriquecimento da base foi realiza utilizando dicionários externos em português.

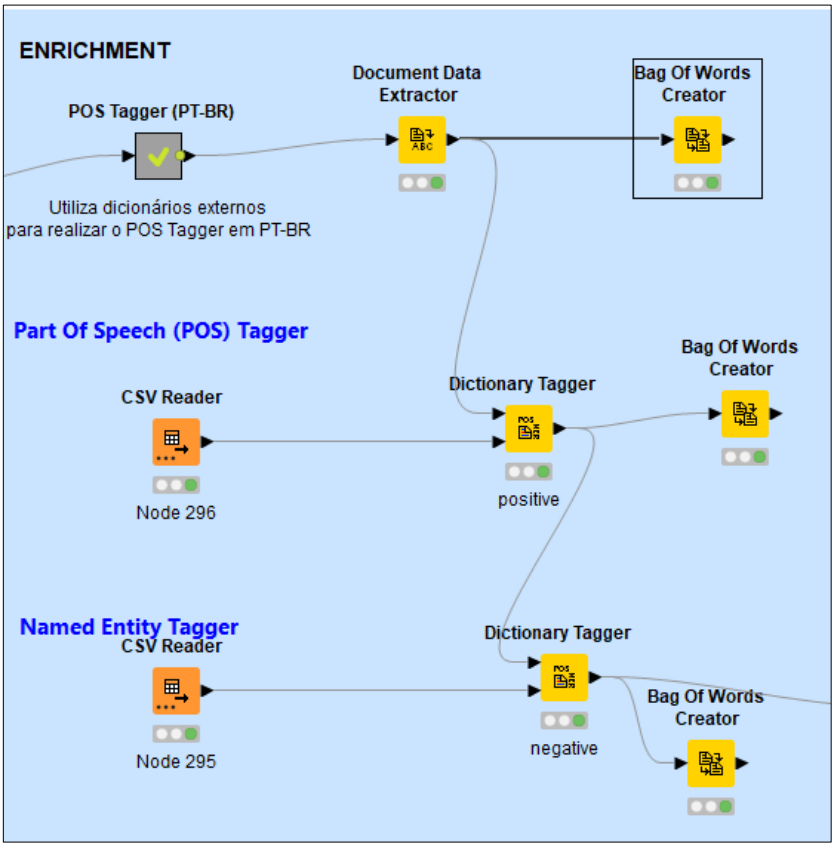

Figura 7 Enriquecimento dados

Após o enriquecimento da base, foram realizadas as etapas de processamento da base e transformação. As *stopwords* foram removidas de acordo com o dicionário em português.

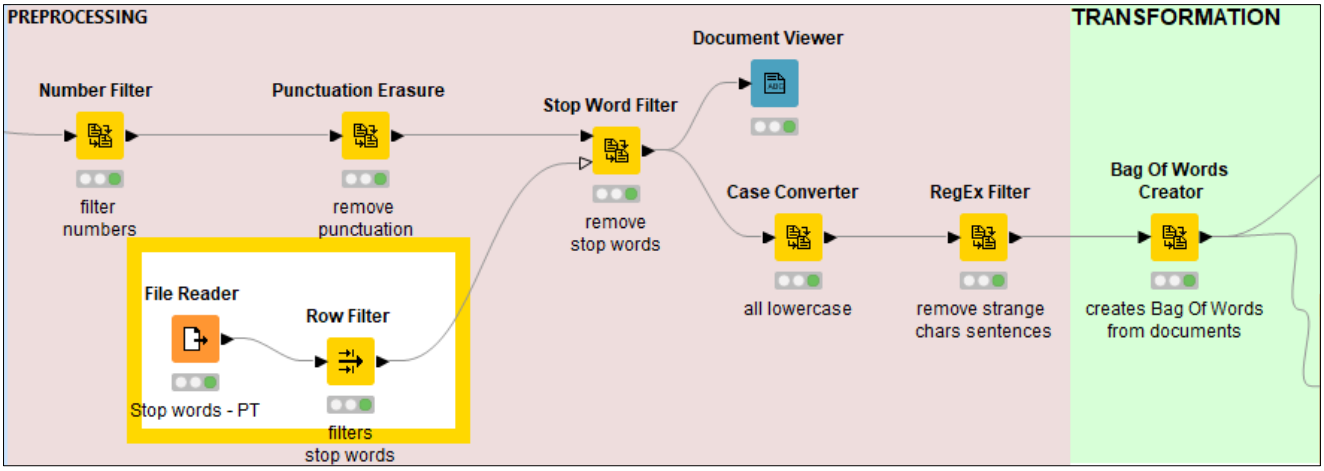

Figura 8 Processamento e transformação dos dados

Em sequência, foi calculada a frequência relativa dos termos do documento. Após essa etapa, foram realizados os filtros de forma que possibilitem visualizar os termos de acordo com o sentimento, seja ele positivo ou negativo. O sentimento neutro é dado ao termo que não foi definido em nenhuma das classes definidas anteriormente.

