UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Elizangela Santos da Costa

Avaliando Atributos de Credibilidade de Páginas Web utilizando Aprendizagem de Máquina

Manaus
Maio de 2020

Elizangela Santos da Costa

Avaliando Atributos de Credibilidade de Páginas Web utilizando Aprendizagem de Máquina

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para obtenção do grau de Mestre em Informática.

Orientador: Prof. Dr. Eduardo Luzeiro Feitosa

Manaus
2020

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Costa, Elizangela Santos da
C837a Avaliando atributos de credibilidade de páginas Web utilizando Aprendizagem de Máquina : elaboração de um método de avaliação de credibilidade por meio da extração de atributos / Elizangela Santos da Costa . 2020
54 f.: il. color; 31 cm.

Orientador: Eduardo Luzeiro Feitosa
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. credibilidade. 2. avaliação. 3. web. 4. atributos. I. Feitosa, Eduardo Luzeiro. II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Avaliando Atributos de Credibilidade de Páginas Web utilizando Aprendizagem de Máquina"

ELIZANGELA SANTOS DA COSTA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Eduardo Luzeiro Feitosa - PRESIDENTE

Prof. Rafael Giusti - MEMBRO INTERNO

Prof. Altair Olivo Santin - MEMBRO EXTERNO

Manaus, 11 de Maio de 2020

Resumo

As informações compartilhadas na Web se propagam rapidamente, sejam elas verdadeiras ou não. O objetivo de reproduzir informações incorretas ou falsas está relacionado a diversos fatores como manipulação política, obtenção de benefícios financeiros, disseminação de difamações, entre outros. Sendo assim, verificar a credibilidade das informações disponíveis na Web acaba sendo uma tarefa obrigatória. Dentre as diversas soluções desenvolvidas para detectar se uma página Web pode ser acreditada ou não, as baseadas em aprendizagem de máquina são a mais empregadas. Esta dissertação visa avaliar e definir atributos empregáveis em um futuro modelo de avaliação de credibilidade de páginas Web, por meio da extração de características do conteúdo da página e da rede, com o auxílio de classificadores de aprendizagem de máquina, possibilitando assim maior certeza sobre a credibilidade de páginas Web. Como resultado, esta dissertação concluiu que o classificador Random Forest teve o melhor resultado para avaliação de credibilidade de páginas web com 95.36% de acurácia. Além de disponibilizar um script de extração de atributos, apontou também quais são os atributos mais relevantes e de fácil extração que podem ser obtidos de qualquer URL, para isso utilizou 3 métodos de seleção de atributos: Select kbest, Seleção RFE e Seleção RFECV, no qual este último apresentou o melhor resultado com 95.33% de acurácia.

Palavras-chave: Credibilidade, Avaliação de credibilidade, Web, Atributos.

Abstract

Information shared on the Web propagates quickly, whether true or not. Credibility in this context refers to the level of trust a user places subjectively on a Web page. The purpose of reproducing incorrect information is related to several factors such as political manipulation, obtain financial benefits, disseminate malicious defamation, among others. Therefore, verifying the credibility of the information available on the Web ends up being a mandatory task. Among the various techniques developed to detect whether a Web page can be accredited or not, machine learning is the most used in comparison to the assessment of credibility manually. The purpose of this work is to evaluate and define attributes that can be used in a future model for assessing the credibility of Web pages, by extracting characteristics from the content of the page and the network, with the help of machine learning classifiers, thus enabling greater certainty on the credibility of web pages.As a result, this dissertation concluded that the Random Forest classifier had the best result for assessing the credibility of web pages with 95.36% accuracy. In addition to providing an attribute extraction script, also pointing out which are the most relevant and easy extraction attributes that can be selected for any URL, for that, 3 attribute selection methods are used: Select the best, RFE Selection and Selection RFECV, the last result with 95.33% accuracy.

Keywords: credibility, credibility assessment, web, information.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Hoje em dia se obtém na Internet grande parte das informações necessárias para a maioria das atividades humanas. Assim, as pessoas têm empregado mais e mais informações on-line em seu cotidiano para compras, negócios, vida social, relacionamentos, entre outros (Aggarwal, 2014; Papaioannou and Aberer, 2012).

Contudo, a Internet também é uma fonte de informações falsas e alegações maliciosas, tanto em mídias sociais quanto em sites e portais, com a capacidade de atingir rapidamente milhões de usuários. Segundo Schwarz (2011), embora haja uma grande quantidade de informações úteis disponíveis on-line, páginas Web enganosas e incorretas continuam a proliferar. Assim, não se pode confiar em todas as informações publicadas na Web, sejam elas textuais, visuais (imagens e vídeos) ou auditivas. É comum encontrar informações contraditórias publicadas em sites diferentes ou até mesmo no mesmo site (Wahsheh et al., 2013). Essa desinformação ocorre de várias formas: citações errôneas, dossiês sobre políticos ou empresas, falsas resenhas sobre produtos ou serviços, notícias sobre celebridades e assim por diante. Detectar falsas alegações é um desafio para os seres humanos (Popat, 2017b).

Nesse sentido, a avaliação da credibilidade de páginas Web está se tornando um aspecto cada vez mais importante do âmbito do conhecimento da informação, pois a falta de controle de qualidade permite que informações incorretas ou de baixa qualidade sejam publicadas (Pattanaphanchai et al., 2013). Se por um lado a credibilidade é uma característica importante para autores e editores que querem tornar seus materias (livros e sites, por exemplo) credíveis, dado que fontes de alta credibilidade são mais valiosas, páginas Web não confiáveis podem ter consequências graves quando as pessoas usam essas informações como base para decisões em domínios críticos como política, finanças e saúde (Schwarz, 2011).

Com o crescente número de boatos e rumores, vários serviços e sites de checagem da verdade (Snopes[1], Politifact[2], Truthorfiction[3], entre outros) tornaram-se populares. De forma geral, eles compilam artigos escritos por especialistas que investigam manualmente afirmações controversas, determinando a procedência e a autenticidade de várias fontes e fornecendo um veredito com evidência de apoio (Popat, 2017a). Entretanto, essa abordagem traz a desvantagem de ser feita manualmente e técnicas mais eficientes devem ser elaboradas.

[1] http://www.snopes.com
[2] http://www.politifact.com
[3] http://www.truthorfiction.com

1.1 Motivação

É difícil para os usuários analisarem, subjetivamente, a credibilidade da informação sem conhecimento e experiência. A incerteza da veracidade das informações tem sido um grande obstáculo para que usuários obtenham e compartilhem informações verdadeiras na Internet (Pattanaphanchai et al., 2013). Via de regra, no momento da verificação, os usuários tendem a basear seus julgamentos em critérios como a apresentação visual da informação ao invés de utilizar critérios normativos robustos. Isso comprova que os usuários precisam de critérios de credibilidade para auxiliá-los a julgar rapidamente a informação.

Existem trabalhos na literatura, como o de (Bezerra, 2015), que avaliam atributos de páginas Web, mas o foco não é credibilidade, e sim *spam*, *phishing*, entre outras atividades maliciosas. Também existem trabalhos ligados a detecção comptucional de credibilidade de páginas Web (Papaioannou and Aberer, 2012; Liu, 2013; Schwarz, 2011), que tentam definir critérios para a avaliação e validação da qualidade de páginas, mas não possuem o foco na análise dos atributos.

Considerando o problema apresentado, a pesquisa proposta nesta disertação visa aplicar classificadores de aprendizagem de máquina em características extraídas do conteúdo de páginas Web e da rede, para dar apoio no processo de verificação de credibilidade na Web, diminuindo assim possíveis enganos dos usuários.

1.2 Objetivos

O objetivo geral desta dissertação é avaliar atributos para mensurar a credibilidade das informações em páginas Web, por meio da combinação de técnicas de extração de conteúdo com aprendizagem de máquina, a fim de fornecer uma avaliação de credibilidade empregável em futuras soluções.

Como objetivos específicos, pretende-se:

- Identificar atributos de páginas Web passíveis de mensuração de credibilidade;
- Elaborar um método de extração de conteúdo e mensuração de credibilidade apto a funcionar como entrada para modelos de verificação de fatos.

1.3 Contribuições Esperadas

Espera-se que, ao final do trabalho, as seguintes contribuições sejam alcançadas:

- Elaboração de uma taxonomia de abordagens que analisam a credibilidade de sites na Web;
- Um modelo de avaliação de credibilidade para sites tendo a URL como entrada.

1.4 Estrutura do Documento

Este documento está organizado em 5 capítulos. O Capítulo 2 apresenta os conceitos necessários para a compreensão desta proposta. O Capítulo 3 descreve os trabalhos relacionados sobre avaliação da credibilidade na Web e as principais abordagens na área. O Capítulo 4 detalha os atributos, a visão geral e metodologia. O Capítulo 5 apresenta o protocolo experimental e resultados. Por fim, o Capítulo 6 expõe a conclusão do trabalho.

Capítulo 2

Conceitos Básicos

Neste capítulo são apresentados os principais conceitos relacionados ao tema de pesquisa sobre credibilidade da informação na Web. São abordadas definições sobre credibilidade e verdade, aprendizagem de máquina e processamento de linguagem natural, visando auxiliar o melhor entendimento sobre o desenvolvimento do trabalho.

2.1 Credibilidade

Há uma confusão entre os conceitos de credibilidade e verdade. Informações podem ser verdadeiras e não credíveis, verdadeiras e credíveis, falsas e credíveis ou falsas e não credíveis. Credibilidade é descrito como um estado mental subjetivo inerente dos humanos enquanto verdade é um conceito frequentemente entendido como universal e objetivo (Wierzbicki, 2018).

Em Aggarwal (2014), a credibilidade é definida como o nível de confiança que um usuário coloca numa determinada página Web disponível, baseado em vários fatores objetivos e subjetivos. Para Ginsca (2015) existem três tipos de credibilidade: fonte, mídia e mensagem. De acordo com os autores, **credibilidade da fonte** diz respeito a um ato de comunicação envolvendo uma fonte e um receptor de informação (uma mensagem). A credibilidade da fonte é sempre avaliada por um receptor. A **credibilidade da mídia** se baseia no fato de que as várias mídias disponiveis na Web (páginas comuns, blogs, mídias sociais e sites de perguntas e respostas, por exemplo) têm seus próprios aspectos específicos e, por isso, podem influenciar a avaliação da credibilidade. Em outras palavras, embora não dependa da relação entre a fonte e o receptor, as várias mídias podem modificar o impacto da credibilidade da fonte e da credibilidade da mensagem nas avaliações gerais de credibilidade. A **credibilidade da mensagem** é um estado mental dependente do contexto do receptor da mensagem. É avaliada não apenas pelo receptor, mas também pelo remetente da mensagem. A credibilidade da mensagem é definida como um sinal que pode fazer um receptor acreditar que a informação é verdadeira.

Ainda segundo Ginsca (2015), no domínio Web é possível identificar quatro tipos de credibilidade em relação à credibilidade da fonte e da mensagem:

- **Credibilidade presumida**: Baseada em suposições gerais na mente dos usuários (por exemplo, fidedignidade dos identificadores do domínio).
- **Credibilidade da superfície**: Derivada da inspeção de um site, é frequentemente baseada na primeira impressão que um usuário tem de um site e muitas vezes influenciada pelo quão profissional é o design do site Web.

- **Credibilidade adquirida**: Refere-se à confiança estabelecida ao longo do tempo e muitas vezes influenciada pela facilidade de uso de um site e sua capacidade de consistentemente fornecer informações confiáveis.

- **Credibilidade de renome**: Refere-se a opiniões de terceiros do site, como quaisquer certificados ou prêmios que o site ganhou.

Fogg (1999) descrevem credibilidade como uma qualidade percebida e composta de múltiplas dimensões. É uma qualidade percebida, pois não reside em um objeto, pessoa ou informação. Essa percepção resulta de múltiplas dimensões simultaneamente. Os autores afirmam que embora a literatura varie de quantas dimensões contribuem para as avaliações de credibilidade, a grande maioria dos pesquisadores identifica dois componentes chave da credibilidade: a *expertise* e a fidedignidade. A ***expertise*** captura o conhecimento e a habilidade percebidos da fonte e é definida por termos como "bem-informado", "experiente", "competente"e assim por diante. Já a **fidedignidade** captura a bondade percebida ou a moralidade da fonte e é definida pelos termos "bem-intencionados", "verdadeiros", "imparciais"e assim por diante.

Na ciência da computação, a maioria das abordagens ligadas à fidedignidade enfatiza fortemente a autoridade, onde a fonte conhecida é usada para informar a credibilidade. Fontes fidedignas são usadas como um indicador da credibilidade de uma dada informação. Na ausência de autoridade externa explícita, o usuário deve confiar no próprio conteúdo.

Ginsca (2015) adicionam dois componentes de interesse particular no julgamento da credibilidade: qualidade e confiabilidade. As percepções de Qualidade estão intimamente associadas à credibilidade, com alguns trabalhos identificando qualidade como conceito super-ordenado, alguns visualizando os dois como associados a categorias separadas e alguns em relação à qualidade como subordinados à credibilidade. Para Ginsca (2015), a qualidade também pode estar ligada ao interesse que determinado conteúdo pode causar. Ao lidar com dados textuais de qualquer tamanho, uma das características mais importantes para estimar a credibilidade da mensagem transmitida é a qualidade do texto. Isso é especialmente importante quando há pouca ou nenhuma informação sobre a origem do texto ou quando a veracidade do conteúdo não pode ser facilmente verificada. Já a confiabilidade comumente se refere a algo percebido como confiável e consistente em qualidade, geralmente em um eixo temporal. Mais especificamente, a confiabilidade do conteúdo de texto pode ser definida como o grau em que o conteúdo do texto é percebido como verdadeiro. A confiabilidade do conteúdo é um critério que, seguindo a relevância do tópico, é um dos aspectos mais influentes que devem ser considerado para avaliar a relevância de uma publicação na Web (Ginsca, 2015).

É importante deixar claro que a credibilidade utilizada neste trabalho utilizará atributos de credibilidade da superfície, pois são aqueles que podemos visualizar em uma página Web. A fonte da credibilidade será a Web e a credibilidade estudada será a credibilidade da mensagem visto que o interesse é na análise da veracidade da notícia.

2.2 Aprendizagem de Máquina

A aprendizagem de máquina é parte da inteligência artificial que usa a teoria da estatística na construção de modelos matemáticos, cuja tarefa principal é a inferência de uma amostra. Como resultado, tem-se um modelo definido, baseado em alguns parâmetros. O modelo pode ser preditivo (para fazer previsões do futuro), descritivo (para conhecimento dos dados) ou ambos (Alpaydin, 2010).

De acordo com (Alpaydin, 2010), a aprendizagem de máquina pode ser de três tipos: supervisionada, não supervisionada e por reforço. Aplicações em que os dados de treino compreendem

exemplos dos vetores de entrada, juntamente com os seus vetores alvo correspondentes, são conhecidos como problemas de **aprendizagem supervisionados** (Bishop, 2006). Tanto regressão quanto classificação são problemas de aprendizagem supervisionada, onde o objetivo é aprender o mapeamento da entrada para a saída. A classificação é usada quando o objetivo é atribuir um vetor de entrada para um vetor de um número finito de categorias discretas. Um exemplo de aplicação é o problema de reconhecimento de dígitos. A regressão é usada quando a saída desejada consiste em uma ou mais variáveis contínuas. Um exemplo de aplicação é a predição do preço de um carro.

Em outros problemas de reconhecimento de padrões, os dados de treinamento consistem em um conjunto de vetores de entrada x sem nenhum valor alvo correspondente. O objetivo em tais problemas de **aprendizado não supervisionado** pode ser descobrir grupos de exemplos semelhantes dentro dos dados, onde ele é chamado de agrupamento, determinar a distribuição de dados dentro do espaço de entrada, conhecido como estimativa de densidade, ou projetar os dados de um espaço de alta dimensão até duas ou três dimensões para fins de visualização.

Finalmente, a técnica de **aprendizado por reforço** está relacionada ao problema de encontrar ações adequadas a serem tomadas em uma dada situação, a fim de maximizar uma recompensa.

A aprendizagem de máquina tem sido o método mais utilizado nos problemas de avaliação de credibilidade de notícias. Nesta pesquisa pretende-se usar aprendizagem supervisionada, onde os vetores de entrada são as características extraídas de uma página Web e o vetor de saída é a avaliação da credibilidade.

Uma **árvore de decisão** é um modelo hierárquico para aprendizado supervisionado, em que a região local é identificada em uma sequência de divisões recursivas em um número menor de etapas. Uma árvore de decisão é composta pelo nó de decisão, nó de decisão interno e pelas folhas terminais. Dada uma entrada, em cada nó, um teste é aplicado e um dos ramos é obtido dependendo do resultado. Esse processo inicia na raiz e é repetido recursivamente até que um nó folha seja atingido, ponto no qual o valor gravado na folha constitui a saída (Alpaydin, 2010), conforme Figura 2.1.

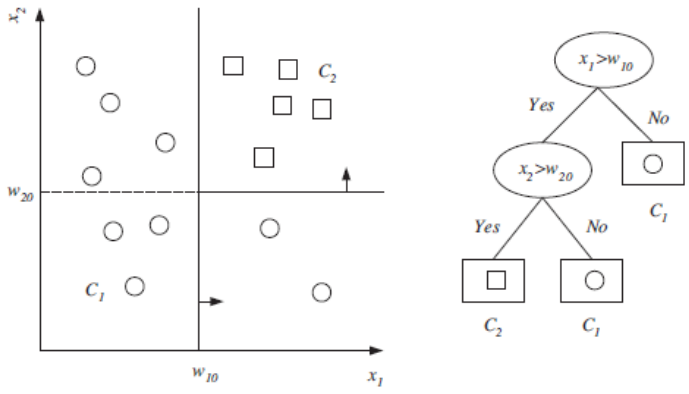

Figura 2.1: Árvore de decisão
Fonte: (Alpaydin, 2010)

De acordo com (Alpaydin, 2010), uma árvore de decisão também é um modelo não paramétrico no sentido de que não assume nenhuma forma paramétrica para as densidades de classe e

a estrutura da árvore não é fixada a priori, mas a árvore cresce, ramos e folhas são adicionados durante a aprendizagem dependendo da complexidade do problema inerente aos dados.

Em **Support Vector Machine (SVM)**, a solução é fundamentalmente um classificador de duas classes. Eles são baseados em um algoritmo que encontra um tipo especial de modelo linear: o hiperplano de margem máxima. Dado um conjunto de dados de duas classes cujas classes são linearmente separáveis, ou seja, há um hiperplano no espaço de instâncias que classifica todas as instâncias de treinamento corretamente. O hiperplano de margem máxima é aquele que dá a maior separação entre as classes - não se aproxima mais do que é necessário. Um exemplo é mostrado na Figura 2.2 , em que as classes são representadas por círculos abertos e preenchidos, respectivamente.

O hiperplano de margem máxima é definido pela localização dos vetores de suporte que são as instâncias mais próximas do hiperplano - aquelas com distância mínima. Há sempre pelo menos um vetor de suporte para cada classe e muitas vezes há mais. O importante é que o conjunto de vetores de suporte define exclusivamente o hiperplano de margem máximo para o problema de aprendizado. O limite de decisão é escolhido para ser aquele para o qual a margem é maximizada (Bishop, 2006; Witten, 2011).

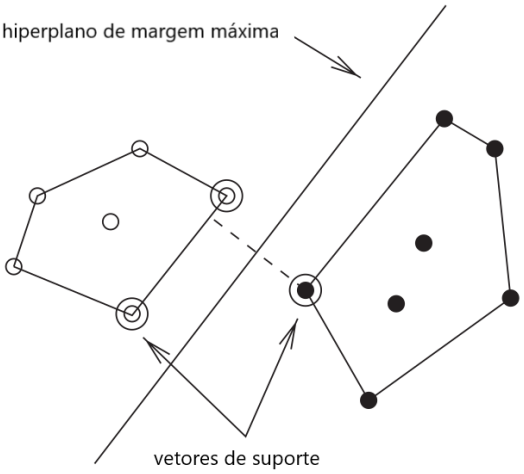

Figura 2.2: SVM
Fonte: (Witten, 2011)

Relativo a essa margem, existe o parâmetro C que controla o nível de complexidade do modelo. Basicamente o que esse parâmetro faz é definir se o modelo deverá ser mais rígido ou menos rígido com relação aos erros de classificação. O modelo mais rígido irá penalizar as fronteiras de separação que tem erros de classificação, com isso, esse modelo cria fronteiras de separação mais perfeitas. O modelo menos rígido irá penalizar menos as fronteiras de separação que tem erros de classificação, com isso, este modelo cria fronteiras de separação menos perfeitas, porém mais amplas e simples.

O SVM também pode ser utilizado em problemas de separação não linear, para isso, utiliza-se o truque de kernel que são funções matemáticas que ocupam um espaço de entrada dimensional baixo e o transformam em um espaço dimensional mais alto, isto é, ele converte um problema não separável em um problema separável.

Random Forest, introduzido por Breiman (2001), é um método de aprendizado de máquina que faz predições baseadas nos resultados de múltiplas árvores de decisão independentes, sendo

capaz de resolver problemas de classificação, regressão e outras tarefas. Segundo (Treeratpituk, 2009), cada árvore de decisão dentro da floresta é montada com uma amostra de inicialização diferente, extraída do conjunto de dados original. Cada árvore é então construída até o tamanho máximo sem qualquer poda. A seleção de variáveis para cada divisão na árvore é realizada em um subconjunto de características selecionadas aleatoriamente, em vez de no conjunto completo de características, como normalmente é feito na árvore de decisão tradicional. Uma vez que a floresta é construída, a classificação pode ser feita simplesmente agregando os votos de todas as árvores.

Existem apenas dois parâmetros para ajustar *random forest*: T, o número de árvores, e M, o número de características a serem consideradas ao dividir cada nó. A taxa de erro de uma *random forest* depende de dois fatores: a correlação entre as árvores na floresta e a força de cada árvore individual. Quanto mais correlacionadas forem as árvores, maior será a taxa de erro. Quanto mais forte for a árvore individual (alta precisão), menor será a taxa de erro.

Para avaliação dos atributos do modelo de credibilidade, esta dissertação utiliza os algoritmos SVM (Support Vector Machine), árvore de decisão e Random Forest.

2.2.1 Medidas de Aprendizagem de máquina

Nesta seção são discutidos os tipos de avaliação de desempenho dos classificadores. Contudo, antes de apresentar essas métricas, se faz necessário explicar os conceitos de verdadeiro positivo (TP), verdadeiro negativo (TN), falso positivo (FP) e falso negativo (FN) (Witten, 2011).

Os verdadeiros positivos (TP) e verdadeiros negativos (TN) são classificações corretas. Um falso positivo (FP) ocorre quando o resultado é incorretamente previsto como positivo quando na verdade não o é (negativo). Um falso negativo (FN) ocorre quando o resultado é incorretamente previsto como negativo quando é realmente positivo

Esses quatro conceitos permitem calcular efitivamente medidas de desempenho, como a **acurácia**, que representa o total geral de acerto; a **precisão**, que é o percentual de documentos recuperados que são relevantes; a **revocação** (também chamada de recuperação ou *recall*), que diz respeito ao percentual de documentos relevantes que são retornados; a **medida F** que sintetiza as informações das medidas de precisão e recall obtendo dessa maneira a média harmônica poderada entre elas, e **sensibilidade x especificidade**, a sensibilidade é a mesma que o recall. Especificidade é o quão bem detectamos os negativos. O cálculo dessas medidas podem ser visualizadas na Tabela 2.1

Tabela 2.1: Medidas de desempenho da aprendizagem de máquina

Medida	**Fórmula**
Acurácia	TP+TN/(TP+FP+FN+TN)
Precisão	TP/(TP+FP)
Recuperação/revocação/recall	TP/(TP+FN)
Medida F	(2×recuperação×precisão)/(recuperação+precisão)
Sensibilidade × especifidade	(TP / (TP + FN)) × (TN / (FP + TN))

Além dessas medidas, na área de apredizagem de máquina também se utilizam as curvas ROC (*Receiver Operating Characteristic*) e AUC (*area under the ROC curve*). Ambas serão utilizada nesta dissertação.

A curva ROC mostra o quão bom o modelo criado pode distinguir entre duas coisas (já que é utilizado para classificação). Essas duas coisas podem ser 0 ou 1, ou positivo e negativo.

Os melhores modelos conseguem distinguir com precisão o binômio. A curva ROC possui dois parâmetros: taxa de verdadeiro positivo (TP) e a taxa de falso positivo (FP).

Para simplificar a curva ROC, foi criada a AUC. A AUC resume a curva ROC num único valor, calculando a "área sob a curva". Sabendo que, idealmente, um classificador deve ter taxa TP de 1 e taxa FP de 0, quanto mais próxima a curva AUC for de 1 (quanto mais se aproxima do canto superior esquerdo), melhor é o resultado obtido.

Por fim, também será empregada a **matriz de confusão**, uma predição de multiclasse, cujo resultado em um conjunto de testes é geralmente exibido como uma matriz bidimensional com uma linha e coluna para cada classe. Bons resultados correspondem a grandes números nos elementos diagonal e pequeno, idealmente zero, fora da diagonal. A Figura 2.3 mostra um exemplo numérico com três classes. Nesse caso, o conjunto de testes possui 200 instâncias (a soma dos nove números na matriz) e 88 + 40 + 12 = 140 deles são previstos corretamente, portanto a taxa de sucesso é de 70%.

		Classe Prevista			
		a	b	c	Total
Classe atual	a	88	10	2	100
	b	14	40	6	60
	c	18	10	12	40
	Total	120	60	20	

Figura 2.3: Matriz de confusão
Fonte: (Witten, 2011)

Nesta dissertação, usam-se todas as medidas de avaliação.

2.3 Processamento de Linguagem Natural

A linguagem natural representa uma linguagem que é usada para comunicação cotidiana por humanos. Em contraste com as linguagens artificiais, como as linguagens de programação e as notações matemáticas, as linguagens naturais evoluíram à medida que passam de geração para geração e são difíceis de definir com regras explícitas. O Processamento de Linguagem Natural - ou PLN - pode ser tão simples quanto contar frequências de palavras para comparar diferentes estilos de escrita como também entender expressões humanas completas, pelo menos até o ponto de ser capaz de dar respostas úteis a elas (Bird and Klein, 2009).

Segundo (Jurafsky, 2008), o que distingue os aplicativos de processamento de linguagem de outros sistemas de processamento de dados é o uso do conhecimento da linguagem. O conhecimento da linguagem necessário para se envolver no comportamento complexo da linguagem pode ser separado em seis categorias distintas:

- **Fonética e fonologia**: O estudo dos sons linguísticos
- **Morfologia**: O estudo dos componentes significativos das palavras
- **Sintaxe**: O estudo das relações estruturais entre palavras.
- **Semântica**: O estudo do significado.

- **Pragmática**: O estudo de como a linguagem é usada para atingir objetivos.
- **Discurso**: O estudo das unidades linguísticas maiores que um único enunciado.

No contexto desta pesquisa, o processamento de linguagem natural será utilizado para extrair atributos e informações de páginas Web, uma vez que são de natureza amplamente textual e devem possuir características importantes sobre o autor, corpo editorial, qualidade do texto, entre outros atributos que podem ser analisados para determinar a credibilidade.

A Tabela 2.2 relaciona os elementos de PLN que serão utilizadas no processo de extração de atributos.

Tabela 2.2: Definições da PLN

Elemento	Descrição
Tokens	Lista de palavras e pontuação. Omite espaços em branco, quebra de linha e linhas em branco.Exemplo: ["Eles", "se", "recusam", "a", "nos", "permitir","."]
Stopwords	Corpo de palavras irrelevantes, isto é, palavras de alta frequência como, por exemplo, "para"e "também"que, por vezes, são retiradas de um documento antes de continuar o processamento. *Stopwords* geralmente têm pouco conteúdo léxico e sua presença em um texto não consegue distinguir de outros textos.
Normalização	Transformação de todas as letras maiúsculas em minúsculas.
Dicionário	Criação de uma lista de palavras únicas.
Tagging	O processo de classificar palavras em suas partes de fala e rotulá-las. Partes de fala também são conhecidos como classes de palavras ou categorias lexicais. Exemplo: [(Comprei, VB),(uma, PR), (casa, S)] onde VB significa verbo, PR - pronome, S - substantivo.
Chunking	Subconjunto de tokens. Usado para reconhecimento de entidades. Exemplo: Nós vimos um cachorro amarelo. Chunk: ["Nós", "um cachorro amarelo"].
Stemming	Consiste em reduzir uma palavra ao seu radical. Exemplo: Paisagem - pais; copiar - copi.
Lematização	Reduz a palavra ao seu lema, que é a forma no masculino e singular. No caso de verbos, o lema é o infinitivo. Por exemplo, as palavras “gato”, “gata”, “gatos” e “gatas” são todas formas do mesmo lema: “gato”. Igualmente, as palavras “tiver”, “tenho”, “tinha”, “tem” são formas do mesmo lema “ter”.
Frequência de palavras	Quantidade de cada palavra no texto.
Bigramas	Lista de pares de palavras. Por exemplo: ["mais é", "do que", "disse então", "é dito"].
Collocations	Sequência de palavras que ocorrem juntas com frequência. Por exemplo: ["Estados Unidos", "anos atrás"].

2.4 Considerações sobre o Capítulo

Neste Capítulo foram vistos os conceitos necessários para a compreensão desta proposta. Primeiro foram apresentados os tipos de credibilidade, bem como a distinção sobre verdade e

credibilidade. Os três tipos de credibilidade vistos foram: fonte, mídia e mensagem. O tipo de credibilidade focada nesta dissertação é a da mensagem. Dentro dela, é possível definir quatro categorias: credibilidade presumida, da superfície, adquirida e de renome. A categoria que melhor se encaixa nesta pesquisa de credibilidade de mensagem é a credibilidade da superfície, pois se baseia em atributos percebidos no conteúdo.

Para entender como será a fase de avaliar o modelo, os conceitos de aprendizagem de máquina foram abordados, assim como seus tipos: aprendizagem supervisionada, não supervisionada e por reforço. A descrição das principais técnicas de aprendizagem, como SVM, Random Forest e Árvore de Decisão, foram explicadas. Nesta pesquina as três técnicas serão utilizadas para o tipo de aprendizagem supervisionada.

