UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GEOVANE DO NASCIMENTO SILVA

ANÁLISE DE CORRELAÇÃO ENTRE COMENTÁRIOS E CURTIDAS/DESCURTIDAS DE VÍDEOS DO YOUTUBE POR MEIO DE ANÁLISE DE SENTIMENTOS

CAMPINA GRANDE - PB

2021

GEOVANE DO NASCIMENTO SILVA

ANÁLISE DE CORRELAÇÃO ENTRE COMENTÁRIOS E CURTIDAS/DESCURTIDAS DE VÍDEOS DO YOUTUBE POR MEIO DE ANÁLISE DE SENTIMENTOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Eanes Torres Pereira.

CAMPINA GRANDE - PB

2021

S586a Silva, Geovane do Nascimento.
Análise de correlação entre comentários e curtidas/descurtidas de vídeos do You Tube por meio de análise de sentimentos. / Geovane do Nascimento Silva. – 2021.

12 f.

Orientador: Prof. Dr. Eanes Torres Pereira.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. You Tube. 2. Análise de sentimentos. 3. Curtidas e descurtidas – you tube. 4. Comentários – you tube. 5. Rede social – You Tube. 6. Análise de dados. 7. Aprendizagem de máquina. 8. Mídias sociais. I. Pereira, Eanes Torres. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

GEOVANE DO NASCIMENTO SILVA

ANÁLISE DE CORRELAÇÃO ENTRE COMENTÁRIOS E CURTIDAS/DESCURTIDAS DE VÍDEOS DO YOUTUBE POR MEIO DE ANÁLISE DE SENTIMENTOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Eanes Torres Pereira

Orientador – UASC/CEEI/UFCG

Professora Dr. Leandro Balby Marinho

Examinador – UASC/CEEI/UFCG

Professor Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 20 de outubro de 2021.

CAMPINA GRANDE - PB

ABSTRACT

Sentiment Analysis is a task to identify the sentiments expressed in certain texts, labeling them as positive, negative or neutral. The rise of social media raised this technique to an industrial level of great interest, since companies need to know if their marketing strategies have had the expected effect on the audience target. Focusing on demonstrating the use of sentiment analysis in textual data, more specifically in comments of social media, this work aims to carry out a data analysis and check if there is a correlation between the sentiments recognized in the comments and the amount of likes and dislikes of Youtube videos.

Análise de correlação entre comentários e curtidas/descurtidas de vídeos do Youtube por meio de Análise de Sentimentos

Geovane do Nascimento Silva
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande Campina Grande, Paraíba, Brasil
geovane.silva@ccc.ufcg.edu.br

Eanes Torres Pereira
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande Campina Grande, Paraíba, Brasil
eanes@computacao.ufcg.edu.br

RESUMO

A análise de sentimentos é uma tarefa para identificar os sentimentos expressos em determinados textos, rotulando-os como positivos, negativos ou neutros. A ascensão das mídias sociais elevou essa técnica a um patamar industrial de grande interesse, uma vez que as empresas necessitam saber se suas estratégias de marketing têm tido o efeito esperado no público alvo. Tendo em foco demonstrar o uso da análise de sentimento em dados textuais, mais especificamente em comentários de mídias sociais, o presente trabalho tem como objetivo realizar uma análise de dados e verificar se existe correlação entre os sentimentos reconhecidos nos comentários e a quantidade de curtidas e descurtidas de vídeos do Youtube.

Palavras-chave

Análise de sentimentos, Aprendizagem de Máquina, Mídias Sociais, Youtube, Análise de Dados.

1. INTRODUÇÃO

Com o crescente aumento e popularização de redes sociais, e também a facilidade de acesso à internet nas últimas 2 décadas, tem se tornado comum encontrar diversas modalidades de profissoẽs que surgiram junto com elas. Empresas cada dia mais têm visto o potencial desse público, que consome e gera uma quantidade massiva de conteúdo. É um mercado dinâmico, onde profissões surgem ao mesmo tempo que outras perdem força. Dentro de tantas profissões uma das que se destacam é a de produtores de conteúdo para o Youtube. Cada vídeo contém milhares de comentários, podendo chegar a milhões. Atualmente, o segundo vídeo[1] mais assistido da plataforma possui 4.127.961 de comentários[2].

Esse novo nicho logo chamou atenção de empresas que começaram a usar o Youtube como uma plataforma para divulgar suas campanhas de marketing. Com a consolidação da plataforma como um dos maiores sites de compartilhamento de vídeos, credibilidade não é um problema, o site é atualmente o 2º mais acessado da Internet, pertencente à família Google, conta com 721,9 milhões de visitas por dia e já provou ser responsável pelo lançamento de muitos sucessos da música, como é o caso do cantor sul-coreano Psy e seu hit *Gangnam Style*.