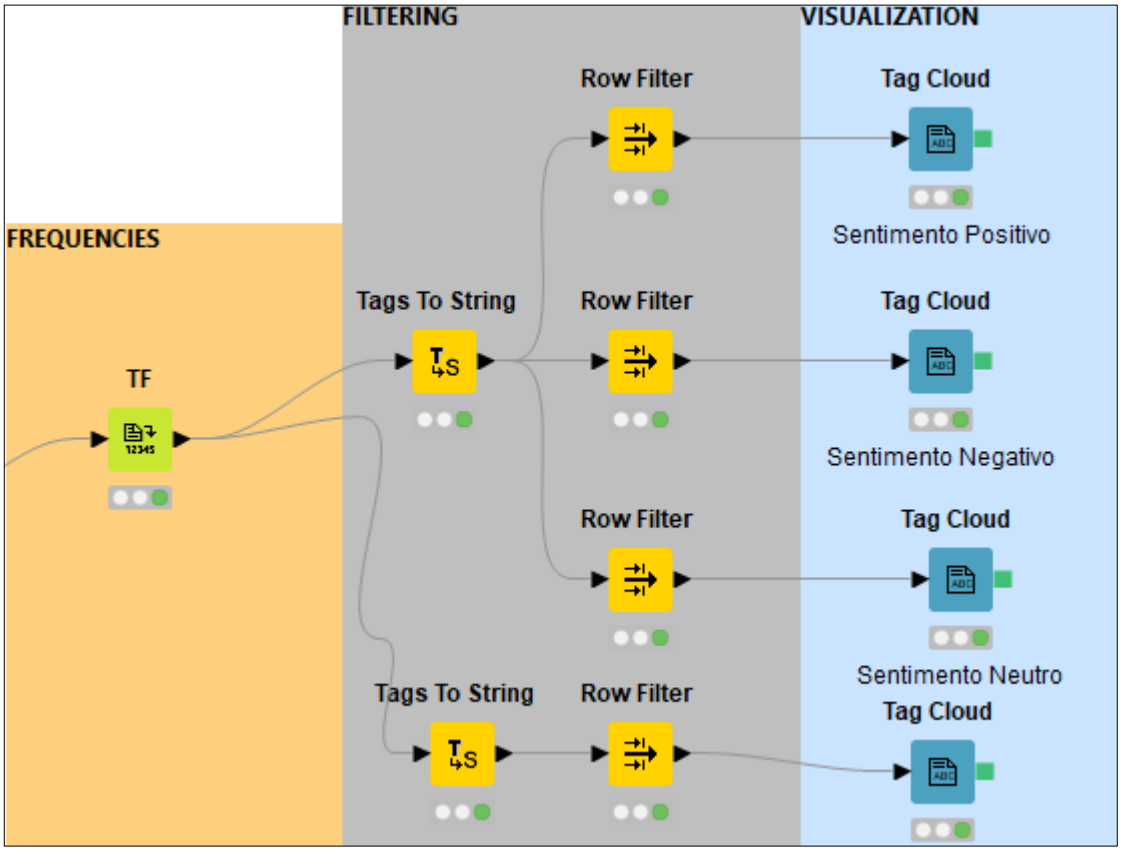

Figura 9 Fluxo de cálculo de frequência, filtragem e visualização

Uma vez filtradas, foi possível realizar a nuvem de palavras que foram classificadas como positivas. Destacam-se as palavras “importante”, “amizade”, “beleza”, “bom”, “amo”, “feliz”, “perfeito”. Na figura 10 é possível observar as demais palavras que compõem a nuvem de sentimentos positivos.

Figura 10 Nuvem de sentimentos positivos

A figura 11 apresenta as palavras que compõem a nuvem de sentimentos negativos. As palavras “Ansioso”, “defeito”, “comum”, “saudade”, se destacam.

Vale frisar a palavra “Ultimato”, que também aparece como sentimento negativo. Essa palavra se refere ao último filme da saga dos Vingadores, o encerramento da fase três do MCU. O filme foi bastante citado nos tweets coletados, uma vez que o universo da série *Falcão e o Soldado Invernal* se passa seis meses após os eventos finais de Ultimato. Uma hipótese para que essa palavra tenha se enquadrado como sentimento negativo é que, embora o filme seja considerado um sucesso pela crítica especializada, os tweets associados a ele estão ligados às palavras “drama”, “trauma” e contextos negativos relacionados à saga.

Figura 11 Nuvem de sentimentos negativas

A figura 12 apresenta as nuvens de palavras. Em vermelho, as que se relacionam aos sentimentos negativos e em verde, as palavras que se relacionam aos sentimentos positivos.