Por fim, para extrair atributos é necessários conhecer algumas técnicas de processamento de linguagem natural - PLN. Essa técnicas são fundamentais para tratar os elementos textuais das notícias, assim como sintaxe e morfologia.

Capítulo 3

Trabalhos Relacionados

Neste Capítulo são apresentadas as abordagens existentes no processo de verificação de credibilidade. Os trabalhos são agrupados em abordagens que usam Aprendizagem de Máquina, as baseadas no feedback do usuário e uma abordagem híbrida.

3.1 Abordagens baseadas em Aprendizagem de Máquina

Horne (2018) criaram uma ferramenta, chamada NELA, que extrai informações do título e do conteúdo de uma página Web. A ferramenta é modular, baseada em *pipelines* e permite que pesquisadores adicionem novas funcionalidades. Para construir sua base, os autores coletaram dados de notícias em páginas Web por 6 meses, totalizando 92 fontes e 136 mil artigos. Os atributos extraídos pela NELA estão listados na Tabela 3.1.

Tabela 3.1: Atributos utilizados na pesquisa de Horne (2018)

Categoria	Atributos
Complexidade e estilo de escrita	Diversidade léxica; legibilidade SMOG[1]; número de *stopwords*; números de sinais de exclamação; tamanho médio de caracteres em uma palavra; etc.
Sentimentos e emoções	Análise de sentimento; número de palavras positivas, negativas e neutras; entre outros.
Atributos tendenciosos	Número de palavras tendenciosas; probabilidade de subjetividade; número de palavras opinativas (positivas e negativas); etc.
Características de psicologia	Número de palavras de certeza; número de palavras de dúvidas; etc.
Recursos de engajamento	Número de compartilhamentos, comentários e reações no Facebook.
Funcionalidades da fundaçao moral	Atributos morais retirados de um léxico (por exemplo, justiça, fidelidade, autoridade, prejuízo e degradação)
Recursos de parte da fala	Números de verbos, advérbios, pronomes, interjeições, etc.
Outros tópicos específicos	Número de palavras religiosas; número de palavras de tempo (fim, até, temporada); etc.

A partir desse conjunto de dados gerais de notícias, quatro módulos foram desenvolvidos. O módulo de predição de notícias confiáveis é um classificador *Random Forest* treinado com notícias rotuladas. O segundo módulo, chamado de viés ou subjetividade, é composto de dois classificadores independentes: um classificador *Random Forest,* treinado em atributos baseados em conteúdo para prever artigos com tendências políticas, e um classificador *Naive Bayes*,

treinado em sentenças rotuladas objetivas e subjetivas. O terceiro é o módulo de predição de interesse comunitário, que prevê quais grupos online estão interessados em um artigo usando comunidades de notícias no reddit.com. O último módulo analisa as notícias em nível de origem, em vez de nível do artigo, ou seja analisa a fonte de onde veio a informação.

Na implementação do módulo 1, o classificador foi treinado com 4504 artigos e testado com 1130 artigos, alcançando 0,89 ROC/AUC. Na implementação do módulo 2, em Random Forest, o classificador foi treinado com 6158 artigos e testado com 1539 artigos, alcançando 0,92 ROC/AUC. Em naive Bayes, alcança 92% de precisão usando validação cruzada. Na implementação do módulo 3, três classificadores binários foram usados, cada classificador foi treinado em 2000 artigos e testado cada um em 500 artigos, obtendo 0,77 ROC/AUC em média. O último módulo não possui dados de implementação.

O trabalho de Wawer (2014) possui o foco em prever a credibilidade em conteúdos textuais em páginas Web. Para tanto, emprega a ferramenta *General Inquirer*, que consiste de uma aplicação e um dicionário com 11767 significados de palavras mapeadas em 183 categorias. A lista de categorias inclui, por exemplo, categorias baseadas em tópicos (política, economia, religião), categorias relacionadas à emoção (prazer, dor, sentimentos ou excitação). A associação à categoria é binária: as palavras pertencem a uma categoria ou não. Cada documento da base gera um vetor de 183 números, representados pela categoria. O método foi testado em dois cenários de classificação e um cenário de regressão.

Para avaliar o desempenho da abordagem foi usada uma base de dados construída pela Microsoft, que consiste de 1000 URLs e suas taxas (*ratings*) de credibilidade. Os atributos utilizados estão listados na Tabela 3.2. Foram definidas cinco categorias que exibem conteúdos Web credíveis e não credíveis, usando uma escala Likert de 5 pontos, onde 1 significa não credível e 5 significa muito credível. Duas configurações foram aplicadas: (1) verificar se uma página Web é credível ou não. Nesse caso a verificação de credibilidade é tratada como um problema de classificação binária; (2) verificar o nível de credibilidade em uma página Web em uma escala *likert* de 5 pontos, que é tratado como um problema de regressão. Na configuação de regressão, as medidas de desempenho utilizadas foram R^2, RMSE (*root mean square error*), MAE (*mean absolute error*), Evar (*explained error*). Apenas os valores de classificação serão apresentados pois é o método focado nesse trabalho.

Tabela 3.2: Atributos utilizados na pesquisa de (Wawer, 2014)

Categoria	**Atributos**
Texto	Número de: exclamação "!" no texto, vírgulas ","pontos ".", interrogação "?", Comprimento do texto como o número de palavras, polaridade 0 se a página for negativa, 1 se a página for positiva, Número de sentenças positivas, negativas, subjetivas, número de frases objetivas, Número de erros de ortografia, complexidade do texto pela entropia, exclusividade do conteúdo da página em relação a outras páginas, Medição estatística da legibilidade do texto (smog), categoria de categoria da Web e.g. Entretenimento, Negócios etc., número de substantivos no texto, verbos no texto, adjetivos, advérbios, determinantes.
Aparência	Número de anúncios na página da web, a área em pixels do maior anúncio, proporção da área de todos os anúncios para a área da página, número de definições de estilo CSS da página da Web.
Meta informação	tipo do domínio URL .gov, .edu, etc.
Popularidade social	Número de compartilhamentos do Facebook para uma URL da página da Web,Número de curtidas no Facebook para uma URL de página da web, Número de comentários no Facebook para uma URL de página da web,Número de cliques do Facebook para um URL da página da web,Total de compartilhamentos, curtidas, comentários e cliques do Facebook,Número Tweets mencionando um URL de página da web,Número de cliques em URL curtos para uma página da web,Número de sites com Bitly short URL para uma página da web, Número de favoritos deliciosos para um URL de página da web.
Popularidade Web	Alexa rank.
Estrutura do link	Número de linkagem para o website, Google PageRank.

Na classificação, há um experimento com três classes, onde os dados são divididos de acordo com sua fidedignidade. As pontuações abaixo de 2 são não-credíveis (164 casos), valor 3 é rotulado como verdade média (191 casos), pontuações maiores que 4 são marcadas como altamente confiável (524 casos). Três métricas diferentes são apresentadas: precisão, recall e f1 com os valores 0.66%, 0.65% e 0.65% respectivamente . Já no experimento com duas classes, os dados foram divididos em sites com pontuação menor que 2 (355 casos) e sites com pontuação maior ou igual a 4 foram marcados como altamente credível (524 casos). As métricas foram as mesmas (precisão, recall e f1), agora com os valores 0.62%, 0.73% e 0.67% para as classes de pontuação menor que 2. E 0.79%, 0.69% e 0.74% para as classes de pontuação maior ou igual a 4.

O CredEye é uma ferramenta, desenvolvida no trabalho de (Popat, 2018), que aceita uma declaração de linguagem natural como entrada do usuário e calcula sua avaliação de credibilidade junto com evidências como saída. Os autores desenvolveram e testaram três métodos para verificação de credibilidade: um pipeline de classificadores, um modelo de inferência e uma rede neural profunda baseada em LSTM (*Long short-term memory*). A arquitetura de pipeline teve o melhor desempenho e por isso o trabalho se concentrou nessa configuração. Seu núcleo é a análise da credibilidade da afirmação, com base na evidência geral ou na contra-evidência de um conjunto de artigos da Web recuperados automaticamente.

Os classificadores são treinados por aprendizado supervisionado usando 5.000 sentenças do site Snopes, cada uma identificada como verdadeira ou falsa, e 30 artigos Web recuperado para cada um deles. A arquitetura do sistema consiste nas seguintes etapas: (i) Recuperação de artigos de diversas fontes da Web, enviando o texto da afirmação para uma pesquisa motor, (ii) Detecção de Posturas para entender a postura de cada artigo, (iii) Análise de Conteúdo para entender a credibilidade de cada artigo, utilizando o estilo de linguagem e recursos relacionados à posição, (iv) Agregação de credibilidade para mesclar essas avaliações por artigo para calcular a pontuação geral da afirmação sendo verdadeira ou falsa, e (v) Extração de Evidência como apoio na forma de trechos informativos dos artigos relevantes da Web.

A acurácia das afirmações verdadeiras atingiram 83.20%, a acurácia das afirmações falsas atingiram 80.78% e a média macro atingiu 82%.

Tabela 3.3: Atributos utilizados na pesquisa de Popat (2018)

Atributos	**Descrição**
Atributos linguísticos	Verbos assertivos (66) e factivos (27) capturam o grau de certeza ao qual uma proposição se sustenta. Hedges (100) são as palavras atenuantes que suavizam o grau de compromisso com uma proposição. Palavras implícitas (32) acionam a pressuposição em uma enunciação. Os verbos de relatório (181) enfatizam a atitude em relação ao fonte da informação. Marcadores de discurso (13) capturam o grau de confiança, perspectiva e certeza nas declarações. Léxico de subjetividade e viés (8770) capta a atitude e emoções do escritor enquanto escrevia um artigo.
Atributos de confiança	PageRank AlexaRank

No trabalho de (Wahsheh et al., 2013) é proposta uma metodologia para avaliar a classificação de credibilidade e confiança de sites diferentes que aborda três (3) métricas principais da Web: reputação (estática e dinâmica), popularidade e métricas de spam (10 atributos). É feito um estudo de caso em um site governamental e de uma universidade.

Na pesquisa, 669 páginas foram coletadas da universidade de Jordânia e consideradas como páginas Web confiáveis, enquanto 1331 páginas de sites governamentais foram coletadas para teste. Os resultados dos algoritmos Naive Bayes e J48 são comparados, usando a ferramenta Weka, para avaliar a metodologia. Os autores descobriram que o algoritmo da árvore de decisão

(J48) é mais eficaz do que os Naive Bayes na determinação do grau de confiança e credibilidade para as páginas da Web.

O algoritmo Naive Bayes atinge 95,4412% de instâncias classificadas corretamente e 4,5588% de instâncias classificadas incorretamente. O algoritmo Árvore de Decisão (J48) alcança 99.7059% de instâncias classificadas corretamente e 0,2941% de instâncias classificadas incorretamente.

Tabela 3.4: Atributos utilizados na pesquisa de Wahsheh et al. (2013)

Atributos	**Descrição**
Métricas de reputação	Métricas fixas: nome do domínio .gov, .edu, etc Métricas dinâmicas: menções em redes sociais.
Métricas de popularidade do link	medida pelo número de páginas recuperadas da maioria das três principais pesquisas motores na Web (ou seja, Google, Bing e Yahoo).
Métricas de spam web	Palavras-chave sem significado (árabe / inglês). Taxa de compactação para páginas da Web. Número de imagens. Número de links de imagens. Comprimentos médios de palavras em árabe / inglês dentro das páginas da web. Comprimento da URL (número total de caracteres na URL). Número de links quebrados. Número de links redirecionados na página em consideração. Número de texto de link vazio (links sem texto âncora). Número de links vazios (texto âncora sem links).

3.2 Abordagens baseadas no feedback do usuário

(Yamamoto, 2017) propôs um sistema que sugere informações suspeitas, promovendo informações mais cuidadosas para os usuários durante a pesquisa na Web e navegação. Estuda a relação entre os avisos de credibilidade e o comportamento do usuário. Esse relacionamento tem implicações no design para aprimorar o envolvimento do usuário na busca cuidadosa de informações. O estudo centrou-se na sugestão de tópicos controversos como um tipo de recurso de ameaça. Os tópicos discutidos são aqueles que algumas pessoas afirmam serem suspeitas, independentemente dos tópicos serem realmente verdadeiros ou não.

O autor destaca que vários tipos de sistemas de suporte foram propostos para julgamento de credibilidade em informações da web. No entanto, as pessoas frequentemente não sentem a necessidade de sistemas de suporte. Assim, deve-se considerar como melhorar a busca cuidadosa de informações.

No estudo, há duas situações para sugerir tópicos controversos: verificar uma lista de resultados de pesquisa na Web e navegar em páginas da Web na lista. Um estudo de usuário on-line foi realizado, onde os participantes pesquisam informações de saúde usando sistemas experimentais com serviço de *crowdsourcing*.

O estudo revelou que o momento da sugestão do tópico pode influenciar os comportamentos de busca dos participantes. Isso significa que a sugestão de tópico antes de visitar as páginas da Web pode incentivar os pesquisadores a buscar informações com mais cuidado. Os resultados indicam que o foco no pré-alarme antes de acessar a informação é importante no projeto de sistemas de apoio para aumentar a disposição do pesquisador de se envolver na busca cuidadosa da informação.

Rejmund et al. (2014) trataram a credibilidade como uma função de distribuição de probabilidade e avaliaram a relação do valor de credibilidade com a similaridade semântica de sentenças. O trabalho criou um conjunto de sentenças com credibilidades conhecidas e então calculou a similaridade semântica de cada par de sentenças para tentar verificar a credibilidade de sentença usando uma sentença similar.

Para o trabalho foram selecionas 150 páginas Web, cada página foi pontuada 19 vezes por voluntários. O experimento é dividido em 3 estágios (A, B e C). No estágio A, os avaliadores dão um valor de credibilidade às páginas. No estágio B, os avaliadores respondem a questões da importância e credibilidade de sentenças extraídas do texto. No estágio C novamente os avaliadores dão credibilidade às páginas, com o objetivo de descobrir se a opinião mudou após as sentenças avaliadas.

A credibilidade foi avaliada em escala Likert de 5 pontos, onde 5 significa altamente credível e 1 significa altamente não credível. Devido ao fato de que, para uma dada informação, o valor da sua classificação de credibilidade depende de da pessoa que avalia, a credibilidade é aproximada por meios de distribuições de classificação.

3.3 Abordagens híbrida

Papaioannou and Aberer (2012) propuseram um sistema de recomendação social distribuído e colaborativo para avaliação da credibilidade de páginas Web. O sistema proposto trabalha com três componentes de avaliação. O primeiro deles é um componente de filtragem baseado na interpretação de usuários sobre uma página Web, chamado *Social Component*. O segundo é um componente de filtragem colaborativa baseado em itens importantes, que considerar os recursos do conteúdo da página Web chamado *Content Component*. O terceiro componente se baseia no ranqueamento da página Web nos resultados de consultas, chamado *Search-Ranking Component*. Os autores argumentam que a abordagem proposta é genérica o suficiente para incluir recursos adicionais de páginas Web.

O sistema propoosto assume que os usuários são conectados entre si por uma rede ponto a ponto (P2P), onde um nó P2P é um plug-in no navegador do usuário e, assim, consegue montar grafos sociais dos usuários e grafos de similaridade do conteudos (sejam elas vinculadas aos usuários ou explicitamente a outras página. Desta form, o *Social Component* estima a credibilidade de um página Web com base nas classificações dos relacionamento dos usuários sobre esta página. Já o *Content Component* avalia a credibilidade da página Web baseado em atributos do conteúdos da página, como semânticos (por exemplo, categoria, entidades, palavras-chave etc.), PNL (sentimentos, subjetividade, etc.), sintáticas (multiplicidades de tags em parte do discurso, sinais de pontuação, erros de ortografia etc.), anúncios, layout da página, entre outros. Por fim, o *Search-Ranking Component* estima a credibilidade da página Web com base algoritmo page-rankind do google. Ao final, todas essas estimativas são adequadamente combinadas em uma única avaliação baseada em pesos adaptativos.

O sistema foi implementado como plug-in do Firefox e foram usados dois conjuntos de dados: o Microsoft corpus **?**, que contém uma única avaliação por página Web, e o Reconcilie - um conjunto criado por pesquisadores do Instituto Japonês de Tecnologia da Informação (PJIIT) de Warszawa, no âmbito do projeto Reconcile - que agrega avaliações de 90 alunos do PJIIT que classificaram 9 páginas Web cada um a partir de 85 documentos poloneses (todos relacionados a tópicos de saúde). Como resultados, os autores apresentam apenas uma comparação do *Content Component*, em seus vários segmentos, com o classificador SVM, em termos de *recall* e precisão, bem como a curva AUC. Os autores sugerem que toda a experimentação realizada sugere que a abordagem proposta supera as abordasgens clássicas de avaliação de conteúdo.

3.4 Discussão

O artigo de Horne (2018) apresenta os resultados da primeira versão da ferramenta proposta e os autores afirmam que os dados continuarão sendo coletados para uso no kit de ferramentas

em seu lançamento posterior. A originalidade do trabalho em relação à credibilidade se deve ao fato de analisar a fonte da notícia, ou seja, a instituição geradora do conteúdo. A forma modular em que foi projetado faz com que o trabalho desenvolvido seja bastante abrangente, porém a falta de foco no estudo dos atributos de credibilidade pode ser apontada como desvantagem em relação a trabalhos que possuem essa preocupação.

As métricas do artigo de Wawer (2014) mostram que os resultados são aperfeiçoados usando a ferramenta de vocabulários General Inquirer nos dois cenários (regressão e classificação). Como trabalho futuro, o autor sugere avaliar se o foco em um assunto único (por exemplo, medicina) ou tipo de página Web (blogs, portais de notícias, comércio eletrônico) influenciará a precisão e o recall e qual seria a eficiência dos mesmos recursos linguísticos para sites escritos em idiomas diferentes do inglês. A abordagem baseada em texto apesar de ser promissora não chega a atingir 70% de acurácia na classificação em 2 classes.

De acordo com Popat (2018), uma das limitações do CredEye é a falta de compreensão profunda do escopo exato e tom mais refinado de afirmações. As evidências e contra-evidências são recuperadas automaticamente de páginas Web, porém essa atividade de recuperar páginas suficientes é outro problema onde há dependência dos resultados dos mecanismos de busca.

Wahsheh et al. (2013), em sua proposta, visam como trabalho futuro aumentar o número de métricas de credibilidade Web e construir um sistema para calcular o nível de credibilidade. Sua principal contribuição foi a organização e propostas de características. Apenas dois algoritmos foram utilizados Nayve Bayes e Árvore de Decisão. O estudo de caso não é amplo, dado que foi realizado apenas em um site do governo e em um site de universidade, além do mais o número de amostras é muito baixo, tornando a pesquisa bastante limitada.

Sobre o método utilizado para a coleta de dados, Horne (2018) e Wahsheh et al. (2013) foram os únicos entre os autores que montaram sua própria base manualmente. Wawer (2014) utilizou uma base da Microsoft. Popat (2018) utilizou uma base pronta a partir de um site que coletou dados da universidade. Em todos os trabalhos, os atributos foram divididos em categorias. A Tabela 3.5 apresenta uma comparação entre as características principais encontradas nos artigos de aprendizagem de máquina.

Tabela 3.5: Trabalhos Relacionados com Aprendizagem de Máquina

Autor	**Técnica de AM**	**Métrica**
Horne et al. (2018)	Random Forest	ROC/AUC
	Naive Bayes	Acurácia
Wawer (2014)	Regressão linear	R^2, RMSE, MAE, EVar
	Classificação	Precisão, Recall, medida F
Popat et al. (2018)	Regressão logística	Acurácia
Wahsheh et al. (2013)	Naive Bayes	TP, FP, precisão, recall, medida F, ROC/AUC
	Árvore de Decisão J48	

Já os trabalhos de Yamamoto (2017) e Rejmund et al. (2014) dependem fundamentalmente do julgamento e experiência do usuário. Estes não precisaram de uma base de páginas Web rotuladas, considerando o fato de que a classificação é feita pelo próprio usuário. O trabalho de Yamamoto (2017) forneceu uma solução visual de credibilidade, melhorando a maneira em que os resultados da web são exibidos. É um modelo de julgamento de credibilidade baseada na predição do usuário. A desvantagem encontrada neste trabalho, por meio de análise de log e questionário, é que se os usuários não são familiarizados com os tópicos de pesquisa, eles não podem julgar e usar as informações que o sistema fornece melhor do que se eles estiverem familiarizados com

os tópicos. No trabalho de Rejmund et al. (2014) foram formuladas 3 hipóteses referente à semelhança textual e credibilidade, a partir dessas hipóteses conclui-se que a credibilidade da sentença não depende do contexto, pois a sentença pode ser tratada como uma parte atômica da página Web, frases semelhantes não tem a mesma credibilidade, porém a similaridade pode prever a credibilidade. Sua contribuição agregou conhecimento teórico, mas não disponibilizou artefatos práticos para serem incorporados em outras soluções.

No trabalho de Papaioannou and Aberer (2012) foi proposto um sistema de avaliação de credibilidade decentralizada formado por três componentes, que, assim como a abordagem baseada no *feedback* do usuário, é preciso contar com a boa intenção dos usuários na classificação da notícia. Porém, o trabalho inovou e dividiu o comportamento dos usuários em perfil confiável e perfil malicioso. A abordagem de filtro colaborativa desenvolvida foi comparada a uma abordagem de aprendizagem de máquina, entretanto, apenas um classificador foi utilizado, podendo levar a um resultado tendencioso.

Por fim, ao realizar uma comparação entre os trabalhos descritos e o desta dissertação pode-se afirmar a não possibilidade de reprodução dos experimentos apresentados, devido à indisponibilidade de bases ou falta das APIs empregadas. Porém, percebe-se o diferencial na avaliação dos atributos deste trabalho com os trabalhos relacionados. Um destaque é a tentativa de criar um modelo de avaliação para páginas em português, introduzindo uma base de URL rotuladas para esse idioma.

3.5 Considerações finais

Este Capítulo relatou uma descrição de trabalhos encontrados na área de credibilidade de informação agrupados em três abordagens previamente identificadas em uma revisão sistemática da literatura: aprendizagem de máquina, *feedback* do usuário e abordagens híbridas.

Observa-se que entre as abordagens que utilizam aprendizagem de máquina é preciso definir as classes de saída, atributos que contribuem para avaliação de credibilidade e mecanismos para extraí-los, base de dados e métricas de análise do modelo. As métricas de avaliação geralmente são as mesmas e nota-se uma variação na configuração dos modelos em questão de classificadores ou pode haver um conjunto de módulos de classificadores com diferentes finalidades. As bases de dados frequentemente são extraídas de sites de verificação de notícias conhecidos.

Nas abordagens de feedback do usuário, há a necessidade de um esforço manual para uma eficiente confiabilidade dos resultados e a garantia de eficácia é baixa visto que cada ser humano possui seu ponto de vista acerca da percepção de credibilidade, sua abrangência também é baixa levando em consideração o baixo número de páginas avaliadas comparadas com abordagens que são feitas de forma automática.

Independente da abordagem, propriedades semelhantes comumente são encontradas para definir a credibilidade de notícias, como a fonte da notícia, a linguagem usada pelo autor, gramática correta e assim por diante. Neste contexto, espera-se absorver as vantagens e facilidades das técnicas empregadas atualmente, principalmente aprendizagem de máquina, para unir na solução proposta de uma avaliação de credibilidade de informações na Web.

Capítulo 4

Atributos de Credibilidade

Este Capítulo descreve os atributos, identificados na literatura, capazes de auxiliar na identificação da credibilidadde de informações encontradas na Web. Além disso, também é apresentada uma metodologia para coleta, extração e análise dos atributos.

4.1 Atributos

Segundo (Olteanu, 2013), os usuários podem ter expectativas diferentes sobre as normas e estilos de escrita, o que pode afetar sua percepção e, assim, a interpretação da credibilidade do conteúdo da Web. Nesta pesquisa, os atributos baseados em texto se enquadram perfeitamente neste problema. Já Liu (2013) afirmam que a análise de links de páginas da Web tem sido amplamente utilizada na área de pesquisa na Web. Os autores acreditam que os atributos relacionados à estrutura de links podem ser um indicador promissor para a avaliação da credibilidade da Web.

Por esta razão, nesta dissertação decidiu-se dividir os atributos extraídos em duas categorias: atributos de conteúdo e atributos de rede. Foram selecionados 14 atributos de conteúdo e 3 atributos de rede a partir de diversos artigos.

4.1.1 Atributos de Conteúdo

Os atributos de conteúdo são relacionados com a notícia principal do site, incluem elementos textuais, sintáticos e quantitativos. Tais atributos são:

Autor: O atributo autor é baseado nos artigos de (Aggarwal, 2014; Horne, 2018; Horne and Adali, 2017; Pattanaphanchai et al., 2013) e contém a informação sobre autor do documento, página ou site. Seu valor é representado por 1 (quando existe um autor) ou 0 (quando não existe autor).

Quantidade de sinais no título: O atributo quantidade de sinais no título é encontrado nos trabalhos de (Papaioannou and Aberer, 2012; Liu, 2013; Olteanu, 2013) e auxilia na análise de padrões de escrita. Papaioannou and Aberer (2012) dividem os atributos nas categorias sintáticos, léxicos, semânticos, PLN e Rank de pesquisa. Sendo a quantidade de sinais um atributo sintático. Segundo Olteanu (2013), o uso não padronizado de gramática e pontuação é considerado um bom indicador de conteúdo de baixa qualidade e baixa percepção de credibilidade.

Quantidade de sinais no corpo: Assim como o atributo anterior, a quantidade de sinais no corpo da informação é encontrado em (Papaioannou and Aberer, 2012; Liu, 2013; Olteanu, 2013) e representa a quantidade de sinais de pontuação. Pode auxiliar na análise de padrões de escrita.

Tamanho do Título: Modificado do trabalho de (Olteanu, 2013), representa a quantidade de caracteres que compõem o título da página. Assim como já foi dito, o uso não padronizado de gramática é considerado um bom indicador de conteúdo de baixa qualidade e baixa percepção de credibilidade.

Análise de Sentimento ou Subjetividade: Atributo retirado dos artigos de (Aggarwal, 2014; Liu, 2013; Papaioannou and Aberer, 2012). Frases subjetivas geralmente se referem à opinião pessoal, emoção ou julgamento, enquanto o objetivo se refere à informação factual. A subjetividade é um *float* que fica no intervalo de [0,1], onde 0 significa muito objetivo e 1 muito subjetivo.

Diversidade léxica: Encontrada nos artigos de (Horne, 2018; Horne and Adali, 2017), a diversidade léxica (*Type-token Ratio* - TTR) indica o número médio de vezes que uma determinada palavra foi repetida em todo o texto, representando assim a diversidade de palavras (quantidade de palavras) únicas encontradas. Pode ser obtida pela razão entre o número total de palavras e palavras únicas.

Legibilidade: É o cálculo sobre o quão claro um texto é. Há várias formas de calcular a legibilidade: FORCAST, McLaughli's SMOG, Fry readability graph, The Gunning fog, The Dale-Chall, The Flesch, Flesch-Kincaid (FK), entre outros (Cavaco, 2010). Nesta proposta, será adotado o índice de legibilidade SMOG e legibilidade FK. A medida de **legibilidade SMOG** (*Simple Measure of Gobbledygook*) realça a interação entre o texto e um grupo de leitores com características conhecidas, tais como a competência de leitura, o conhecimento prévio e a motivação. Segundo Cavaco (2010) tem uma fórmula, proposta por Harry G. McLaughlin em 1969, que estima os anos de escolaridade no contexto educacional Norte-Americano necessários para compreender completamente um excerto escrito.

A fórmula SMOG é descrita por:

$$1040 * \sqrt{(número\_de\_polissilabos * (30/número\_de\_frases) + 3.1291)}$$

Onde 3.1291 e 1040 são constantes da fórmula.

Na Tabela 4.1 são apresentados os valores de SMOG, a sua correspondência aos níveis de escolaridade nos E.U.A. e a correspondência destes com os níveis de escolaridade em Portugal. Esse atributo é encontrado nos trabalhos de (Horne, 2018; Horne and Adali, 2017; Olteanu, 2013).

Quantidade de verbos no título: Calcula, para cada título, a quantidade de verbos. Foi usado em (Liu, 2013). Horne and Adali (2017) afirmam que títulos que possuem muitos verbos e nomes de entidade são mais propensos a serem falsos.

Quantidade de verbos no corpo: Calcula, para cada artigo, a quantidade de verbos no corpo da notícia. Foi usado em (Liu, 2013).

Quantidade de palavras no título: Modificado do trabalho de Olteanu (2013), este atributo descreve o comprimento do título. Em outras palavras, calcula a quantidade de palavras

Tabela 4.1: Tabela SMOG

Índice *SMOG*	Escolaridade E.U.A	Escolaridade Portugal
0 - 6	*Low-literate*	Competência básica (ler e escrever)
7	*Junior high school*	7º, 8º e 9º ano de escolaridade
8	*Junior high school*	-
9	*Some high school*	10 º ano de escolaridade
10	*Some high school*	11 º ano de escolaridade
11	*Some high school*	12 º ano de escolaridade
12	*High school graduate*	Terminou ensino secundário
13 - 15	*Some college education*	Frequência do ensino superior
16	*University degree*	Conclusão do curso universitário
17 - 18	*Post-graduate studies*	Estudos pós graduados sem grau acadêmico
19	*Post-graduate degree*	Todos os graus acadêmicos pós licenciatura ou mestrado integrado

no título.

Quantidade de palavras no corpo: Olteanu (2013) descreve como o comprimento do texto. Calcula para cada artigo a quantidade de palavras no corpo.

Quantidade de imagens: Representa o número de imagens embutidas na notícia. Retirado dos artigos de (Wahsheh et al., 2013; Putri Ghaisani et al., 2017).

Quantidade de vídeos: Atributo sugerido no artigo de (Putri Ghaisani et al., 2017), como um fator determinante de credibilidade, representa o número de vídeos embutidos na notícia.

Ano de publicação: Representa o ano em que a notícia foi publicada. Pattanaphanchai et al. (2013) citam este atributo como data de publicação.

A descrição de todos os atributos de conteúdo encontra-se na Tabela 4.2.

4.1.2 Atributos de rede

Os atributos de rede são relacionados com o domínio do site, incluindo elementos sintáticos e quantitativos. Tais atributos são: autoridade do domínio, status e quantidade de sinais na URL.

Autoridade do domínio: Atributo encontrado em diversos artigos como (Schwarz, 2011; Wahsheh et al., 2013; Olteanu, 2013; Xu et al., 2011; Aggarwal, 2014). Autoridade do domínio é representado pelo sufixo do domínio na URL e representa o tipo de domínio. Segundos os autores listados, o tipo de domínio de uma página da Web pode sugerir que a página pertence ou não a um conjunto de páginas confiáveis. Por exemplo, Um domínio .gov indica que seu conteúdo é aprovado por uma instituição governamental enquanto um domínio .edu pertence a um grupo de instituição educacional (Wahsheh et al., 2013).

Comprimento da URL: Atributo retirado do trabalho de (Wahsheh et al., 2013). Conta quantos caracteres a URL possui.

Tabela 4.2: Atributos de conteúdo

Atributos	**Descrição**
Autor	Informação sobre autor
Quantidade de sinais no título	Quantidade de sinais de pontuação no título pode auxiliar na análise de padrões de escrita
Quantidade de sinais no corpo	Quantidade de sinais de pontuação no texto pode auxiliar na análise de padrões de escrita
Tamanho do Título	Quantidade de caracteres que compõem o título da URL. Pode auxiliar na detecção de títulos irregulares
Análise de Sentimento (Subjetividade)	Análise da subjetividade do texto, sendo mensurada no intervalo [0.0;1.0], onde 0 significa muito objetivo e 1 muito subjetivo.
Diversidade	Atributo que representa a diversidade palavras utilizadas no texto. Ou seja, a quantidade de palavras únicas encontradas. Intervalo [0;1]
Legibilidade SMOG e FK	Cálculo sobre o quão claro um texto está.
Quantidade de verbos no título	Calcula para cada título dos artigo a quantidade de verbos.
Quantidade de verbos no corpo	Calcula para cada artigo a quantidade de verbos no corpo da notícia.
Quantidade de palavras no título	Calcula para cada artigo a quantidade de palavras no título.
Quantidade de palavras no corpo	Calcula para cada artigo a quantidade de palavras no corpo.
Quantidade de imagens	Número de imagens embutidas na notícia
Quantidade de vídeos	Número de vídeos embutidos na notícia (Youtube e Vimeo)
Quantidade de links	Número de links no meio da notícia (Youtube e Vimeo)
Ano de publicação	Ano em que a notícia foi publicada
Quantidade de palavras únicas	Ocorrência de palavras únicas no texto.
Quantidade de palavras em maiúsculo	Quantidade de palavras com todos os caracteres em maiúsculo.
SuBJ/OBJ	Classificação do texto em subjetivo ou objetivo

Quantidade de sinais na URL: Atributo adaptado de (Papaioannou and Aberer, 2012; Liu, 2013; Olteanu, 2013). Conta quantos sinais de pontuação há na URL.

A descrição dos atributos de rede encontra-se na Tabela 4.3. Novos atributos foram identificados, adapatados e incluídos no modelo. Quantidade de sinais no título, quantidade de sinais no corpo e quantidade de imagens também foram calculados em porcentagem. Espera-se que a proporção da imagem em relação ao texto tenha algum equilíbrio, ou seja, seria inusitado uma notícia Web ter mais imagens do que texto.

Tabela 4.3: Atributos de rede

Atributos	Descrição
Autoridade do domínio	.gov, .com, .edu etc
Comprimento da URL	Quantidade de caracteres na URL
Quantidade de sinais na URL	Quantidade de sinais de pontuação na URL

4.2 Metodologia

Uma vez que o objetivo proposto nesta pesquisa é avaliar atributos, extraidos de páginas Web, que permitam mensurar a credibilidade de páginas Web, uma metodologia foi elaborada para auxiliar no processo de coleta, extração e avaliação dos atributos.

A metodologia proposta é ilustrada na Figura 4.1.

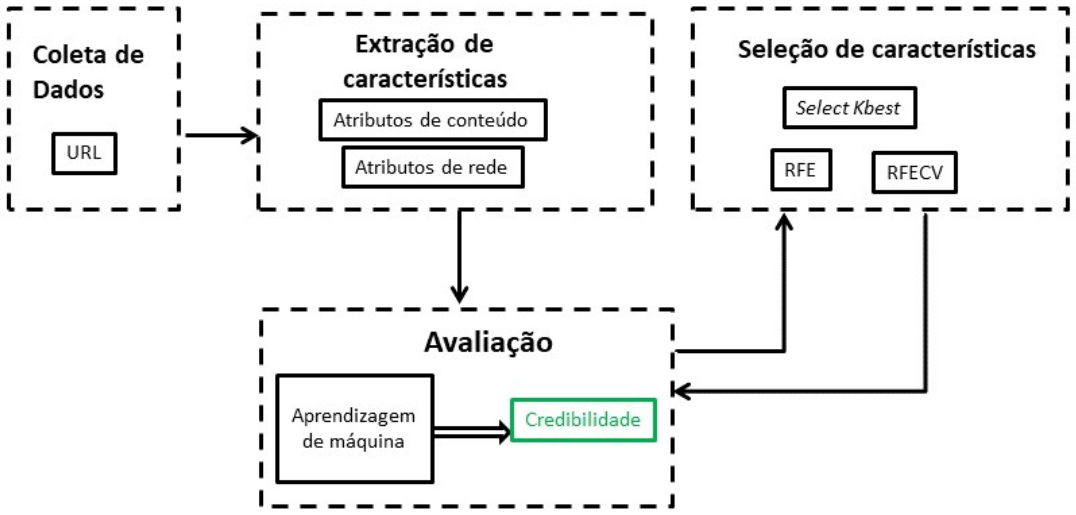

Figura 4.1: Etapas metodológicas
Fonte: Autor.

4.2.1 Coleta de dados

A **coleta de dados** é a etapa da metodologia que reúne dados de URL que serão classificadas de acordo com sua veracidade. Tal coleta é obtida através da busca e armazenamento do conteúdo de páginas Web. Para tanto, um processo de *Web crawling* é realizado, tendo como entrada diferentes URL (*seeds*). Existe no processo de coleta, Um componente para verificar se as URL coletadas são válidas ou não. Tal procedimento permite eliminar URL mal formadas, bem como entradas duplicadas e ausência futura de atributos.

4.2.2 Extração de Características

A **extração de características** consiste no uso de ferramentas, técnicas e scripts para extrair atributos. A extração visa separar os atributos de interesse (atributos de conteúdo e atributos

da rede).

Como mencionado anteriormente, os atributos de conteúdo são baseados em propriedades textuais da página Web. Grande parte desses atributos será extraído utilizando técnicas de processamento de linguagem natural (PLN) e scripts python. Os atributos de rede são características extraídas da URL. Tal extração será feita por scrips python.

As Tabelas 4.2 e 4.3 apresentam os atributos empregados nesta pesquisa.

4.2.3 Seleção de atributos

A **seleção de atributos** envolve definir quais atributos são bons candidatos de acordo com os métodos de seleção de atributos Select Kbest, RFE e RFECV. Antes de utilizar essas técnicas, um score de cada atributo é fornecido a fim de analisar a importância de cada atributo individualmente e comparar se os atributos classificados como mais importantes individualmente são os mesmos da seleção automática.

4.2.4 Avaliação do modelo

A etapa de **avaliação** é o componente principal do processo, onde as características extraídas são reunidas, preparadas e usadas pelo classificador de aprendizagem de máquina no treinamento e teste. Como resultado final, as URL avaliadas são categorizadas como credíveis ou não. Os resultados encontrados nos trabalhos relacionados da seção de aprendizagem de máquina poderão ser utilizados como métricas de comparação para avaliar se o modelo proposto nesta pesquisa atinge valores superiores de eficácia de acordo com as métricas de acurácia.

4.3 Considerações sobre o Capítulo

Neste capítulo, foi descrito a organização de extração de atributos e o porquê de ser dividido nas categorias de conteúdo e rede. Cada atributo de conteúdo foi especificado e em qual trabalho da literatura este foi utilizado, o mesmo foi feito para os atributos de rede. A escolha dos atributos como autor, atributos relacionados com o título e o corpo, imagens, vídeos, ano de publicação, baseou-se na frequência em que estes foram encontrados na literatura e no quantos eles são comuns em páginas Web. O atributos de rede são características adicionais e alternativas para deduzir credibilidade em notícias.

A metodologia também foi exposta neste capítulo, no qual foram detalhadas as suas etapas. Os dados atualmente são provenientes de bases de dados, porém, almeja-se que futuramete seja implementado o processo de Web crawling juntamente com um componente verificador de URL válida. Com os dados adquiridos, os atributos propostos desta pesquisa são extraídos. Uma avaliação feita com o conjunto completo de características é realizado para poder ser feita uma comparação posterior das métricas do modelo. A seleção de atributos visa empregar menos atributos e conquistar um valor próximo do valor de acurácia de todos os atributos juntos. Então a avaliaçao do modelo é realizada novamente.

Em cada etapa de seleção, seja de atributos ou de classificador, foi importante estudar cada parâmetro e cada técnica, para prover o melhor resultado de precisão possível. O tratamento dos dados também deve ser feito com cautela para evitar influências negativas.

Capítulo 5

Protocolo Experimental e Resultado

Neste Capítulo será apresentado o protocolo experimental (ambiente de experimentação e base de dados) utilizado na avaliação da metodologia proposta, bem como os resultados encontrados.

5.1 Protocolo Experimental

Para investigar a capacidade de validar/classificar um conjunto de atributos como credíveis é necessária a realização de vários experimentos com diferentes classificadores. Nesta dissertação foram escolhidos SVM, Árvore de Decisão e Random Forest, os classificadores mais empregdos encontrados na abordagem de aprendizagem de máquina vistos no Capítulo 3. Contudo, antes de apresentar esses resultados é preciso descrever o ambiente, as bases de dados, o processo de extração das características, as métricas de avaliação e os ajustes nos classificadores em questão. Esta seção descreve todo o protocolo experimental necessário para se atingir o objetivo proposto.

5.1.1 Ambiente de Experimentação

Os experimentos realizados na elaboração desta dissertação foram executados em um computador com sistema operaciona. Ubuntu 64 bits, 8GB de memória RAM, disco de 500GB e um processador Intel Core 5 de 2.5Gz. Para a execução dos algoritmos de classificação e análise foi utilizado o software Jupyter Notebook.

5.1.2 Base de Dados

Esta pesquisa utiliza duas bases retiradas do site kaggle[1]. A primeira se chama *Fake News Sample*[2] com 42220 valores. Destes, 13139 são sentenças verdadeiras e 29081 são sentenças falsas. Foram separadas 29554 para treino e 12666 para teste (70% e 30% da base respectivamente). A extração de característica é aplicada nesta base para obter os atributos desejados para análise de credibilidade desta pesquisa. Nesta base serão feitas todas as etapas de treino, validação e

[1] https://www.kaggle.com/
[2] https://www.kaggle.com/pontes/fake-news-sample/data

teste para descobrir quais os melhores parâmetros de cada classificador. A segunda se chama *Fake News Detection*[3]. Ela possui 3352 instâncias e com os seguintes dados: URL, título, texto e rótulo (Label), no qual 0 significa notícia falsa e 1 notícia verdadeira. Através da preparação de dados percebu-se que 1138 entradas estavam duplicadas, o que implicou em uma base de 2871 valores únicos, onde 1215 são sentenças verdadeiras e 1656 são sentenças falsas. Foram separadas 2009 para treino e 862 para teste (70% e 30% da base respectivamente).

A primeira base coleta artigos de 2003 a 2019, enquanto que a segunda possui artigos de 2012 a 2017. A atualidade da base pode influenciar na classificação, o comportamento dos atributos relativos à credibilidade talvez mude ao longo do tempo, os termos do momento podem ser alterados, porém este é um estudo futuro a se fazer.

A pesquisa levou em consideração 25 atributos: autor, quantidade de sinais no título, quantidade de sinais no corpo, tamanho do título, análise de sentimento ou subjetividade, diversidade léxica, legibilidade, quantidade de verbos no título, quantidade de verbos no corpo, quantidade de palavras no título, quantidade de palavras no corpo, quantidade de imagens, quantidade de vídeos, ano de publicação, domínio, quantidade de sinais na url, status.

5.2 Implementação

Na tabela 5.1, são descritas de modo breve as ferramentas utilizadas na implementação deste trabalho. Para facilitar a etapa de extração, a página HTML de cada URL é extraída, assim como a notícia principal da página Web. Por meio desses, todos os atributos foram extraídos a partir de bibliotecas do python e os scripts elaborados pelo ambiente Jupyter Notebook.

Tabela 5.1: Ferramentas utilizadas

Função/atributo	**Biblioteca**
Baixar HTML	Requests
Baixar artigo	Newspaper3k
Tipo de domínio	Tldextract
Quantidade de sinais na URL, no título e no corpo	Re.findall
Comprimento da URL, tamanho do título	len
Autor, ano de publicação, título, quantidade de vídeos	Article
Classificação	Sklearn
Diversidade léxica	Lexical_diversity
Quantidade de verbos no título e no corpo, normalização do texto	NLTK
Quantidade de imagens e links	BeautifulSoup
Legibilidade SMOG e FK	Readcalc
Índice de subjetividade, subjetivo/objetivo	Textblob

5.3 Escolha do Classificador e dos Atributos

Para obter o melhor resultado sobre o conjunto de dados para cada classificador, foram realizados treinamentos onde os valores dos principais parâmetros de cada classificador foram ajustados até a obtenção do valor mais adequado.

Esta seção apresenta esses resultados para o Random Forest, Árvore de Decisão e SVM. Os resultados alcançados levaram em consideração o texto sem *stopwords*, sem sinais e em minúsculo. Utilizou-se o método de treino, validação e teste na avaliação do modelo. Foi aplicado *hold out*

[3]https://www.kaggle.com/jruvika/fake-news-detection

para separar a base em 70% para treino e 30% para teste. A Validação Cruzada foi feita em 70% do treino em 5 combinações. Finalmente, foi realizado o teste nos 30% da base. As amostras do treino não balanceado possuem 20357 falsas e 9197 verdadeiras.

5.3.1 Random Forest

A Tabela 5.2 apresenta os parâmetros utilizados para ajustar o classificador Random Forest. O parâmetro *max_depth* representa a profundidade de cada árvore na floresta. Assim, quanto mais profunda a árvore mais ela se divide e captura mais informações sobre os dados. Já o parâmetro *n_estimators* define o número de árvores no classificador.

Tabela 5.2: Validação cruzada em Random Forest na base Fake News Sample

Teste	max_depth	n_estimators	acurácia	Tempo (em segundos)
1	30	100	95.02%	34.96
2	40	100	95.02%	34.94
3	**50**	**100**	**95.02%**	**34.33**
4	60	100	95.02%	34.08
5	30	200	95.20%	70.86
6	**40**	**200**	**95.21%**	**68.08**
7	50	200	95.21%	69.75
8	60	200	95.21%	70.53

Observa-se que o melhor valor de acurácia foi para a sexta configuração, onde a profundidade de árvore é 40 e o número de estimadores é 200. Porém na terceira configuração (50,100) o tempo de execução é reduzido para a metade, sendo assim, estes parâmetros foram selecionados para o teste.

Com esses valores de parâmetros definidos, a base *Fake News Sample* foi testada no classificador Random Forest, agora com *hold out* de 30% da base. A Tabela 5.3 apresenta o resultado obtido.

No resultado observado na Tabela 5.3, a **acurácia** alcançou 95.39%. Vale relembrar que a acurácia representa a proporção de casos que foram corretamente previstos, sejam verdadeiros positivos ou verdadeiros negativos. Já a **matriz de confusão** mostra que, do total do número de amostras de teste (12666), o modelo acerta 8507 e erra 217 da classe verdadeira, enquanto erra 371 e acerta 3571 da classe falsa. Esse valor representa um maior número de verdadeiros positivos e verdadeiros negativos. A **precisão** obteve 0.95 e representa um cálculo sobre todas as classificações corretas dada pelo modelo, quantas realmente eram corretas. A **sensibilidade** ou *recall* obteve 0.94 e representa a proporção de casos positivos que foram identificados corretamente. A **medida F** obteve 0.95 e representa a média harmônica entre precisão e recall. Por fim o ROC é criado a partir da taxa de verdadeiro positivo contra os falsos positivos e o AUC que varia de 0 a 1 é o resumo do ROC em um único valor, Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1. No modelo desevolvido, a medida **ROC/AUC** possui o valor de 0.94.

Tabela 5.3: Hold out com Random Forest na base *Fake News Sample*

Teste	max_depth	n_estimators	acurácia	matriz	precisão	recall	medida F	ROC/AUC
1	50	100	95.39%	$\begin{pmatrix} 8513 & 211 \\ 373 & 3569 \end{pmatrix}$	0.95	0.94	0.95	0.94

Além da acurácia, foi realizada uma verificação da importância de cada atributo no classificador. Para tanto, foi utilizada a funcionalidade *feature_importances_* para classificadores baseados em árvore, da biblioteca *scikit learn*, e empregada no Jupyter Notebook. Vale destacar que o *feature_importances_* retorna um *array* onde cada elemento dele é uma $feature$ do seu modelo. Ele informa, em proporções, quão importante aquela $feature$ é para o modelo, onde quanto maior o valor, mais importante a $feature$ é para o modelo. A ideia é visualizar os atributos mais importantes, o que permite analisar posteriormente essas características em páginas Web e assim definir quais as mais indicadas para identificar credibilidade.

A Tabela 5.4 exibe a importância dos atributos em ordem decrescente.

Tabela 5.4: Importância dos atributos em Random Forest base *Fake news Sample*

Atributo	**importância**
quantidade de imagens	0.238271
quantidade de links	0.168421
autor	0.143172
quantidade de imagens (em porcentagem)	0.057525
quantidade de sinais na url	0.037896
ano de publicação	0.028733
quantidade de links (em porcentagem)	0.024261
diversidade	0.023845
quantidade de palavras em maiúsculo	0.023652
legibilidade smog	0.022315
subjetividade	0.020935
quantidade de palavras únicas	0.020087
domínio	0.019868
quantidade de palavras no corpo	0.018985
quantidade de sinais no corpo	0.018895
quantidade de vídeos embutidos	0.018816
quantidade de sinais no título	0.018280
quantidade de verbos no corpo	0.018176
quantidade de palavras no título	0.017892
quantidade de sinais no corpo (em porcentagem)	0.016054
quantidade de sinais no título (em porcentagem)	0.011391
quantidade de verbos no título	0.007631
tamanho do título	0.006772
comprimento da url	0.006154
classificação do texto em subjetivo ou objetivo	0.005837

Observa-se que o atributo **quantidade de imagens** foi o mais importante no modelo e a **classificação do texto em subjetivo/objetivo** foi o menos importante na avaliação das credibilidades de página Web. Assim, a *quantidade de imagens* deveria ser o foco principal de análise na página e o atributo texto subjetivo/objetivo não deveria ganhar tanta atenção.

5.3.2 Árvore de Decisão

A Tabela 5.5 apresenta os parâmetros utilizados para ajustar o classificador Árvore de Decisão. O parâmetro *max_depth*, assim como no classificador Random Forest, representa a profundidade máxima da árvore, ou quantas perguntas a árvore faz até tomar uma decisão. É importante ressaltar que a profundidade não pode ter um valor elevado, pois isto provocaria um modelo viciado, onde o treino acerta bastante, mas o teste não gera bons resultados para novos dados.

Observa-se que o melhor valor de acurácia foi a primeira configuração, onde a profundidade de árvore é 10 comprovando que a acurácia vai diminuindo a medida que a profundidade

Tabela 5.5: Validação cruzada em Árvore de Decisão base *fake news sample*

Teste	max_depth	acurácia	Tempo em segundos
1	**10**	**93.11%**	**1.83**
2	20	92.08%	2.81
3	30	91.85%	3.03
4	40	91.85%	3.02
5	50	91.85%	3.00
6	60	91.85%	3.29
7	70	91.85%	3.05
8	80	91.85%	3.1

aumenta. Com esse parâmetro definido, a base *Fake News Sample* foi testada no classificador Árvorde de Decisão, agora com *hold out* de 30% da base. A Tabela 5.6 apresenta o resultado obtido.

Tabela 5.6: Hold out na árvore de decisão base fake news sample

Teste	max_depth	acurácia	matriz de confusão	precisão	recall	medida F	ROC/AUC
1	10	93.60%	$\begin{pmatrix} 8334 & 390 \\ 421 & 3521 \end{pmatrix}$	0.93	0.92	0.93	0.92

No resultado observado na Tabela 5.6, a **acurácia** alcançou 93.60%. Já a matriz de confusão mostra que, do total do número de amostras de teste (12666), o modelo acerta 8334 e erra 390 da classe verdadeira, enquanto erra 421 e acerta 3521 da classe falsa. Isto representa um maior número de verdadeiros positivos e verdadeiros negativos. A **precisão** obteve 0.93. O *recall* obteve 0.92. A **medida F** obteve 0.93 e representa a média harmônica entre precisão e recall. A medida **ROC/AUC** obteve 0.92.

Até aqui o classificador Random Forest dispõe o melhor desempenho em comparação à Árvore de Decisão.

Por fim, a Tabela 5.7 exibe, em ordem decrescente, a importância de cada atributo no classificador. Observa-se que o atributo mais importante e o menos importante foram os mesmos encontrados em Random Forest, que são **quantidade de imagens** e **classificação do texto em subjetivo/objetivo** respectivamente, variando apenas alguns atributos da Tabela.

5.3.3 SVM

Diferente dos classificadores anteriores, o SVM foi ajustado em relação parâmetro C que determina um ponto de equilíbrio razoável entre a maximização da margem e a minimização do erro de classificação. O kernel utilizado foi o "rbf", um kernel padrão de uso para o SVM. Além disso, o parâmetro de folga γ (*gamma*), que é um coeficiente para controlar o raio do kernel RBF, foi fixado em 0.001. Quanto maior o valor de γ mais tentará se ajustar exatamente ao conjunto de dados de treinamento, isto é, erro de generalização e causará problemas de ajuste excessivo. Por fim, o parâmetro C, que determina um ponto de equilíbrio razoável entre a maximização da margem e a minimização do erro de classificação, foi testado para se definir o melhor valor. A Tabela 5.8 aponto os resultados do ajuste do parâmetro C na base *fake news sample*.

Observa-se que o melhor valor de acurácia foi a oitava configuração, onde o parametro C é 400, porém o tempo de execução é muito alto, assim o parâmetro adotado foi o de 70 (segunda configuração). Com o valor do parâmetro C definido, a base *Fake News Sample* foi testada no classificador SVM, agora com *hold out* de 30% da base (Tabela 5.9).

Tabela 5.7: Importância dos atributos na árvore de decisão base *fake news sample*

Atributo	importância
quantidade de imagens	0.488830
quantidade de links	0.337496
ano de publicação	0.079917
tamanho do título	0.014143
comprimento da url	0.010276
quantidade de vídeos embutidos	0.009517
autor	0.008066
quantidade de imagens (em porcentagem)	0.006520
quantidade de palavras em maiúsculo	0.006160
diversidade	0.005364
legibilidade smog	0.004508
subjetividade	0.004381
quantidade de links (em porcentagem)	0.003509
quantidade de sinais no título (em porcentagem)	0.003402
quantidade de verbos no corpo	0.003106
quantidade de sinais no corpo (em porcentagem)	0.002964
quantidade de sinais na url	0.002749
quantidade de sinais no corpo	0.002540
quantidade de palavras únicas	0.002134
domínio	0.001072
quantidade de palavras no corpo	0.001034
quantidade de sinais no título	0.000894
quantidade de verbos no título	0.000533
quantidade de palavras no título	0.000505
classificação do texto em subjetivo ou objetivo	0.000383

Tabela 5.8: Validação cruzada em SVM base fake news sample

Teste	Regularizador C	acurácia	Tempo em segundos
1	60	87.41%	187.37
2	**70**	**87.56%**	**171.18**
3	80	87.67%	171.26
4	90	87.74%	171.86
5	100	87.83%	171.84
6	200	88.21%	192.91
7	300	88.46%	217.83
8	**400**	**88.55%**	**257.94**

Tabela 5.9: Hold out em SVM base *fake news sample*

Teste	Regularizador C	acurácia	matriz de confusão	precisão	recall	medida F	ROC/AUC
1	70	87.69%	$\begin{pmatrix} 8061 & 663 \\ 896 & 3046 \end{pmatrix}$	0.86	0.85	0.85	0.85

No resutlado da Tabela 5.9, a **acurácia** obtida foi 87.69%, a menor entre os classificadores. A matriz de confusão mostra que, do total do número de amostras de teste (12666), o modelo acerta 8061 e erra 663 da classe verdadeira, enquanto erra 896 e acerta 3046 da classe falsa. A **precisão** obteve 0.86. O *recall* obteve 0.85. A **medida F** obteve 0.85 e representa a média harmônica entre precisão e recall. A medida **ROC/AUC** obteve 0.85.

Por fim, este classificador não possui parâmetro para listar a importância dos atributos.

5.3.4 Melhor classificador

Após a realização dos experimentos com o ajuste de parâmetros, é possível realizar a comparação entre os classificadores Random Forest, Árvore de Decisão e SVM nas duas bases. Na Tabela 5.10 é possível observar qual o melhor parâmetro encontrado e a acurácia equivalente na base *fake news sample*. Nota-se que o classificador Random Forest obteve o melhor resultado de avaliação do modelo. A Figura 5.1 ilustra as curvas ROC/AUC da base *Fake news sample*, onde AUC obteve um valor de 0.94 muito proximo do valor desejado (1).

Tabela 5.10: Melhor classificador da base fake news sample

Classificador	Parâmetros	Acurácia
Random Forest	max_depth = 50, estimators=100	95.39%
Árvore de Decisão	max_depth = 10	93.60%
SVM	C=70	87.69%

Analisando as particularidades dos algoritmos, o **SVM** é muito bom quando em casos onde o número de características é maior que o número de amostras. Não funciona bem quando o conjunto de dados tem muitos ruídos, ou seja, onde as classes estão sobrepostas ou quando o conjunto de dados é muito grandes, pois exige inversão de matriz, aumentando a complexidade computacional. Por essas razões, é possível compreender porque este algoritmo não teve o melhor desempenho.

As regras geradas pela **árvore de decisão** são fáceis de interpretar, entretanto, é um algoritmo que sofre com o problema de sobreajuste (*overfitting*). O problema *overfitting* ocorre quando o modelo se adaptou muito bem aos dados com os quais está sendo treinado, porém, não generaliza bem para novos dados.

Random Forest, trabalha com diferentes tipos de entrada, tais como binárias, categóricas ou numéricas. Possui métodos para equilibrar erros em conjuntos de dados onde as classes são desequilibradas. Como é um conjunto de árvore de decisões, possui no geral as mesmas vantagens e desvantagens, contudo, escolhe-se o resultado da melhor àrvore.

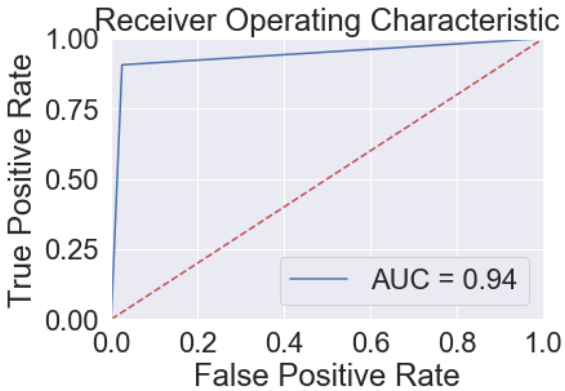

Figura 5.1: ROC/AUC base fake news sample
Fonte: Autor.

5.3.5 Escolha dos Atributos

Após a etapa de escolha do classificador é o momento de selecionar os melhores atributos. De acordo com (Kohavi and John, 1997) há duas abordagens principais para a seleção de atributos: Filtro e *Wrapper*. A abordagem filtro introduz um processo de separação, que ocorre antes da aplicação do algoritmo de aprendizagem propriamente dito. Em outras palavras, a idéia é separar ("filtrar") atributos irrelevantes, segundo algum critério, antes do aprendizado ocorrer. A abordagem *wrapper* ocorre externamente ao algoritmo básico de aprendizagem, mas utiliza este algoritmo como uma espécie de "caixa preta" para analisar, a cada iteração, o subconjunto de atributos até ali selecionados. Em outras palavras, os métodos *wrapper* geram um subconjunto candidato de atributos selecionado do conjunto de treinamento, e utilizam a precisão resultante do algoritmo de aprendizado para avaliar o subconjunto de atributos em questão.

A partir do resultado do melhor classificador, foram selecionados três métodos de seleção de características da abordagem wrapper: *Select Kbest*, Seleção RFE e Seleção RFECV. O método Select Kbest (Scikit-learn, 2020) é uma função do *sklearn*, onde se informa quantas características devem ser selecionadas, as melhores são escolhidas por base em um teste estatístico. A ideia básica do método da eliminação recursiva de características, denominado de *Recursive Feature Elimination* (RFE) é que a cada passo do processo, um número fixo de componentes é eliminado e o classificador é retreinado. A eliminação recursiva de uma característica por vez gera um classificador com menos erros esperados, quando comparada à remoçãoo de mais de uma caracterítica ao mesmo tempo (Villela et al., 2011). A Seleção RFECV é a eliminação recursiva de características por validação cruzada. Nesta técnica não definimos o número de características, mas sim o próprio modelo retorna qual a melhor configuração.

Na Tabela 5.11 é possível observar quais foram os atributos selecionados de acordos com cada técnica de seleção de feature para a base *Fake News Sample*. Foram definidos 5 atributos para o método select kbest e seleção RFE. Como vimos a seleção rfecv define seu próprio número de características. Os 5 melhores atributos escolhidos foram: comprimento da url, ano de publicação, quantidade de imagens, quantidade de imagens em porcentagem em relação ao tamanho da página e quantidade de links. O método seleção RFE foi melhor que seleção kbest por usar a mesma quantidade de atributos e resultar no maior valor de acurácia. Já o método seleção RFECV optou por usar todo o conjunto de atributos.

Tabela 5.11: Métodos de classificação de atributos na base fake news sample

Método	**Atributos**	**Acurácia**	**matriz de confusão**
Select Kbest	comprimento da url, ano de publicação, quantidade de imagens, quantidade de imagens (em %), quantidade de links (%)	91.89%	$\begin{pmatrix} 8268 & 456 \\ 571 & 3371 \end{pmatrix}$
Seleção RFE	comprimento da url, ano de publicação, quantidade de imagens, quantidade de imagens (em %), quantidade de links	94.64%	$\begin{pmatrix} 8459 & 292 \\ 387 & 3528 \end{pmatrix}$
Seleção RFECV	todos	95.18%	$\begin{pmatrix} 8525 & 226 \\ 384 & 3531 \end{pmatrix}$

5.4 Avaliação e Resultados

Após verificar quais os melhores parâmetros dos classificadores, adotou-se o classificador Random Forest para ser aplicado na segunda base.

Na Tabela 5.12 é possível observar os resultados antes de utilizar as técnicas de seleção de atributos. Empregando todos os atributos a acurácia obtida foi de 91.42%. A base *fake news detection* como vimos, possui 862 amostras de teste, a matriz de confusão mostra que 326 e 462 foram os valores das classificações corretas das classes verdadeira e falsa respectivamente, enquanto que 39 e 35 foram os números das classificações incorretas das classes verdadeira e falsa respectivamente. O valor da medida ROC/AUC também foi muito próximo do desejável, este reflete que o valor de verdadeiros positivos e negativos é elevado.

Tabela 5.12: Random Forest na base Fake News Detection

Teste	max_depth	n_estimators	acurácia	matriz	ROC/AUC
1	50	100	91.42%	$\begin{pmatrix} 326 & 39 \\ 35 & 462 \end{pmatrix}$	0.91

Na Tabela 5.13 é possível observar os 5 melhores atributos com o método Select Kbest, e seleção RFE. Comparando novamente os métodos select kbest e seleção RFE, mais uma vez o método de seleção RFECV se sobressaiu. Já na seleção RFECV, observou-se um resultado interessante em que a acurácia foi a maior entre os métodos retornando 17 atributos como mais importantes para o modelo. Nota-se que a acurácia foi maior com 8 atributos a menos: dominio, quantidade de sinais do titulo, legibilidade fk, quantidade de imagens, quantidade de verbos no corpo, subj/obj, % sinais no titulo e % sinais no corpo.

Tabela 5.13: Métodos de classificação de atributos na base fake news detection

Método	Atributos	Acurácia	matriz de confusão
Select Kbest	ano de publicação,quantidade de sinais no corpo, quantidade de palavras únicas, quantidade de palavras no corpo e quantidade de links	70.65%	$\begin{pmatrix} 242 & 123 \\ 130 & 367 \end{pmatrix}$
Seleção RFE	comprimento da url, ano de publicação, tamanho do título, legibilidade smog, quantidade de sinais no corpo	90.60%	$\begin{pmatrix} 318 & 50 \\ 31 & 463 \end{pmatrix}$
Seleção RFECV	17 atributos: quantidade de sinais na url, comprimento da URL, autor, ano de publicação, tamanho do titulo, quantidade de sinais no corpo, subjetividade, legibilidade fk, legibilidade smog, quantidade de palavras em maiusculo, quantidade de palavras únicas, quantidade de palavras no título, quantidade de palavras no corpo, quantidade de vídeo, quantidade de link, quantidade de verbos no título, diversidade	92%	$\begin{pmatrix} 345 & 23 \\ 46 & 448 \end{pmatrix}$

5.5 Avaliando URLs

Uma vez que todo o processo de avaliação até agora foi realizado em uma base de língua inglesa, empregou-se o modelo em testes com páginas Web em Português e Inglês. Foram selecionadas dez (10) URLs em cada língua, sendo 5 verdadeiras e 5 falsas. Vale ressaltar que a classificação

do texto em português não reconhece corretamente as classes gramaticais, dado que as bases utilizadas foram de lingua inglesa.

As Tabelas 5.14 e 5.15 listam as URLs em português e em inglês, respectivamente, classificadas como verdadeiras e utilizadas para testar o classsificador e os artibutos selecionados.

Tabela 5.14: URLs em português verdadeiras e falsas selecionadas para o experimento

#	Classificação	URL
1	Verdadeira	https://g1.globo.com/mundo/noticia/2020/02/29/por-que-o-coronavirus-levou-a-falta-de-papel-higienico-no-japao.ghtml
2	Verdadeira	https://www.opovo.com.br/noticias/politica/2020/03/01/pms-aceitam-nova-proposta-do-governo-e-motim-termina.html
3	Verdadeira	https://www.tecmundo.com.br/ciencia/150850-sinais-vida-nasa-descobre-caverna-subterranea-marte.htm
4	Verdadeira	https://noticias.r7.com/tecnologia-e-ciencia/facebook-fecha-escritorio-apos-diagnostico-de-coronavirus-05032020
5	Verdadeira	https://www.uol.com.br/tilt/noticias/redacao/2020/03/06/iphones-lentos-geram-multa-a-apple-no-mundo-mas-caso-se-arrasta-no-brasil.htm
6	Falsa	https://www.dolanguenews.com.br/curitibano-registra-o-nome-de-seu-filho-de-ameno/
7	Falsa	https://minilua.com/assustador-ventriloquo-que-utilizava-corpo-crianca-como-boneco/?fbclid=IwAR2tzOqY4_v-STrQEB2uV5RFT2CXQvUhC5WxEhG3OSNcEdtupTu2pgt1dQ
8	Falsa	https://gazetamt.net/2020/01/13/prefeito-coloca-cameras-no-quarto-para-filmar-fantasmas-e-acaba-pegando-a-esposa-com-amantes/
9	Falsa	http://www.receitasnaturais.com.br/especialistas-alertam-para-risco-de-cancer-causado-for-fones-de-ouvido-sem-fio/
10	Falsa	https://bocadopovonews.com.br/aviao-cai-e-todos-os-missionarios-a-bordo-sobreviveram/

Tabela 5.15: URLs em inglês verdadeiras e falsas selecionadas para o experimento

#	Classificação	URL
1	Verdadeira	https://www.nytimes.com/2020/03/02/well/live/coronavirus-spread-transmission-face-touching-hands.html
2	Verdadeira	https://finance.yahoo.com/news/protein-found-scorpion-venom-just-042141124.html
3	Verdadeira	https://techxplore.com/news/2020-03-closer-batteries.html
4	Verdadeira	https://www.cnet.com/news/these-high-tech-contacts-may-help-correct-color-blindness/
5	Verdadeira	https://www.cnbc.com/2020/03/06/coronavirus-cases-surpass-100000-worldwide.html
6	Falsa	:https//www.express.co.uk/news/science/1249990/Asteroid-warning-NASA-tracks-4KM-killer-asteroid-hit-Earth-end-civilisation-asteroid-news
7	Falsa	https://perma.cc/8JBN-XN87
8	Falsa	https://thereisnews.com/dog-travels-more-than-100-km-to-bite-its-owner-after-being-abandoned/
9	Falsa	https://newspunch.com/keanu-reeves-breaking-free-matrix/
10	Falsa	https://www.dailymail.co.uk/sciencetech/article-7626887/Instagram-Facebook-ban-sexual-emojis-including-eggplant-peach.html

O modelo aplicado (classificador Random Forest e seus parâmetros de configuração) foram aplicados nas URLs listadas nas Tabelas 5.14 e 5.15, onde foram extraídos 17 atributos de cada URL de acordo com o método RFECV.

Tabela 5.16: Classificação de URL em português

Atributos	**URL1**	**URL2**	**URL3**	**URL4**	**URL5**	**URL6**	**URL7**	**URL8**	**URL9**	**URL10**
1. Qtde de sinais na URL	23	21	16	16	27	16	16	23	20	15
2. Comprimento da URL	115	111	98	111	133	81	152	125	114	85
3. Autor	N	S	S	N	S	N	S	S	N	S
4. Ano de publicação	2020	2020	0	2020	2020	2020	0	2020	2019	0
5. Tamanho do título	63	53	58	57	75	50	73	92	78	54
6. Qtde de sinais no corpo do artigo	4	14	9	4	3	10	12	5	5	4
7. Indice de Subjetividade	0	0	0	0	0	0	0	0	0	0.33
8. Legibilidade FK	13	9.58	14.16	21.9	9.78	9.78	13.44	23.33	24.33	0.98
9. Legibilidade SMOG	13.02	11.21	16.53	18.24	12.46	12.46	13.82	22.08	22.08	11.21
10. Qtde de palavras em MAIUSCULO	2.78	12.12	9.38	6.67	9.68	11.43	3.85	3.7	7.41	3.23
11. Qtde de palavras únicas	34	27	30	17	29	31	26	25	25	27
12. Qtde de palavras no título	12	9	9	7	13	9	12	15	13	9
13. Qtde de palavras no corpo do artigo	36	33	32	30	31	35	26	27	27	31
14. Qtde de vídeo embutido	0	1	0	0	0	0	0	0	0	1
15. Qtde de links	0	0	0	0	0	0	0	0	0	0
16. Qtde de verbos no título do artigo	1	0	1	0	2	1	0	3	1	1
17. Diversidade	0.97	0.91	0.97	0.77	0.97	0.94	1	0.96	0.96	0.94
Credibilidade	V	F	V	V	V	V	F	F	V	V

Em relação ao idioma português (Tabela 5.16), a **quantidade de sinais na URL** ficou no intervalo de 15-27. O **comprimento da URL** situou-se no intervalo 81-152. No campo **autor**, 6 URLs possuem autor e 4 não. O **ano de publicação** são dos anos 2019 e 2020, onde o valor 0 significa que a notícia não possui ano de publicação. O **tamanho do título** varia entre 50-92. **Quantidade de sinais no corpo** do artigo varia entre 3 e 14. **Índice de subjetividade** tem o valor 0 na maioria dos atributos. **Legibilidade FK** intervalo 0-21. **A legibilidade SMOG** varia entre 11 e 22. **Quantidade de palavras em maiúsculo** variou entre 2 e 12. **Quantidade de palavras únicas** variou de 17 a 34. A **quantidade de palavras no corpo** variou de 26 a 36. A **quantidade de vídeos embutidos** é no máximo 1. Nenhum **link** foi encontrado nas URL. **Quantidade de verbos no título** varia de 0 a 3. E por fim, a **diversidade** varia de 0.94 a 1.

O mesmo comportamento foi observado na Tabela 5.17, no idioma inglês, em que não é possível dividir o intervalo de valores dos atributos para verdadeiro e falso. Em relação ao valores, a **quantidade de sinais na URL** ficou no intervalo de 6-18. O **comprimento da URL** situou-se no intervalo 56-136. No campo **autor**, 8 URLs possuem autor e 4 não. O **ano de publicação** são dos anos 2018 a 2020, onde o valor 0 significa que a notícia não possui ano de publicação. O **tamanho do título** varia entre 24-133. **Quantidade de sinais no corpo** do artigo varia entre 4 e 14. **Índice de subjetividade** tem o valor 0.16 a 0.59. **Legibilidade FK** intervalo 10-22. **A legibilidade SMOG** varia entre 12 e 18. **Quantidade de palavras em maiúsculo** variou entre 0 e 14. **Quantidade de palavras únicas** variou de 15 a 32. A **quantidade de palavras no corpo** variou de 4 a 22. A **quantidade de vídeos embutidos** é no máximo 1. Nenhum **link** foi encontrado nas URL. **Quantidade de verbos no título** varia de 0 a 3. E por fim, a **diversidade** varia de 0.86 a 1.

5.5.1 Nova coleta de URLs

Uma base com 200 URL brasileiras foi coletada, sendo 100 amostras da classe verdadeira (V) e 100 amostras da classe falsa (F), chamada de *Credibilidade de URL brasileira*[4]. A rotulagem foi feita baseada em notícias retiradas de sites de verificação de fatos, como o e-farsas, e a verificação de notícias do *Google* e de sites confiáveis. As Tabelas 5.18 e 5.19 possuem alguns exemplos com as primeiras 20 URL de cada classe.

Das 200 URL, o modelo proposto usando a técnica RFE acertou 63 classificações verdadeiras e 63 classificações falsas. Já usando a técnica RFECV, acertou 65 classificações verdadeiras e 52 classificações falsas. Este resultado mostra que os 5 atributos escolhidos na RFE são mais relevantes do que todos os atributos empregados no RFECV, especialmente em URLs classificadas como falsas. Este resultado mostra que esse estudo sobre classificação ainda está em aberto.

5.6 Análise da avaliação do modelo

O resultado da Credibilidade no final da Tabela 5.16, percebeu-se que os intervalos de atributos entre URLs verdadeiras e falsas se sobrepuseram muitas vezes, o que pode ter deixado o modelo confuso na avaliação da classe. As 5 primeiras URLs da Tabela 5.16 deveriam ser "V"e as 5 últimas "F", mas isso não foi visto. O modelo clasifica bem os verdadeiros positivos, mas precisa melhorar os verdadeiros negativos. Analisando as 2009 amostras de treino, verificou-se que 1159 eram instâncias verdadeiras e 850 instâncias falsas, o que pode ter ocasionado na preferência do modelo em rotular novas URL como verdadeiras. É comum o número de

[4]https://www.kaggle.com/ecosta/credibilidade-de-url-brasileiras

Tabela 5.17: Classificação de URL em inglês

Atributos	**URL1**	**URL2**	**URL3**	**URL4**	**URL5**	**URL6**	**URL7**	**URL8**	**URL9**	**URL10**
1. Qtde de sinais na URL	17	13	10	16	14	22	6	18	10	18
2. Comprimento da URL	101	79	57	84	79	136	26	93	56	122
3. Autor	S	S	S	S	S	S	N	N	S	S
4. Ano de publicação	2020	2020	0	0	2020	2020	0	2018	2018	2019
5. Tamanho do título	24	81	57	57	66	89	133	68	58	82
6. Qtde de sinais no corpo do artigo	13	6	4	4	10	14	3	6	8	7
7. Indice de Subjetividade	0.53	0.16	0.59	0	0.25	0.2	0.5	0.7	0.26	0.55
8. Legibilidade FK	12.63	21.42	14.8	13.9	21.1	11.47	10.43	13.33	21.78	22.55
9. Legibilidade SMOG	12.16	14.55	17.12	13.82	18.24	13.82	8.84	14.55	15.9	18.24
10. Qtde de palavras em MAIUSCULO	0	0	0	0	6.67	14.71	0	10.34	3.85	0
11. Qtde de palavras únicas	32	28	27	26	28	28	15	25	22	26
12. Qtde de palavras no título	4	15	8	8	10	14	22	13	10	13
13. Qtde de palavras no corpo do artigo	34	28	27	35	30	34	17	29	26	30
14. Qtde de vídeo embutido	0	0	0	0	0	0	0	0	1	0
15. Qtde de links	0	0	0	0	0	0	0	0	0	0
16. Qtde de verbos no título do artigo	0	2	0	2	2	3	2	3	0	1
17. Diversidade	0.97	1	1	0.86	0.97	0.91	0.94	0.93	0.92	0.93
Credibilidade	V	F	V	V	V	F	V	V	V	F

Tabela 5.18: URLs em português verdadeiras selecionadas para o experimento

#	Classificação	URL
1	Verdadeira	https://g1.globo.com/mundo/noticia/2020/02/29/por-que-o-coronavirus-levou-a-falta-de-papel-higienico-no-japao.ghtml
2	Verdadeira	https://www.opovo.com.br/noticias/politica/2020/03/01/pms-aceitam-nova-proposta-do-governo-e-motim-termina.html
3	Verdadeira	https://www.tecmundo.com.br/ciencia/150850-sinais-vida-nasa-descobre-caverna-subterranea-marte.htm
4	Verdadeira	https://noticias.r7.com/tecnologia-e-ciencia/facebook-fecha-escritorio-apos-diagnostico-de-coronavirus-05032020
5	Verdadeira	https://www.uol.com.br/tilt/noticias/redacao/2020/03/06/iphones-lentos-geram-multa-a-apple-no-mundo-mas-caso-se-arrasta-no-brasil.htm
6	Verdadeira	https://noticias.uol.com.br/ultimas-noticias/agencia-brasil/2020/05/26/voluntarios-consertam-mais-de-mil-respiradores-e-devolvem-a-hospitais.htm
7	Verdadeira	https://g1.globo.com/rj/rio-de-janeiro/noticia/2020/05/26/paulo-marinho-presta-novo-depoimento-na-pf-do-rio-nesta-terca-feira.ghtml
8	Verdadeira	https://g1.globo.com/natureza/noticia/2020/05/26/amazonia-perdeu-em-media-21-mil-hectares-de-floresta-por-dia-em-2019-aponta-levantamento.ghtml
9	Verdadeira	https://g1.globo.com/economia/tecnologia/noticia/2020/05/26/tres-em-cada-quatro-brasileiros-tem-acesso-a-internet-e-mais-da-metade-das-classes-baixas-esta-conectada-aponta-pesquisa.ghtml
10	Verdadeira	https://economia.uol.com.br/noticias/estadao-conteudo/2020/05/26/petrobras-anuncia-alta-de-5-no-preco-da-gasolina-e-de-7-no-diesel-na-refinaria.htm
11	Verdadeira	https://www.bbc.com/portuguese/brasil-52801691
12	Verdadeira	https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2020/05/27/esteira-de-r-44-mil-que-governo-quer-comprar-e-de-excelente-nivel-diz-mourao.htm
13	Verdadeira	https://www1.folha.uol.com.br/poder/2020/05/bolsonaro-convoca-ministros-para-discutir-reacao-ao-stf-e-aliados-defendem-lei-de-abuso.shtml
14	Verdadeira	https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/05/27/governo-de-sp-autoriza-reabertura-de-shoppins-com-restricoes-na-capital.htm
15	Verdadeira	https://economia.uol.com.br/noticias/redacao/2020/05/26/aneel-decide-nao-cobrar-tarifa-extra-de-energia-ate-dezembro-de-2020.htm
16	Verdadeira	https://noticias.uol.com.br/politica/ultimas-noticias/2020/05/27/quem-governa-e-a-familia-nao-bolsonaro-diz-marinho.htm
17	Verdadeira	https://noticias.uol.com.br/colunas/josias-de-souza/2020/05/27/wanderson-havia-uma-pedra-no-meio-do-caminho.htm
18	Verdadeira	https://economia.uol.com.br/faq/auxilio-emergencial-prorrogacao-meses-parcelas-r-600.htm
19	Verdadeira	https://noticias.uol.com.br/cotidiano/ultimas-noticias/2020/05/27/ap-capitao-e-militares-responderao-por-homicidio-em-naufragio-que-matou-42.htm
20	Verdadeira	https://entretenimento.uol.com.br/noticias/redacao/2020/05/27/estudio-uol-em-casa-apresenta-melim.htm
...	Verdadeira	...

verdadeiros positivos ser maior, em comparação com o mundo real, é muito difícil que as classes sejam balanceadas. Outra possível causa dos erros de modelo é em relação ao ano das notícias coletadas comparadas ao ano das URLs aleatórias selecionadas. A credibilidade de cinco atrás talvez tenha um comportamento diferente da credibilidade do ano atual.

Tabela 5.19: URLs em português falsas selecionadas para o experimento

#	Classificação	URL
1	Falsa	https://www.dolanguenews.com.br/curitibano-registra-o-nome-de-seu-filho-de-ameno/
2	Falsa	https://minilua.com/assustador-ventriloquo-que-utilizava-corpo-crianca-como-boneco/?fbclid=IwAR2tzOqY4_v-STrQEB2uV5RFT2CXQvUhC5WxEhG3OSNcEdtupTu2pgt1dQ
3	Falsa	https://gazetamt.net/2020/01/13/prefeito-coloca-cameras-no-quarto-para-filmar-fantasmas-e-acaba-pegando-a-esposa-com-amantes/
4	Falsa	http://www.g17.com.br/caso-neymar-traz-a-importancia-de-namorar-com-testemunhas
5	Falsa	http://www.g17.com.br/lula-pede-aumento-da-pena-para-nao-sair-para-trabalhar
6	Falsa	http://www.g17.com.br/bolsonaro-nao-sera-multado-por-transportar-filhos-sem-cadeirinha
7	Falsa	https://www.sensacionalista.com.br/2020/05/02/uso-de-mascara-tambem-e-recomendado-para-o-queixo-nao-cair-toda-vez-que-bolsonaro-fala/
8	Falsa	https://www.sensacionalista.com.br/2020/04/24/eu-nao-tenho-que-pedir-autorizacao-de-ninguem-para-violar-a-lei-diz-bolsonaro/
9	Falsa	https://www.sensacionalista.com.br/2020/04/24/bolsonaro-fala-por-uma-hora-e-esquece-de-renunciar/
10	Falsa	https://www.sensacionalista.com.br/2020/04/24/paulo-guedes-era-o-unico-de-mascara-em-discurso-porque-e-o-proximo-a-ir-para-a-rua/
11	Falsa	http://www.g17.com.br/annita-e-scoby-doo-estao-juntos
12	Falsa	https://conexaoufo.com.br/piramides-encontradas-na-russia-seriam-as-mais-antigas-do-mundo/
13	Falsa	https://noticiaspt.net/2020/05/18/pescador-da-florida-perdido-no-mar-por-14-dias-afirma-que-foi-violado-por-sereias/
14	Falsa	https://fatosdesconhecidos.ig.com.br/hackers-podem-ter-gravado-voce-durante-visita-ao-xvideos/
15	Falsa	https://www.joselitomuller.com/lula-pede-perdao-pelo-vacilo-foi-mal-tava-biritado/
16	Falsa	https://www.joselitomuller.com/homem-investe-duzentos-mil-e-confeccao-de-mascaras-mas-so-vende-quatro-e-vai-a-falencia/
17	Falsa	https://veja.abril.com.br/blog/sensacionalista/governo-manda-abrir-saloes-para-maquiar-contas-do-cartao-corporativo/
18	Falsa	https://www.ovnihoje.com/2019/11/17/keanu-reeves-acha-que-a-humanidade-esta-prestes-a-se-libertar-da-matriz/
19	Falsa	https://extra.globo.com/noticias/bizarro/fabricabritanica-lanca-cerveja-feita-de-viagra-para-celebrar-casamento-real-1626632.html
20	Falsa	https://enfu.com.br/ingestao-de-esperma-nas-primeiras-horas-da-manha-acelera-o-emagrecimento/
...	Falsa	...

5.7 Considerações Finais

Neste Capítulo, o desempenho de três classificadores foram estudados, avaliando acurácia, matriz de confusão, precisão, recall e ROC/AUC.

Sem utilizar os métodos automáticos de seleção de atributos, em Random Forest, os atributos "quantidade de imagens", "quantidade de links"e "autor"foram os atributos mais importantes dos 25 atributos extraídos obtendo 95.36% de acurácia, enquanto que em árvore de decisão os atributos mais importantes em ordem de importância foram "quantidade de imagens", "quantidade de links"e "ano de publicação"obtendo 93.60% de acurácia, os atributos de SVM não foram possíveis de analisar devido o método ser caixa-preta, sabe-se apenas o valor da acurácia

de 88.50%.

Dentre os três classificadores, o melhor resultado obtido para avaliação da credibilidade foi o Random Forest, este foi selecionado para continuar a avaliação dos atributos de credibilidade.

Com os métodos Select Kbest e Seleção RFE utilizou-se apenas cinco atributos: comprimento da url, ano de publicação, quantidadede imagens, quantidade de imagens, quantidade de link. A acurácia obtida para cada foi 91.87% e 93.60% respectivamente, um bom resultado reduzindo significativamente a quantidade de atributos. Nota-se que tanto atributo de rede quanto atributo de conteúdo foram importantes para a avaliação de credibilidade. A seleção RFECV não descartou nenhum atributo do modelo.

Após utilizar a segunda base, a acurácia obtida antes da seleção de atributos foi de 91.42%. A seleção RFECV retornou um valor maior de 91.76% removendo 8 atributos. Isso mostra a importância do método de seleção e que para cada base os melhores atributos variam. A seleção Kbest e seleção RFE definiram atributos diferentes da primeira base, porém trouxe a mesma conclusão de que é possível reduzir a quantidade de atributos e obter um resultado próximo de quando se utiliza todos os atributos extraídos.

Capítulo 6

Conclusão

As dificuldades que as notícias falsas causam em diversos contextos na política, na obtenção diária de informação, na verificação de qualidade de notícia foram contempladas e motivaram a conceber o desenvolvimento desta proposta. Como estímulo também há o fator de querer se concentrar em um trabalho que avalie atributos que sejam associados ao julgamento de credibilidade no contexto de uma página .

Os trabalhos relacionados com credibilidade foram brevemente descritos sobre o que faziam, organizados em qual abordagem pertenciam, as métricas utilizadas foram descritas para que baseadas nelas muitos mecanismos estivessem disponiveis e o modelo proposto seja bem avaliado por eles. Dentre essas métricas estão: acurácia, matriz de confusão, medida F, precisão, recall e ROC/AUC. Além disso investigou-se a presença de todos os atributos empregados nestes trabalhos. Verificou-se que os trabalhos encontrados na literatura que envolvem aprendizagem de máquina não se dedicam a elementos textuais e de domínio simultâneamente. Notou-se que a quantidade de atributos empregadas também é pequena dada a dificuldade de obtenção destes.

Três abordagens foram elaboradas: abordagens baseadas em aprendizagem de máquina, abordagens baseadas no feedback do usuário e abordagens híbridas. Sendo a abordagem baseada em aprendizagem de máquina o foco deste trabalho. Nesta abordagem, o classificador de aprendizagem de máquina Random Forest foi o melhor em relação à Árvore de Decisão e SVM. Três métodos de seleção automática de atributos foram utilizados, percebeu-se que é possível extrair atributos e reduzir a quantidade deles selecionando apenas os mais importantes de acordo com a base empregada. Para tanto, quanto mais atributos extraídos, melhor a chance de encontrar quais atributos irão fornecer o melhor grupo de características de acordo com a base.

6.1 Contribuições Alcançadas

Esta dissertação apresentou uma solução para a falta de método eficaz para verificar a credibilidade de páginas Web por meio da criação de um modelo de avaliação utilizando aprendizagem de máquina que define atributos relevantes de credibilidade para serem testados em relação à confiabilidade da informação.

Além do modelo proposto, como contribuição foi elaborada uma organização do trabalho em três grupos: abordagens baseadas em aprendizagem de máquina, abordagens baseadas no *feedback* do usuário e abordagens híbridas. As abordagens baseadas em aprendizagem de

máquina têm o aspecto de fornecer resultados exatos enquanto as baseadas no feedback do usuário tem uma natureza mais empírica do usuário.

Foram apresentadas comparações entre os métodos de seleção de atributos Select kbest, seleção RFE e seleção RFECV. Fixando o número de atributos dos métodos select kbest e seleção RFE, constatou-se que a acurácia da seleção RFE frequentemente é superior. Contudo o método de seleção RFECV sempre é o maior entre os 3, neste método a quantidade de atributos é selecionada automaticamente. Já em URL coletadas aleatoriamente, percebeu-se que a técnica RFE classificou muito bem as URL falsas, mesmo utilizando apenas 5 atributos.

Uma base de páginas em português inicialmente foi desenvolvida e disponibilizada, dado que nenhuma base neste idioma foi encontrada na literatura, espera-se que esta contribuição sirva como apoio para que seja incrementada em trabalhos futuros.

6.2 Dificuldades Encontradas

O tratamento de cada base utilizada neste trabalho exigiu um esforço considerado, visto que para o aprendizado é necessário dados sem ruídos para uma boa classificação. Encontrar uma base rotulada com um número de amostras elevado foi uma das adversidades encontradas, além do mais, nenhuma base em português foi encontrada. A classificação de páginas em português é um trabalho futuro a ser considerado.

O valor padrão dos atributos de cada classe verdadeira ou falsa ainda não foi capaz de ser definido, por exemplo, ainda não se sabe para qual comprimento de url uma notícia pode ser considerada credível ou nao, assim como para todos os outros atributos.

6.3 Trabalhos Futuros

A identificação de novas características ainda pode contribuir para resultados cada vez mais satisfatórios, especialmente os atributos relacionados à rede para ter uma quantidade próxima aos atributos de conteúdo. Novos métodos de seleção de atributos também podem ser incorporados ao estudo para avaliar os fatores determinantes na classificação da página e seus resultados serem alvo de comparação com os métodos atuais.

Posteriormente, seria interessante usar técnicas de reamostragem para balancear as classes e melhorar as predições, como exemplo tem-se a abordagem de *sampling* a qual visa minimizar a discrepância entre as classes. A implementação de outros algoritmos de aprendizagem de máquina também podem ser realizadas. Tendo exemplo O *Random Forest* utilizando neste trabalho, que é um método *ensemble* do tipo *bagging*, no qual as àrvores são treinadas em paralelo, poderiam também ser testado o método ensemble do tipo *boosting*, neste as arvores são treinadas de um modo sequencial.

O estudo de termos temporais da base também se encaixa como trabalho futuro, na base utilizada neste trabalho foram coletadas notícias de 2003 a 2019, são 16 anos em que o comportamento dos atributos pode ter variado muito e tornado a classificação uma tarefa difícil de se realizar.

O foco do trabalho foi notícias Web, contudo pode se ainda reduzir a especialidade e trabalhar com a categoria da notícia, fonte ou tipo de página. Para isso, espera-se que o trabalho desenvolvido nesta pesquisa sirva como apoio para avaliar a credibilidade de qualquer informação ou possivelmente ter um impacto maior sendo empregado na área de jornalismo para checar uma grande quantidade de notícias em tempo real.

Referências Bibliográficas

Aggarwal, Sonal e Oostendorp, H. e. R. R. Y. e. I. B. (2014). Providing web credibility assessment support.

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Bezerra, M. A. (2015). Uma investigação do uso de características na detecção de urls. Instituto de Computação.

Bird, S. and Klein, Ewan e Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cavaco, