**UNIVERSIDADE FEDERAL FLUMINENSE**

**INSTITUTO DE COMPUTAÇÃO**

**DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO**

**PHELIPE GONÇALVES MARTINS**

**ANÁLISE EXPERIMENTAL COMPARATIVA DE ABORDAGENS TRADICIONAL E BASEADA EM DEEP LEARNING PARA A IDENTIFICAÇÃO DE FAKE NEWS**

**NITERÓI – RJ**

**2018**

**PHELIPE GONÇALVES MARTINS**

**ANÁLISE EXPERIMENTAL COMPARATIVA DE ABORDAGENS TRADICIONAL E BASEADA EM DEEP LEARNING PARA A IDENTIFICAÇÃO DE FAKE NEWS**

Monografia apresentada ao Departamento de Ciência da Computação da Universidade Federal Fluminense, como requisito para

obtenção do Grau de Bacharel em Ciência da Computação.

Orientador: Prof. José Viterbo Filho

**NITERÓI – RJ**

**2018**

**AGRADECIMENTOS**

Lorem ipsu.

Conteúdo

[**1.** **INTRODUÇÃO** 1](#_Toc529900130)

[**1.1.** **DEFINIÇÃO DO PROBLEMA** 2](#_Toc529900131)

[**1.2.** **OBJETIVOS** 3](#_Toc529900132)

1. **INTRODUÇÃO**

Ao longo do século XXI, o número de usuários que utilizam a internet tem crescido ano após ano. Um de seus principais usos é a busca por informação, seja através de redes sociais, sites de buscas, blogs, *podcasts* ou qualquer outro meio no qual se consiga conhecimento. À medida que os números têm crescido, um fenômeno social tem ganhado destaque no mundo todo. Tal fenômeno é conhecido como *Fake News*.

A disseminação de notícias falsas não é uma novidade, porém com a popularização das redes sociais, o número de divulgação desse tipo de notícia, que é potencializado através do compartilhamento da mesma pelos usuários, aumentou drasticamente. As *Fake News* podem ocorrer de diversas formas, como através de mentiras, boatos, informações absurdas com o intuito de aumentar o número de acessos de determinados sites a fim de faturar com publicidade digital, disseminação de ódio ou reforçar um pensamento. Sua dispersão é bastante prejudicial, pois podem afetar de maneira negativa pessoas normais, celebridades e políticos. Em épocas de eleições, pode-se notar um aumento ainda maior desse fenômeno, que costuma ser utilizado para difamar determinados candidatos, o que aconteceu frequentemente no Brasil durante o ano de 2018.

Como esse assunto tem se mostrado extremamente prejudicial e interessante ao mesmo tempo, e com o avanço da computação de modo geral, diversas empresas ao redor do mundo tem investido em tecnologia no combate ao problema das *Fake News*, geralmente utilizando técnicas de *Machine Learning* (ML), que é uma área de Inteligência Artificial (IA), na qual sistemas computacionais conseguem aprender através da análise de dados, identificando padrões e tomando decisões, juntamente com técnicas de Processamento de Linguagem Natural (PLN), que é um ramo das áreas de Linguística e também de IA, que auxilia sistemas computacionais na interpretação, entendimento e manipulação da linguagem natural (humana). Sendo assim, se tornou perfeitamente possível, com um nível de acurácia satisfatório, um computador definir se uma determinada notícia é falsa ou verdadeira.

* 1. **DEFINIÇÃO DO PROBLEMA**

Diversas técnicas e algoritmos podem ser utilizados para a resolução do problema das *Fake News*. É de extrema importância que se analise quais destas retornam resultados desejáveis, ou seja, quais possuem o melhor nível de acurácia, dadas determinadas situações, pois uma notícia classificada de maneira errada pode gerar danos à sociedade em geral. Além disso, é importante verificar a quantidade de tempo gasta em cada caso e qual possui melhor desempenho com volume de dados grande, médio e pequeno.

Dentre as variadas abordagens em *Machine Learning*, para o problema de determinar uma notícia como sendo verdadeira ou falsa, que é um problema de classificação, ou seja, dada uma entrada é atribuído um rótulo a ela, de caráter distinto como “verdadeiro” ou “falso”, são utilizadas técnicas de aprendizado supervisionado, onde o programa é treinado sobre um conjunto de dados já classificado e a partir daí é possível predizer a classificação de dados não classificados, com certo nível de exatidão. Alguns dos modelos comumente utilizados são *Support Vector Machine* (SVM) e *Naive Bayes* (NB). Para esses modelos, assim como para todos os outros existentes em aprendizado supervisionado, é necessário realizar um pré-processamento dos dados, utilizando técnicas de PLN.

Outra técnica que está sendo bastante utilizada ultimamente, devido aos seus resultados impressionantes é o *Deep Learning*, que é mais uma abordagem de ML, porém essa utiliza redes neurais artificiais profundas no processamento de informações e aprendizagem. Seu grande diferencial para as demais abordagens, consiste no fato de que com *Deep Learning* é possível trabalhar com análise de dados brutos, possibilitando um campo de atuação mais amplo. Alguns modelos utilizando são *Convolutional Neural Network* (CNN) e *Recurrent Neural Network* (RNN).

Para esse trabalho, as análises serão feitas sobre os modelos citados anteriormente.

* 1. **OBJETIVOS E METODOLOGIAS**

O objetivo desse trabalho consiste em treinar quatro algoritmos de aprendizado de máquina, sendo eles: *Support Vector Machine*, *Naive Bayes*, *Convolutional Neural Network* e *Recurrent Neural Network*. Antes da etapa de treinamento sobre os dados, serão realizadas técnicas de PLN para o pré-processamento dos mesmos. Para os algoritmos de aprendizado supervisionado (SVM e NB) serão utilizados *Bag-of-Words* (BoW) e *Term Frequency-Inverse Document Frequency* (TF-IDF). Já para os de *Deep Learning*, será utilizado *Word Embedding*.

Será utilizada uma base de dados em três versões: com 100.000, 10.000 e 1.000 dados. Com os resultados obtidos, será possível identificar, em cada caso, qual algoritmo de aprendizado supervisionado possui a melhor acurácia e qual a melhor forma de pré-processamento para eles (BoW ou TF-IDF), qual possui melhor acurácia entre CNN e RNN e qual é o que possui melhor nível de exatidão dentre todos os algoritmos. Além disso,