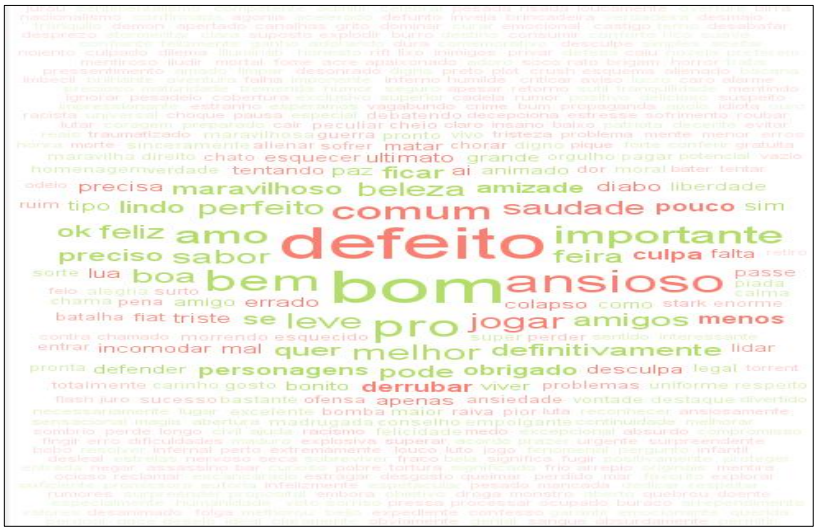

Figura 12 Nuvem de sentimentos

4.3. Bigrama – Análise de Palavras mais ditas em conjunto

Com o auxílio da biblioteca "Natural Language Toolkit" para Python, adicionalmente foi realizada uma análise das palavras mais ditas em conjunto. A tabela 2 apresenta o top das 20 palavras com o resultado dessa análise.

Palavra 1	Palavra 2	Quantidade de vezes ditas em conjunto
falcão	soldado	9.425
Soldado	invernal	9.172
Marvel	studios	1.534
19	março	1.410
Sam	wilson	1.103
#snydercut	sexta	1.082
Steve	rogers	980
#falcãoeosoldadoinvernal	marvel	909
Capitão	américa	881
Sexta	falcão	881
original	marvel	873
Série	original	873
Legado	escudo	816
Honre	legado	816
Vai	ser	810
invernal	estreia	785
#falcãoeosoldadoinvernal	#thefalconandthewintersoldier	783
Trailer	final	774
Estreia	exclusiva	766
exclusiva	19	753

Tabela 2 Bigrama - Top 20 palavras mais ditas

Em primeiro lugar os personagens título da série. É Interessante observar a recorrência da hashtag "#SnyderCut" associada à palavra sexta. Uma possível explicação para isso é o lançamento do filme "Liga da Justiça de Zack Snyder" na quinta-feira 18 de março, dia anterior à estreia da série *Falcão e o Soldado Invernal*.

Figura 13 Exemplo de tweet coletado

5. Conclusão

Nesse trabalho foi possível analisar o conteúdo textual dos tweets coletados no dia do lançamento da série *Falcão e o Soldado Invernal*. Foi interessante observar as palavras em destaque como sentimento positivo, tais como “amigos”, “amizade”, uma vez que a série aborda a relação entre os dois protagonistas. Outro fator interessante é a associação ao sentimento negativo relacionada ao filme *Vingadores – Ultimato*, como dito anteriormente.

Algumas limitações foram encontradas nesse trabalho. Em primeiro lugar, é interessante ressaltar que, embora a análise textual tenha avançado, em português a documentação limitada. Outro ponto que vale a atenção é que, diferentemente das séries de televisão, que normalmente são lançadas em algum dia específico da semana

e horário previamente definido, as séries exclusivas para streaming podem ser assistidas a qualquer momento, o que pode espaçar a audiência e consequentemente o fluxo de tweets. Esse é um fator que pode dificultar o melhor horário de *share*, ou seja, o horário em que realmente as pessoas estão assistindo ao episódio.

Para trabalhos futuros, sugere-se realizar a análise coletando os tweets durante o final de semana, essa avaliação tem sido adotada por alguns especialistas no que tange à quantidade de telespectadores que estão assistindo a algum episódio, por exemplo. Outra sugestão é a realização de um trabalho coletando informações durante toda a temporada. Assim, pode-se analisar o sentimento durante toda a temporada, podendo observar se há algum padrão de sentimento entre episódios.

6. Referências:

COMO RETWEETAR. *In*: **Como retweetar**. [*S. l.*], 2021. Disponível em: https://help.twitter.com/pt/using-twitter/how-to-retweet. Acesso em: 28 mar. 2021.

COZINHA, A. WandaVision se torna a série mais vista no mundo todo. *In*: COZINHA, A. **WandaVision se torna a série mais vista no mundo todo**. [*S. l.*], 13 fev. 2021. Disponível em: https://www.omelete.com.br/marvel-cinema/wandavision-marvel-serie-mais-vista-no-mundo. Acesso em: 28 mar. 2021.