Muito além do ramo da música, os produtores de conteúdo do Youtube são das mais variadas áreas, sejam educativas, infanto-juvenil, influenciadores digitais, curiosidades, entre outros. Independente da área ou do tipo de vídeo, o engajamento do público que consome o conteúdo de um produtor é um fator crucial para definir a longevidade e popularidade do canal na rede, e também o sucesso ou insucesso de um determinado vídeo. Seja o vídeo uma campanha de marketing de uma multinacional, ou um vídeo de humor de um desconhecido, o engajamento é quem dirá o caminho que o produtor deve seguir para melhorar, críticas e *feedbacks* são obtidos de forma rápida e gratuita assim que o vídeo entra na web através do Youtube.

Um portfólio de dados de texto - comentários nos vídeos - tão vasto, torna-se uma ótima fonte para aplicações baseadas em análise de sentimentos. A análise de sentimento é uma técnica desenvolvida por profissionais que trabalham com ciência de dados para gerar informações e *insights* a partir de dados, normalmente textuais. Essa técnica é comumente utilizada para que empresas possam entender como seus clientes se sentem em relação aos seus produtos ou serviços, e com a proliferação de opiniões, comentários e avaliações, a imagem das empresas está cada vez mais exposta nas redes.

A análise de sentimentos é uma mineração contextual de um texto que identifica e extrai informações subjetivas no material de origem. Um sistema de análise de sentimentos para conteúdo textual combina o processamento de linguagem natural (PLN) e técnicas de aprendizado de máquina para atribuir pontuações ponderadas de sentimento às sentenças.

Nesse contexto, o objetivo deste trabalho é extrair informações relacionadas a sentimentos de comentários de vídeos do Youtube e analisar se existe relação entre o tipo predominante de comentários e a quantidade de curtidas e descurtidas do vídeo. Na plataforma de vídeos do Google existem dois botões que os usuários podem utilizar para expressar o seu gosto quanto ao vídeo, um com o dedo indicador para cima que indica que o usuário gostou do vídeo, o outro com o dedo indicador para baixo

[1] Most popular YouTube videos based on total global views as of August 2021. Disponível em <https://www.statista.com/statistics/249396/top-youtube-videos-views/>. Acesso em 4 de outubro de 2021 às 14h48.

[2] Luis Fonsi - Despacito ft. Daddy Yankee. Disponível em <https://www.youtube.com/watch?v=kJQP7kiw5Fk&ab_channel=LuisFonsiVEVO>. Acesso em 4 de outubro de 2021 às 14h48.

que indica que o espectador não gostou ou reprova o conteúdo. Neste trabalho chamaremos o botão com indicador para baixo de descurtir/descurtidas e o com indicador para cima de curtir/curtidas.

2. TRABALHOS RELACIONADOS

No campo de análise de sentimentos voltada principalmente para comentários de vídeos do Youtube pode-se mencionar Bhuiyan [1] que realizou um estudo para classificar buscas textuais de vídeos com base em processamento de linguagem natural (PLN), de forma a descobrir o vídeo mais relevante e popular para cada busca. Chang [2] também investigou a popularidade de vídeos do Youtube baseado na influência dos comentários, segundo ele um grande número de visualizações pode não resultar em comentários positivos, outros fatores podem ser mais relevantes. Há trabalhos também como o de Yang [3] que trata da relação entre a interação dos usuários e os padrões de consumo dos vídeos. Mais precisamente como os comentários afetam o tempo de exibição da mídia. Seu trabalho concluiu que vídeos com mais comentários positivos tendem a ter um tempo de exibição um pouco maior que os vídeos com mais comentários negativos.

3. FUNDAMENTAÇÃO TEÓRICA

Comentários de vídeos do Youtube tem se tornado cada vez mais relevantes para as comunidades de pesquisas. Siersdorfer [5] em seu estudo busca descobrir o quão úteis podem ser esses dados, onde análises da plataforma revelam uma grande quantidade de *feedback* da comunidade do Youtube por meio dos comentários. O estudo trabalhou com uma amostra de mais de 67 mil vídeos e mais de 6 milhões de comentários para verificar dependências entre variáveis como comentários, classificações de comentários, visualizações e categorias de tópicos, e seus resultados foram promissores.

Outro trabalho relevante com foco em comentários do segundo site mais acessado do mundo[3] também faz uma análise aprofundada dos comentários de usuários do Youtube, a pesquisa de Schultes [4] alerta que normalmente a imagem pública dos comentários é muito pobre, dado que os usuários geralmente esperam que a maioria dos comentários seja de pouco valor ou de mau gosto. O estudo propõe uma nova abordagem de classificação e com base na sua proposta conseguiram realizar análises semânticas de vídeos mais rapidamente e com menos exigências de poder computacional. Por fim, o trabalho também indica que as interações de *likes* e *dislikes* dos usuários nos vídeos são de fato influenciadas pela dispersão de comentários valiosos e inferiores.

3.1 Extração de sentimentos de textos

A análise de sentimentos é um processo para extrair opiniões que os usuários expressam em textos, essas opiniões podem ser classificadas em positivas, neutras ou negativas.

[3] Segundo o site https://www.alexa.com/topsites o Youtube é o segundo site mais acessado globalmente.

Figura 1. Possíveis classificações para análise se sentimento.

Fonte. Adaptado de [7]

3.2 TextBlob

O TextBlob[4] é uma biblioteca construída para o Processamento de Linguagem Natural (PLN), inicialmente desenvolvida por Steven Sloria[5]. A biblioteca foi construída a partir do *Natural Language ToolKit* (NLTK) para realizar a maior parte de suas tarefas.

Quando um texto é submetido ao algoritmo de análise de sentimentos do TextBlob, ele retorna valores numéricos para polaridade e subjetividade, que definem respectivamente o quanto um texto é positivo ou negativo, e o quanto o texto é objetivo ou subjetivo.

Para entender como exatamente o TextBlob calcula as características de subjetividade e polaridade, Fahad [7] e Schumacher [8] inspecionaram o código fonte de Sloria [9] indicando que a biblioteca usa como base para seus cálculos um conjunto de léxicos. A análise de sentimentos inicialmente foi inicialmente usando léxicos pré-desenvolvidos e construídos manualmente. No caso do TextBlob a base de léxicos contém para cada palavra informações da subjetividade, polaridade e intensidade.

Tabela 1. Características no léxico para a palavra *great*.

word	polarity	subjectivity	intensity
great	1,0	1,0	1,0
great	1,0	1,0	1,0
great	0,4	0,2	1,0
great	0,8	0,8	1,0

Fonte. Extraído de [8]

Na Tabela 1 podem ser observadas as características para a palavra *great*. Cada linha é uma entrada diferente no léxico da biblioteca. Conforme pode ser observado na Figura 2, o TextBlob ao calcular o sentimento para uma única palavra usa uma média das características das referências encontradas no léxico.

[4] TextBlob. Disponível em <https://textblob.readthedocs.io/en/dev/>. Acesso em 07 de outubro de 2021 às 20h06.
[5] Mais sobre Steven Sloria em https://stevenloria.com/.

Figura 2. Análise de sentimento para a palavra *great*.

```
TextBlob("great").sentiment
## Sentiment(polarity=0.8, subjectivity=0.75)
```

Fonte. Extraído de [8]

Figura 3. Análise de sentimento para a frase *not great*.

```
TextBlob("not great").sentiment
## Sentiment(polarity=-0.4, subjectivity=0.75)
```

Fonte. Extraído de [8]

Na Figura 3 pode-se observar que, quando é identificada uma negação, a biblioteca altera a polaridade e a subjetividade permanece igual. Segundo [8] a polaridade é multiplicada por -0,5.

Tabela 2. Características para a palavra *very.*

word	polarity	subjectivity	intensity
very	0,2	0,3	1,3

Fonte. Extraído de [8]

A intensidade por sua vez é um ponto muito interessante, ela pode ser interpretada como um modificador ou intensificador, normalmente no inglês palavras modificadores são adjetivos. Quando um modificador é identificado o TextBlob ignora a polaridade e a subjetividade e usa apenas os dados da intensidade.

Figura 4. Análise de sentimento para a frase *very great.*

```
TextBlob("very great").sentiment
## Sentiment(polarity=1.0, subjectivity=0.9750000000000001)
```

Fonte. Extraído de [8]

Na Figura 4 é analisado um caso de frase com um modificador, indicado na Tabela 2, a polaridade é 1,0 uma vez que a intensidade é maior que 1, no entanto o valor máximo para polaridade é de 1. Já a nova subjetividade é calculada como sendo $0,75 \times 1,3$.

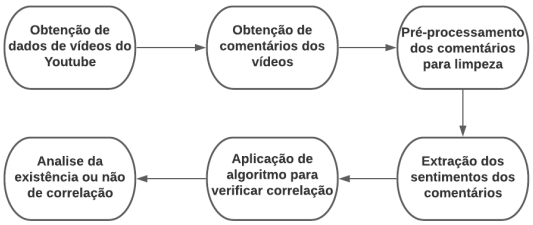
No caso de palavras que não estejam no léxico, o algoritmo apenas as ignora.

4. METODOLOGIA

A metodologia adotada consistiu na aquisição de base de dados, extração de características da base de dados por meio de algoritmo de análise de sentimentos, desenvolvimento de algoritmo para checagem da existência ou não da correlação.

Foram utilizadas bibliotecas de análise de sentimentos e aprendizagem de máquina para os processos de extração de sentimento, análise e comparação. O fluxo completo da metodologia pode ser observado na Figura 5.

