

UNIVERSIDADE FEEVALE

DIEGO GIROTTO ARIGONY

DESENVOLVIMENTO DE UM CLASSIFICADOR DE CLICKBAIT  
BASEADO EM MACHINE LEARNING

(Título Provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo  
2020

DIEGO GIROTTI ARIGONY

DESENVOLVIMENTO DE UM CLASSIFICADOR DE CLICKBAIT  
BASEADO EM MACHINE LEARNING

(Título Provisório)

Anteprojeto de Trabalho de Conclusão de Curso,  
apresentado como requisito parcial  
à obtenção do grau de Bacharel em  
Ciência da Computação pela  
Universidade Feevale

Orientador: Rodrigo Rafael Villarreal Goulart

Novo Hamburgo  
2020

## RESUMO

Clickbait se refere a chamadas curtas de notícia formuladas com o objetivo de induzir os leitores a clicar em um link. O objetivo dessa prática é gerar maior tráfego de acessos a páginas de um site e, conseqüentemente, aumentar a renda a partir de publicidades na página. Isto se faz por meio de mensagens que podem ser vagas e informações exageradas ou errôneas, provocando curiosidade suficiente para o acesso. Em geral, o resultado desta prática vem ao custo de clareza e objetividade na informação. A Bauhaus-Universität Weimar propôs, em 2017, um desafio para o desenvolvimento de um classificador que define a matéria em uma das quatro classes: não clickbait, levemente clickbait, consideravelmente clickbait ou fortemente clickbait. Baseado neste desafio, o presente trabalho busca avaliar o estado da arte na área e aplicá-lo no problema sugerido.

Palavras-chave: classificadores. machine learning. clickbait.

## SUMÁRIO

MOTIVAÇÃO .....	5
OBJETIVOS .....	8
METODOLOGIA .....	9
CRONOGRAMA .....	10
BIBLIOGRAFIA .....	11

## MOTIVAÇÃO

Atualmente, a internet é uma fonte de conhecimento dos mais variados assuntos. Oferecendo acesso fácil a blogs pessoais, notícias nacionais e internacionais, conteúdo técnico dos mais diversos assuntos e fóruns de discussão, a oferta de informações é praticamente inesgotável.

Na atualidade o jornalismo, por exemplo, está em grande parte no meio digital. Dito isso, esse tipo de empresa depende diretamente da renda gerada em seus meios digitais. Com a crescente oferta de informações, é de se esperar uma enorme concorrência pela atenção dos leitores na rede. Buscando aumentar métricas importantes para a receita de publicidades, como quantidade de acessos e tempo de visitaç o, diversos sites vem usando a estrat gia de chamadas *clickbait*.

Clickbait, segundo Potthast et al. (2016), se refere a um certo tipo de propaganda de cont duo web formuladas com o objetivo de induzir o leitor a clicar em seu link. Um resultado imediato dessa pr tica   que, geralmente, a formula o   realizada   custa de objetividade e clareza na informa o da chamada. Se faz por meio de mensagens que podem ser vagas e informa es exageradas ou err neas, provocando curiosidade suficiente para o acesso. Exemplos de chamadas clickbait conforme corpus de Potthast et al. (2018):

- Chinese actress sued for staring too intensely
- As a BASE jumper leaped, his girlfriend snapped photos. Then came confusion, hope and despair
- 10 books to read before you see the movies this summer
- Someone put 1,000 cardboard cutouts of Nicki Minaj's ass on the steps of a Finland Cathedral
- 10 years ago, production errors at the Airbus Group factory crippled the company. Today, it's a model of efficiency:
- Homem tenta abra ar le o, voc  n o vai acreditar no que aconteceu.

Essa prática, além de prejudicial a todo o sistema de notícias online, pode gerar sérios problemas econômicos. Exemplo disso é o caso da Apple (HARGREAVES, 2008), quando chamadas de notícias davam a entender que Steve Jobs, CEO da empresa, havia sofrido um infarto. As ações da empresa despencaram cerca de 10% em valor na bolsa americana.

Na literatura, conforme Geçkil et al. (2020, p. 199), o problema da classificação de notícias clickbait tem sido abordado, no geral, com o uso de métodos de processamento de linguagem natural. Potthast et al (2016) realizou um dos primeiros e mais conhecidos trabalhos na área, propondo um modelo para a detecção clickbait baseado em 215 atributos, possibilitando seu classificador *random forest* atingir a marca de 0.79 ROC-AUC, 0,76 de precisão e 0,76 de recall. O artigo também é responsável pelo primeiro corpus clickbait, com um total 2992 tweets provenientes dos 20 perfis dos que mais geram conteúdo de notícias no Twitter, baseado pela influência em termos de re-tweets. Posteriormente, conforme processo descrito por Potthast et al. (2018), uma evolução do corpus foi realizada, desta vez contendo 38,517 tweets anotados. Denominado Webis Clickbait Corpus 2017.

