**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

Coorientador: Eduardo

**NITERÓI – RJ**

**2018**

**AGRADECIMENTOS**

Lorem ipsu.

**SUMÁRIO**

[1. introduÇÃO 1](#_Toc530513203)

[1.1. DEFINIÇÃO DO PROBLEMA 2](#_Toc530513204)

[1.2. OBJETIVOS E METODOLOGIAS 3](#_Toc530513205)

[1.3. CONFIGURAÇÕES 3](#_Toc530513206)

[2. INTELIGÊNCIA ARTIFICIAL 4](#_Toc530513207)

[3. MACHINE LEARNING 4](#_Toc530513208)

[3.1. APRENDIZADO SUPERVISIONADO 5](#_Toc530513209)

[3.1.1 CLASSIFICAÇÃO 5](#_Toc530513210)

[4. PROCESSAMENTO DE LINGUAGEM NATURAL 6](#_Toc530513211)

[4.1. Pré-processamento dos dados 7](#_Toc530513212)

[4.2. EXTRAÇÃO DE CARACTERÍSTICAS 7](#_Toc530513213)

[5. REFERÊNCIAS BIBLIOGRÁFICAS 8](#_Toc530513214)

# 1. introduÇÃO

Ao longo do século XXI, o número de usuários que utilizam a internet tem crescido ano após ano. “Em 2000, os internautas eram 6,5% da população mundial. Em 2015, esse índice subiu para 43%. A proporção de casas com conexão à rede chegou a 46% no ano. Essa porcentagem é maior na Europa (82,1%) e menor na África (10,7%).” (G1, 2015, p. 1). 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. Dessa maneira, é importante ressaltar que:

A internet é o segundo meio de comunicação usado mais frequentemente pelos brasileiros, atrás da televisão e à frente do rádio, segundo a primeira edição da "Pesquisa Brasileira de Mídia 2014 – Hábitos de Consumo de Mídia pela População Brasileira", divulgada nesta sexta-feira (7) e encomendada ao Ibope pela Secretaria de Comunicação Social da Presidência da República. (G1, 2014, p. 1).

De acordo com Cunha (2018, p.1) em relação às *Fake News:*

A ampla facilidade de acesso, o incremento e a popularização de mecanismos para a criação de conteúdo, aliados à extensão espacial que as mencionadas redes sociais podem conferir ao conteúdo produzido criaram um fenômeno curioso: a elaboração de peças ditas informativas sem critério jornalístico e muitas vezes sem nenhuma correspondência na realidade. Trata-se do que se convencionou chamar fake news (notícias falsas).

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, e dessa maneira o fenômeno social conhecido como *Fake News* pode 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 desde pessoas normais até 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.

Segundo Batista (2018), as empresas Google e Facebook divulgaram que irão combater tal fenômeno, devido ao seu crescimento espantoso, no intuito de frear a dissipação de informações falsas em um universo que é tão difícil de monitorar como o da internet. 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) e Processamento de Linguagem Natural (PLN), propiciando que um computador defina se uma determinada notícia é falsa ou verdadeira, com um nível de precisão satisfatório. Em relação à essas tecnologias:

Empresas [...] estão construindo e treinando modelos de Machine Learning que “leem” artigos denunciados como sendo Fake News e com isso aprendem a identificar certos padrões comuns nesse tipo de atividade. [...] vale ressaltar que a PLN, [...] é principal técnica de identificação de notícias falsas. Este é um campo em rápida expansão. Temos como o Google Tradutor, que nos últimos anos apresentou grandes avanços, justamente fazendo uso das pesquisas mais recentes na área de Processamento de Linguagem Natural.

## 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 ideal 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 utilizados são *Convolutional Neural Network* (CNN) e *Recurrent Neural Network* (RNN). Para esse trabalho, as análises serão feitas sobre os modelos citados anteriormente.

## 1.2. 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, que para esse trabalho estão classificadas como seu tamanho sendo grande, médio e pequeno, respectivamente. Com os resultados obtidos, será possível identificar, em cada caso, qual algoritmo de aprendizado supervisionado (SVM e NB) possui a melhor acurácia e qual a melhor forma de pré-processamento para eles (BoW ou TF-IDF), qual algoritmo de *Deep Learning* (CNN e RNN) possui melhor acurácia e qual é o que possui melhor nível de exatidão dentre todos os algoritmos. Além disso, os tempos de execução para cada caso serão medidos, a fim de verificar se apesar de um possível resultado satisfatório, o tempo com que o mesmo foi obtido torna o algoritmo interessante de ser utilizado ou não. Ainda falando sobre medição temporal, um dado de teste será utilizado para classificação, no intuito de simular quanto tempo levaria para uma aplicação retornar o valor. Uma análise comparativa dos melhores parâmetros para cada algoritmo também será feita.