Figura 5. Fluxograma da metodologia.



Fonte. Gerada pelo autor

4.1 Base de dados

Para o desenvolvimento do trabalho foi necessária a utilização de uma base de dados de comentários de vídeos do Youtube. O ideal seria uma base totalmente aleatória, no entanto, atualmente a API do Youtube não oferece um método que retorne dados de forma randômica. De acordo com os padrões seguidos pela comunidade, a forma mais próxima de alcançar uma aleatoriedade é a de coletar vídeos a partir de buscas por termos. Dessa forma, buscou-se um conjunto de termos na língua inglesa para realizar a busca dos vídeos, para tal, foi utilizada a ferramenta do Google Trends para encontrar uma série de termos para efetuar as buscas. Os termos utilizados podem ser encontrados na página de Pesquisas do ano 2020[6] da ferramenta do Google. Dentre as categorias possíveis, foram escolhidas as seguintes: *Searches, People, Games, Sports Teams, Tv Shows* e *Movies*. A escolha dos termos e vídeos em inglês se deu pelo fato de que algoritmos de análises de sentimentos em textos de língua inglesa são mais conceituados e mais facilmente encontrados. Para cada uma dessas categorias, foram selecionados 10 termos associados. Os termos podem ser apenas uma palavra ou um conjunto de palavras, por exemplo *Coronavirus update* e *Parasite*. Todos os termos e as categorias associadas podem ser observados na Tabela 3.

Tabela 3. Categorias e termos usados nas buscas na API do Youtube.

Searches	***People***	***Games***	***Sports Teams***	***Tv Shows***	***Movies***
Election results	*Joe Biden*	*Among Us*	*Boston Celtics*	*Tiger King*	*Parasite*
Coronavirus	*Kim Jong Un*	*Fall Guys: Ultimate Knockout*	*Miami Heat*	*Cobra Kai*	*1917*
Kobe Bryant	*Kamala Harris*	*Valorant*	*Kansas City Chiefs*	*Ozark*	*Black Panther*
Coronavirus update	*Jacob Blake*	*Genshin Impact*	*Los Angeles Clippers*	*The Umbrella Academy*	*Harley Quinn: Birds of Prey*
Coronavirus symptoms	*Ryan Newman*	*Ghost of Tsushima*	*Dallas Stars*	*The Queen's Gambit*	*Little Women*

[6] Pesquisas do ano no Google - Google Trends. Disponível em <https://trends.google.com.br/trends/yis/2020/US/>. Acesso em 12 de outubro de 2020 às 11h30.

Zoom	*Tom Hanks*	*Animal Crossing*	*Washington Football Team*	*Little Fires Everywhere*	*Just Mercy*
Who is winning the election	*Shakira*	*Assassin's Creed Valhalla*	*Philadelphia Flyers*	*Outer Banks*	*Bad Boys 3*
Naya Rivera	*Tom Brady*	*The Last of Us 2*	*Tampa Bay Buccaneers*	*Ratched*	*Sonic the Hedgehog*
Chadwick Boseman	*Kanye West*	*Madden NFL 21*	*Boston Bruins*	*All American*	*Contagion*
PlayStation 5	*Vanessa Bryant*	*Jackbox*	*San Francisco 49ers*	*The Last Dance*	*Fantasy Island*

Fonte.Gerada pelo autor

4.1.1 Obtenção de dados dos vídeos e comentários

Para cada um dos termos, foi executada uma busca na API do Youtube para recuperar informações de até 15 vídeos. As informações são referentes às estatísticas de cada vídeo. Quantidade de comentários, visualizações, curtidas, descurtidas e favoritos. E também o tópico/categoria do vídeo segundo os dados do Youtube. Como foram selecionados 10 termos para cada categoria indicada na Tabela 3, o esperado era que ao fim das requisições à API, fossem obtidos 900 vídeos. No entanto, nem todos os termos retornaram os 15 resultados, não que os termos não tenham nenhum resultado no Youtube, no entanto as buscas estavam demorando muito e foi implementado no algoritmo de extração uma condição para parar de coletar dados dos termos nessa condição. Inclusive, um dos termos não retornou nenhum resultado de acordo com o algoritmo. Após a execução da extração de informações foram obtidos 829 vídeos. Por fim, os vídeos foram filtrados para permanecer apenas aqueles que tinham pelo menos 30 comentários de acordo com as informações obtidas, depois dessa filtragem, permaneceram apenas 732 vídeos que formam a base de dados.

Para cada um dos vídeos filtrados foi executada uma requisição à API do Youtube para recuperar os comentários. Inicialmente, não foi definida uma quantidade máxima de comentários. No entanto, chegou um ponto em que verificou-se que alguns vídeos tinham uma quantidade muito grande de comentários e isso começou a prejudicar o andamento da obtenção dos dados. Neste momento, o algoritmo foi pausado e adaptado para recuperar um máximo de 5000 comentários para os próximos vídeos. Ao fim das requisições foram recuperados um total de 2.040.158 comentários dos 732 vídeos.

4.1.2 Pré-processamento

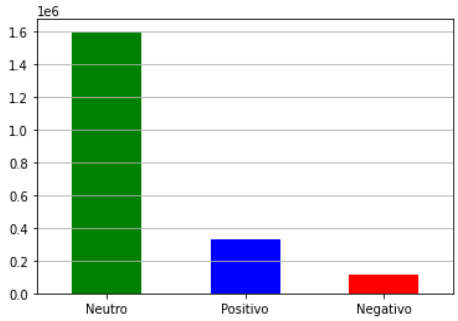
Para toda tarefa relacionada a NLP (*Natural Language Processing*), o recomendado é realizar uma etapa de pré-processamento dos dados para remover caracteres e termos indesejados. Seguindo os passos de Kulkarni e Shivananda [6], foi implementado um algoritmo envolvendo várias técnicas de pré-processamento, sendo elas, conversão do texto para *lowercase,* substituição de urls pelo termo URL e de *usernames* (palavras precedidas por @) por AT_USER, remoção de caracteres não alfanuméricos e de espaços adicionais, remoção de *hashtags*, remoção de números e de múltiplas exclamações, interrogações ou pontos finais e remoção de emoticons. Foi executada também a remoção de *stopwords* que são palavras sem muito significado ou pouco significado quando comparadas com outras palavras-chave, removendo essas palavras o foco estará sob as palavras mais importantes para o contexto do estudo. Para remoção dessas palavras consideradas irrelevantes, foi utilizada a lista de *stopwords* das bibliotecas nltk.corpus[7] e wordcloud[8] em conjunto com outras listas de *stopwords* encontradas na internet totalizando 888 palavras.

Além da remoção de *stopwords* e dos passos já citados, também foi realizada uma etapa de lematização dos termos. A lematização é um processo que extrai uma palavra raiz considerando o vocabulário. Por exemplo, "*good*", "*better*" e "*best*" são lematizados como *good*. Para essa etapa foi utilizado o módulo Word da biblioteca textblob. Após a etapa de pré processamento, foram descartados alguns comentários, que ficaram com dados nulos, restaram 1.916.597 comentários.

4.1.3 Análise de Sentimentos

Para extrair os sentimentos dos comentários, foi utilizada a biblioteca TextBlob, normalmente utilizada para processamento de textos, ela realiza operações de *Part-of-speech tagging*, análise de sentimento e outras tarefas de NLP. O método da biblioteca que extrai os sentimentos do texto retorna duas características, *polarity* e *subjectivity*. *Polarity* é um valor inteiro entre -1 e 1, quanto mais próximo de 1 o texto é considerado mais positivo, quando mais próximo de -1 é considerado mais negativo. Usando uma abordagem conservadora, visando garantir que os comentários estejam mais próximos do sentimento extraído, decidiu-se considerar como positivos apenas comentários que obtiveram um valor de *polarity* maior que 0,4 e negativos aqueles que tiveram polarity menor que -0,4. Na Figura 6, pode ser observada a distribuição dos comentários após a execução do algoritmo de análise de sentimentos.

Figura 6. Distribuição dos sentimentos dos comentários.



Fonte. Gerada pelo autor

[7] *Nltk Corpus Package.* Disponível em <https://www.nltk.org/api/nltk.corpus.html>. Acesso em 07 de outubro de 2021 às 20h11.

[8] *Wordcloud Package.* Disponível em <https://github.com/amueller/word_cloud>. Acesso em 07 de outubro de 2021 às 20h13.

Quanto aos números, 1.595.493 comentários foram classificados como Neutro, 333.300 (78,2%) foram classificados como Positivo (16,3%) e 111.365 (5,5%) como Negativo.

Nas Figuras 7 e 8, podem ser observadas as palavras mais comuns nos comentários positivos e negativos, respectivamente.

Figura 7. Nuvem de palavras dos comentários positivos.

Fonte. Gerada pelo autor

Figura 8. Nuvem de palavras dos comentários negativos.

Fonte. Gerada pelo autor

4.1.4 Testes de Hipóteses

Depois de aplicadas as técnicas para obtenção dos dados, limpeza e análise de sentimentos, a pergunta de pesquisa a ser respondida era "O sentimento dos comentários têm uma relação com a quantidade de *likes/dislikes* do vídeo?". Para responder essa pergunta, utilizou-se de técnicas de análise de dados com foco em verificar a existência ou não de correlação entre características já obtidas. Inicialmente, trabalhou-se com as seguintes hipóteses:

- **Hipótese nula**: os sentimentos dos comentários não têm relação com a quantidade de *likes/dislikes*.
- **Hipótese Alternativa**: Sim. O sentimento do comentário está relacionado com quantidade de *likes/dislikes*.