Utilizando o Webis Clickbait Corpus 2017, o Clickbait Challenge, uma competição organizada em 2017, incentivou os grupos de participantes a criar um modelo para determinar clickbaits. Foram submetidos 16 resultados, onde 10 possuíam o formato de artigo e 8 incluíram também o código-fonte de seu modelo. O estudo deste problema não é algo novo. Contudo, o crescimento da prática com a migração da informação para o plano digital tem impulsionado a pesquisa na área.

Baseado nessa problemática, o presente trabalho busca aprimorar técnicas de classificação utilizada nos modelos submetidos ao Clickbait Challenge. Dentre oportunidades encontradas, destaco o uso aprofundado de paragraph vector - Doc2Vec, um conceito apresentado Le e Mikolov (2014) para criar uma representação numérica de um documento. A técnica é uma alternativa ao *bag-of-words*, técnica comumente utilizada em algoritmos de processamento de linguagem natural.

Glenski et al. (2017), em artigo de resultado submetido ao Clickbait Challenge, sugere o uso de Doc2Vec na busca de similaridade entre texto da chamada (tweet) e título do artigo como uma forma de capturar a relação entre o tweet e o artigo. O conceito utilizado no modelo Doc2Vec, similar ao Word2Vec (MIKOLOV et al., 2013), consiste em um algoritmo de aprendizagem não supervisionada para gerar um vetor representativo de partes de tamanho variado de texto, como sentenças, parágrafos e documentos. Em posse desse vetor, é possível aferir a similaridade entre essas partes. O presente estudo visa desenvolver um modelo de classificação que utilize Doc2Vec, sendo treinado e avaliado de acordo com o Clickbait Challenge.

## OBJETIVOS

### Objetivo geral

O presente estudo tem por objetivo desenvolver um classificador de notícias clickbait utilizando conceitos de Doc2Vec em sua implementação.

### Objetivos específicos

- Pesquisa sobre métodos de aprendizado de máquina adequados ao dataset disponibilizado no Clickbait Challenge.
- Pesquisa sobre a aplicação do conceito Doc2Vec na classificação.
- Desenvolver um modelo de classificador com base nas pesquisas realizadas.
- Avaliar os resultados do classificador desenvolvido e comparar aos resultados obtidos nos trabalhos relacionados.
- Apresentar os resultados obtidos.

## METODOLOGIA

Este trabalho tem como objetivo desenvolver um classificador para a resolução do problema de identificação de notícias clickbait. Pode-se defini-lo, portanto, como um projeto de pesquisa de natureza aplicada, pois conforme Prodanov e Freitas (2013, p. 51): "[...] objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Envolve verdades e interesses locais."

Quanto a classificação de seus procedimentos técnicos, pode ser enquadrado como bibliográfico e experimental. É bibliográfico dada a necessidade de pesquisa acerca do material já publicado, buscando informação relevante em relação ao tema. Também é experimental, pois vai gerar experimentos procurando melhorias em comparação ao resultado de trabalhos anteriores.

A abordagem utilizada é quantitativa. A análise dos resultados obtidos nos experimentos será feita tendo em vista as mesmas métricas utilizadas no Clickbait Challenge: erro médio quadrático, F1 Score, predição, recall e acurácia.

Por fim, em relação ao objetivo do projeto, pode ser definido como exploratório. A pesquisa e desenvolvimento tem o intuito de validar a hipótese de Doc2Vec como uma alternativa de modelo de classificação para o problema destacado no Clickbait Challenge e enriquecer o estudo na área da classificação de notícias clickbait.

## CRONOGRAMA

### Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Mai	Jun
Desenvolvimento do anteprojeto	x	x		
Pesquisa e avaliação de trabalhos relacionados		x	x	x
Formulação da proposta de modelo		x	x	x
Testes iniciais técnicas de classificação			x	x
Redação do Relatório Final		x	x	x
Entrega do Relatório Final				x

### Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Desenvolvimento do modelo	x	x		
Avaliação do modelo desenvolvido		x	x	x
Documentação dos resultados obtidos na avaliação		x	x	x
Redação do Relatório Final	x	x	x	x
Entrega Relatório Final				x
Apresentação a Banca				x

## BIBLIOGRAFIA

GEÇKIL, Ayşe et al. Detecting Clickbait on Online News Sites. In: PUTTING Social Media and Networking Data in Practice for Education, Planning, Prediction and Recommendation. 2020. p. 199-211.

GLENSKI, Maria et al. Fishing for clickbaits in social images and texts with linguistically infused neural network models, 2017.

HARGREAVES, Steve. Apple's stock hit by Web rumor: Online report that CEO Steve Jobs suffered a heart attack was false. SEC said to be investigating.. CNNMoney.com, 2008. Disponível em: <https://money.cnn.com/2008/10/03/technology/apple/>. Acesso em: 4 abr. 2020.

LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In: International Conference on Machine Learning, p. 1188-1196, 2014.

MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space. 2013.

POTTHAST, Martin et al. Clickbait detection. European Conference on Information Retrieval, p. 810-817, 2016.

POTTHAST, Martin et al. Crowdsourcing a large corpus of clickbait on twitter. Proceedings of the 27th International Conference on Computational Linguistics, p. 1498-1507, 2018.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. Metodologia do trabalho científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2. ed. Novo Hamburgo: Universidade Feevale, 2013. 276 p.