## 1.3. CONFIGURAÇÕES

O computador no qual esse trabalho foi desenvolvido com as seguintes configurações:

* Sistema operacional Windows 10 – 64 bits.
* Processador Intel Core i5 – 6200U 2.3 GHz.
* Memória RAM DDR3 de 8GB.
* HD de 1TB

# 2. INTELIGÊNCIA ARTIFICIAL

O ser humano sempre buscou conhecimento. Faz parte de nossa natureza intrínseca querer entender como as coisas funcionam, e é essa característica que nos distingue de qualquer outro ser vivo. Não estamos satisfeitos em saber que algo existe, e sim o porquê. “Durante milhares de anos, procuramos entender como pensamos, isto é, como um mero punhado de matéria pode perceber, compreender, prever e manipular um mundo muito maior e mais complicado que ela própria”. (NORVIG, 2013, p.3).

A área de Inteligência Artificial surgiu, derivada dessa linha de raciocínio, do desejo de construir sistemas computacionais inteligentes, ou seja, que possam realizar tarefas assim como os seres humanos fazem ou até melhor. Tarefas como a capacidade de aprender, perceber e compreender o mundo ao seu redor, tomar decisões e realizar previsões. No entanto, definir o que é Inteligência artificial não é uma tarefa fácil, pois na literatura existem algumas definições. Todas são divididas em duas idéias, cada uma relacionando com o conceito de comportamento e com o processo de pensar e raciocinar, na qual uma define o sucesso de um sistema inteligente como sendo aquela no qual imita o comportamento humano e a outra define o sucesso relacionado à idéia de racionalidade, ou seja, que de acordo com a informação que recebe, toma a atitude correta (NORVIG, 2013).

# 3. MACHINE LEARNING

Estamos vivendo na época do *Big Data*, onde cada vez mais empresas se vêem na necessidade de armazenar um grande amontoado de dados e há a necessidade da extração de informações dessas bases que sejam relevantes para quem as utiliza, a fim de facilitar na tomada de decisões. *Machine Learning* é uma subárea de Inteligência Artificial na qual se podem utilizar métodos automatizados para a análise de dados. Sobretudo, envolve a construção de modelos matemáticos que são os suportes para facilitar o entendimento desses elementos, no qual um programa pode “aprender” através deles, dados alguns parâmetros para os modelos (MURPHY, 2002).

**[Falar sobre seu uso]**

## 3.1. APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado é uma das duas categorias principais (aprendizado não supervisionado é a outra) de *Machine Learning*. Nesta categoria, o modelo tenta encontrar padrões que relacionam as *features* (entradas dos algoritmos de *Machine Learning*) com suas *labels* (saídas). Em seguida, utilizando novos dados que não estão com suas *saídas*, é possível, utilizando o modelo que encontrou um determinado padrão, ou seja, encontrou uma relação entra a entrada e a saída, prever quais são suas *labels*. O aprendizado supervisionado pode ser dividido em dois grupos: classificação e regressão. Para este trabalho, parte dele foi feito utilizando aprendizado supervisionado e a sua categoria de classificação.

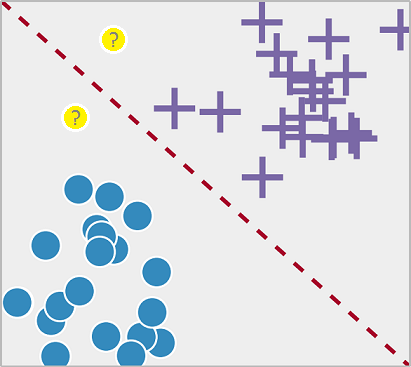
### 3.1.1 CLASSIFICAÇÃO

Basicamente, o problema da classificação consiste em, dados um conjunto de *features* que estão classificadas, ou seja, estão com suas *labels*, queremos classificar um conjunto de novas entradas que não estão classificadas. Em outras palavras, o objetivo da classificação é mapear as entradas em categorias distintas.

De acordo com as conclusões de Murphy (2012, p.3, tradução nossa):

Aqui, o objetivo é aprender o mapeamento das entradas x para as saídas , onde ∈ {1,...,C}, com C sendo o número de classes. Se C = 2, chamamos de classificação binária (caso em que muitas vezes assumimos y ∈ {0,1}); Se C > 2, chamamos de classificação multiclasse. [...] Uma maneira para formalizar o problema é como uma função de aproximação. Assumimos = (x) para alguma função f desconhecida, e o objetivo do aprendizado é estimar a função f dado um conjunto de treinamento rotulado, e então fazer previsões usando = (x). [...] Nosso objetivo principal é fazer previsões sobre novos insumos, o que significa que são aqueles que não foram vistos anteriormente [...].

Figura 1 - Exemplo de classificação binária



Fonte: (BAIA, 2016, p. 1)

A figura 1 representa um exemplo de classificação binária, onde temos dois conjuntos (círculo e cruz) e dadas duas entradas não rotuladas, queremos saber a qual conjunto elas pertencem. Para a identificação de *Fake News* neste trabalho, a utilização da classificação binária se encaixa perfeitamente, visto que temos algumas bases de dados contendo notícias que estão rotuladas como real ou falsa.