4.1.5 Abordagem Proposta

Em um primeiro momento, foram calculadas as correlações entre as seguintes características: *likeCount*, *dislikeCount*, *positive* e *negative* que representam respectivamente a quantidade de *likes,* a quantidade de *dislikes,* a quantidade de comentários positivos e a quantidade de comentários negativos dos vídeos. Pode-se observar na Tabela 4, as correlações.

Tabela 4. Correlações entre as variáveis analisadas.

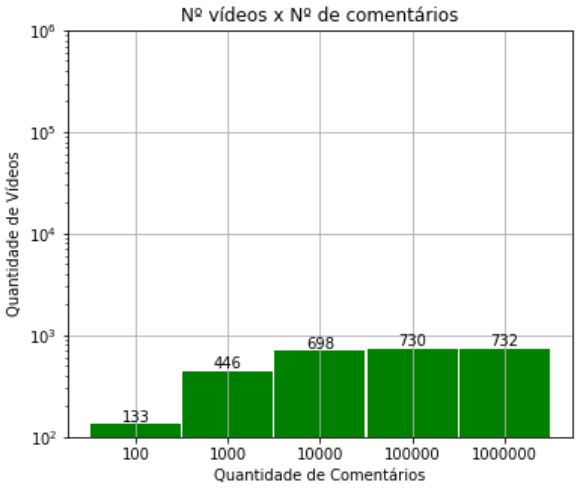
	likeCount	*dislikeCount*	*positive*	*negative*
likeCount	1,000000	0,483155	0,786170	0,746893
dislikeCount	0,483155	1,000000	0,388407	0,629635
positive	0,786170	0,388407	1,000000	0,824901
negative	0,746893	0,629635	0,824901	1,000000

Fonte. Gerada pelo autor

Como observa-se na Tabela 4, as correlações entre *likeCount* e *positive* e *dislikeCount* são fortes, o que era esperado e foi bastante animador. Em contrapartida, a correlação entre *positive* e *negative* também foi forte, o que não era esperado empiricamente, pois não faz muito sentido que em sua maioria os vídeos tivessem uma correlação forte entre essas características.

Para investigar os motivos da correlação forte entre *positive* e *negative*, foi refeita uma análise dos resultados da análise de sentimentos e também da base de dados. Com uma observação mais apurada da base de dados, foi identificado que existiam alguns vídeos com uma quantidade muito elevada de comentários e outros com uma quantidade bastante inferior. Na Figura 9, pode ser observado uma distribuição da quantidade de vídeos por quantidade de comentários de forma cumulativa. O gráfico da Figura 9 foi plotado com uma escala logarítmica devido aos altos valores de algumas colunas.

Figura 9.Quantidade de vídeos por quantidade de comentários.



Fonte.Gerada pelo autor

Se observarmos, por exemplo, a diferença entre a quantidade de vídeos das colunas 10.000 e 100.000, temos pelo menos 32 vídeos com uma quantidade de comentários muito superior se comparado com as colunas anteriores, já entre as colunas de 10 mil e 100 mil

comentários a diferença é de 252 vídeos e 2 dos 732 vídeos tem mais de 100 mil comentários. Dessa forma, pode-se verificar um desbalanceamento da base.

Diante dessa situação, decidiu-se extrair amostras de vídeos da base de dados com uma quantidade fixa de comentários. Nessa abordagem de observar os subconjuntos de comentários de vídeos que tivessem uma quantidade mínima de comentários, foram extraídas 3 amostras aleatórias para grupos de vídeos com 500, 1.000, 1.500, 2.000, 2.500, 3.000, 3.500, 4.000, 4.500 comentários. Dessa forma a correlação seria calculada apenas com a mesma quantidade de comentários para cada vídeo da amostra.

Nas Tabelas 5, 6 e 7 podem ser observadas as correlações calculadas para 3 amostras aleatórias contendo os dados apenas de vídeos que continham pelo menos 500 comentários, ou seja, cada amostra contém no máximo 500 comentários, mantendo assim uma amostra balanceada. Podemos observar que a correlação entre comentários positivos e negativos ficou da forma que se esperava. Já a correlação entre *likeCount* e comentários positivos e negativos é desprezível, da mesma forma que a correlação entre *dislikeCount* e comentários positivos e negativos. No entanto, vale ressaltar que a correlação entre *likeCount* e *positive* é ligeiramente maior que *likeCount* e *negative*, o mesmo vale para *dislikeCount* e *negative,* que por sua vez é maior que *dislikeCount* e *positive*.

Tabela 5. Correlações entre as variáveis analisadas para a amostra de 500 comentários número 1.

	likeCount	*dislikeCount*	*positive*	*negative*
likeCount	1,000000	0,468845	-0,041448	-0,071809
dislikeCount	0,468845	1,000000	-0,089296	0,124172
positive	-0,041448	-0,089296	1,000000	-0,484805
negative	-0,071809	0,124172	-0,484805	1,000000

Fonte. Gerada pelo autor.

Tabela 6. Correlações entre as variáveis analisadas para a amostra de 500 comentários número 2.

	likeCount	*dislikeCount*	*positive*	*negative*
likeCount	1,000000	0,468845	-0,009736	-0,081541
dislikeCount	0,468845	1,000000	-0,063238	0,101401
positive	-0,009736	-0,063238	1,000000	-0,491974
negative	-0,081541	0,101401	-0,491974	1,000000

Fonte. Gerada pelo autor.

Tabela 7. Correlações entre as variáveis analisadas para a amostra de 500 comentários número 3.

	likeCount	*dislikeCount*	*positive*	*negative*
likeCount	1,000000	0,468845	-0,002602	-0,069960
dislikeCount	0,468845	1,000000	-0,057067	0,145579
positive	-0,002602	-0,057067	1,000000	-0,488552
negative	-0,069960	0,145579	-0,488552	1,000000

Fonte. Gerada pelo autor.

Observando as outras amostras com quantidades maiores de comentários, o padrão também foi observado. Diante dessa situação foi aplicada uma verificação por meio de intervalos de confiança para ter uma melhor avaliação.

Para aplicar o intervalo de confiança, foram selecionadas 100 amostras com 500 comentários de cada vídeo (aqueles que continham pelo menos 500 comentários).

Na Tabela 8, podemos observar o um *screenshot* das correlações de cada amostra selecionada.

Tabela 8. *Screenshot* do *data frame* com as correlações das 100 amostras analisadas.

	sample	LP	LN	DP	DN	PN
0	0	-0.008568	-0.067111	-0.083121	0.147390	-0.493273
1	1	-0.007465	-0.110271	-0.058959	0.153859	-0.472676
2	2	-0.000069	-0.072958	-0.058398	0.140831	-0.476540
3	3	-0.011694	-0.076340	-0.062433	0.123807	-0.471374
4	4	-0.022301	-0.081101	-0.070632	0.109821	-0.473741
...	...	...	...	...	...	...
95	95	-0.008714	-0.050960	-0.056795	0.143183	-0.455855
96	96	-0.012804	-0.070681	-0.062965	0.154176	-0.481851
97	97	0.011630	-0.078514	-0.049131	0.148875	-0.501066
98	98	-0.000780	-0.095165	-0.055109	0.104998	-0.499260
99	99	-0.001495	-0.063090	-0.063324	0.166299	-0.490654

Fonte.Gerada pelo autor

Na Tabela 8, as colunas representam as correlações entre as características e foram abreviadas para uma melhor visualização, a descrição de cada uma delas pode ser observada na Tabela 9.

Tabela 9. Descrição das colunas da Tabela 8.

Termo	**Descrição**
Sample	Numero da amostra
LP	Correlação entre *likeCount* e comentários positivos
LN	Correlação entre *likeCount* e comentários negativos
DP	Correlação entre *dislikeCount* e comentários positivos

DN	Correlação entre *dislikeCount* e comentários positivos
PN	Correlação entre comentários positivos e negativos

Fonte.Gerada pelo autor

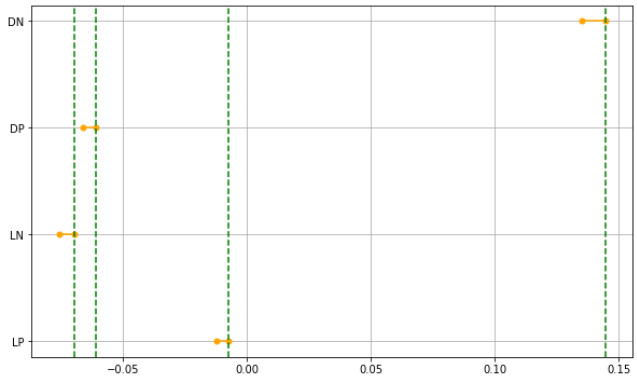
Os resultados da análise por meio de intervalos de confiança foram satisfatórios e podem ser observados na seção de resultados.

Após realizados todos os experimentos, foi possível rejeitar a hipótese nula, pois como apresentado, foram identificados vestígios significativos de correlação entre tipo de comentário e interação do usuário de gostar ou não gostar do vídeo. Mais detalhes sobre podem ser observados nos resultados.

5. RESULTADOS

Na Figura 10, podemos observar o gráfico com os intervalos de confiança para as variáveis DN, DP, LN, LP (indicadas na Tabela 9). Pode-se verificar que nenhum dos intervalos se sobrepõe, indicando que cada uma das correlações são variáveis únicas. Além disso, essa não intersecção entre os intervalos indica a existência de uma significância estatística de modo que podemos afirmar que LP > LN e DN > DP.

Figura 10. Gráfico com plotagem dos intervalos de confiança.



Fonte.Gerada pelo autor

.**Tabela 10**. Detalhamentos dos intervalos de confiança plotados na Figura 10.

Correlação	Limite Inferior	Limite Superior	Média
PN	-0,482476	-0,476578	-0,479527
LP	-0,012207	-0,007534	-0,009870
LN	-0,075674	-0,069682	-0,072678
DP	-0,065862	-0,060943	-0,063402
DN	0,135022	0,144697	0,139859

Fonte.Gerada pelo autor

Podemos observar na Tabela 10 os valores utilizados na aplicação dos intervalos de confiança, onde o valor médio da correlação PN (comentários positivos e negativos) é um valor considerado uma correlação média e negativa, o que indica que as duas variáveis tendem a se afastar, ou seja quanto mais comentários positivos, menos comentários negativos, ou vice-versa.

Para a correlação LP (*likeCount* e comentários positivos) o valor médio é uma correlação muito fraca, quase desprezível, mas ainda assim comparando com LN é maior, que por sua vez também tem um valor médio muito baixo. Isso indica que a quantidade de likes têm uma relação maior com a quantidade de comentários positivos do que negativos.

Observando a correlação DP e DN identificamos um valor considerável para DN e que por sua vez é maior que DP, ou seja a quantidade de *dislikes* de um vídeo tende a ter uma relação com a quantidade de comentários negativos.

Considerando as correlações LP e DN podemos concluir que quando um usuário não gosta do conteúdo de um vídeo, isso tende a se refletir com maior intensidade nos comentários do vídeo.

6. CONSIDERAÇÕES FINAIS

Quanto às contribuições, confirmou-se o que Schultes [4] diz que em sua maioria os comentários do Youtube não têm muito valor em relação a análise de sentimentos dado que a grande maioria dos comentários tratados neste trabalho foram classificados como neutros.

6.1 Limitações

As maiores dificuldades encontradas pode-se dizer que se deram por conta de uma base de dados desbalanceada. Poderia ter sido feito um esforço maior na coleta de dados para se ter uma base com a mesma quantidade de comentários para cada vídeo. Para trabalhos futuros, pode-se pensar na elaboração de uma base de dados equilibrada tanto na quantidade de comentários quanto na distribuição do tipo de comentários. Uma base conceituada ajudaria bastante outros trabalhos de análise de dados e sentimentos de comentários de vídeos do Youtube.

7. REFERENCIAS

[1] H. Bhuiyan, J. Ara, R. Bardhan and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017, pp. 474-478, doi: 10.1109/ICSIPA.2017.8120658.

[2] W. Chang, "Will Sentiments in Comments Influence Online Video Popularity?," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3644-3646, doi: 10.1109/BigData.2018.8621938.

[3] R. Yang, S. Singh, P. Cao, E. Chi and B. Fu, "Video Watch Time and Comment Sentiment: Experiences from YouTube," 2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb), 2016, pp. 26-28, doi: 10.1109/HotWeb.2016.13.

[4] Schultes, Peter; Dorner, Verena; and Lehner, Franz, "Leave a Comment! An In-Depth Analysis of User Comments on

YouTube" (2013). Wirtschaftsinformatik Proceedings 2013. 42.

[5] Siersdorfer, Stefan & Chelaru, Sergiu & Nejdl, Wolfgang & San Pedro, Jose. (2010). How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. Proceedings of the 19th International Conference on World Wide Web, WWW '10. 891-900. 10.1145/1772690.1772781.

[6] Akshay Kulkarni and Adarsha Shivananda. 2019. Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python (1st. ed.). Apress, USA.

[7] Fahad, Abdul Hafeez. Sentiment Analysis — Let TextBlob do all the Work!. 2021. Disponível em: <https://medium.com/red-buffer/sentiment-analysis-let-textblob-do-all-the-work-9927d803d137>. Acesso em 27 de setembro de 2021.

[8] Schumacher, Aaron. TextBlob Sentiment: Calculating Polarity and Subjectivity. 2015. Disponível em: <https://planspace.org/20150607-textblob_sentiment/>. Acesso em 27 de setembro de 2021.

[9] Steven Sloria. TextBlob: Simplified Text Processing. Disponível em : <https://textblob.readthedocs.io/en/dev/>. Acesso em 27 de setembro de 2021.

About the authors:

Geovane do Nascimento Silva. Graduando em Ciência da Computação, com 3 anos de experiência em Desenvolvimento de Software. Atualmente, é Desenvolvedor Junior na empresa Elife trabalhando como Dev Backend na parte de criação de chatbots.

Eanes Torres Pereira. Professor orientador.