# 4. PROCESSAMENTO DE LINGUAGEM NATURAL

A linguagem natural humana é algo fascinante. Nossa habilidade de desenvolver espontaneamente uma língua bem estruturada, com o intuito de nos comunicarmos dentro de uma comunidade, é algo que somente os seres humanos são capazes de fazer. Diferentemente da linguagem formal, como as linguagens de programação, utilizadas nos computadores, a formalização da linguagem natural não se dá de maneira trivial. Dessa forma, desenvolver sistemas computacionais que possam entender nossa língua (falada e escrita) é uma tarefa extremamente complexa. Com esse intuito, surgiu o Processamento de Linguagem Natural (PLN), que é mais uma das subáreas da Inteligência Artificial que tentam resolver esse problema.

Nesse sentido, Goldberg (2017, p.1, tradução nossa) afirma:

Processamento de Linguagem Natural (PLN) é o campo de projetar métodos e algoritmos que pegam como entrada ou produz como uma saída não estruturada, dados de linguagem natural [...]. Entender e produzir linguagem usando computadores é, portanto, altamente desafiador. De fato, o mais conhecido conjunto de métodos para lidar com dados de linguagem é usando aprendizado de máquina supervisionado, que tenta inferir padrões de uso e regularidades a partir de um conjunto de entradas pré-anotadas e pares de saída.

Logo, apesar de ser uma tarefa complicada, existem maneiras já estudadas de se trabalhar com PLN, com o objetivo de que os computadores consigam interpretar, entender, extrair informações e analisar (de maneira sintática, semântica, léxica e morfológica) textos, através do seu processamento.

## 4.1. Pré-processamento dos dados

Quando estamos trabalhando com aplicações PLN, existe a necessidade de transformar textos brutos em algo no qual o modelo de *Machine Learning* consiga processá-los de maneira melhor, ou seja, é necessário realizar determinados processos sobre dados textuais, a fim de refiná-los, de maneira que eles possam servir da forma mais eficiente possível como dados de entrada. Para tal, é necessário que se faça o pré-processamento dos dados.

De maneira precisa, Kęsik (2017, p.1, tradução nossa) diz que:

Pré-processamento de dados é o processo de limpar e preparar o texto para classificação. O texto usualmente contém algum ruído e dados menos significativos. Isso precisa ser identificado e limpo para criar classificadores mais fortes, rápidos e precisos. Se tais palavras insignificantes não forem removidas, isso aumentaria a complexidade, adicionando mais dimensões na extração de recursos.

Sendo assim, quanto mais reduzirmos os ruídos e dados que não são significativos, melhores serão nossas entradas e consequentemente iremos obter melhores resultados. Existem diversas técnicas que podem ser aplicados ao texto como forma de pré-processamento. As mais utilizadas na literatura, e por esse trabalho são conhecidas como: Tokenização, Capitalização, Normalização, Palavras de parada (*Stopwords)* e Lematização.

A Tokenização costuma ser uma das primeiras técnicas aplicadas no texto. Essa técnica consiste em pegar cada texto e separá-los em partes menores. Pode-se separar parágrafos em sentenças e sentenças em palavras ser separados em sentenças ou textos, criando-se assim *tokens*

## 4.2. EXTRAÇÃO DE CARACTERÍSTICAS

# 5. REFERÊNCIAS BIBLIOGRÁFICAS

BAIA, C. Introdução ao Machine Learning. **Carlos Baia**, 17 jul. 2016. Disponivel em: <http://carlosbaia.com/2016/07/17/introducao-ao-machine-learning/>. Acesso em: 10 set. 2018.

G1. Segundo meio de comunicação mais usado é internet, aponta pesquisa. **G1**, 2014. Disponivel em: <http://g1.globo.com/economia/midia-e-marketing/noticia/2014/03/segundo-meio-de-comunicacao-mais-usado-e-internet-aponta-pesquisa.html>. Acesso em: 14 set. 2018.

G1. Mundo tem 3,2 bilhões de pessoas conectadas à internet, diz UIT. **G1**, 2015. Disponivel em: <http://g1.globo.com/tecnologia/noticia/2015/05/mundo-tem-32-bilhoes-de-pessoas-conectadas-internet-diz-uit.html>. Acesso em: 23 ago. 2017.

GOLDBERG, Y. **Neural Network Methods for Natural Language Processing**. 1ª. ed. Toronto: Morgan & Claypool Publishers, 2017.

KĘSIK, A. Text Classification (Topic Categorization, Spam filtering, etc). **Commonlounge**, 2017. Disponivel em: <https://www.commonlounge.com/discussion/ea97b828a7a24a5ca11b50da6e3b4b09/history>. Acesso em: 20 set. 2018.

MURPHY, K. **Machine Learning:** A Probabilistic Perspective. 1ª. ed. Massachussets: The MIT Press, 2012.

NORVIG, P. **Inteligência Artificial**. Tradução de Regina SIMILLE. 3ª. ed. Rio de Janeiro: Elsevier, 2013.

VANDERPLAS, J. **Python Data Science Handbook**. 1ª. ed. Califórnia: O'Reilly, 2017.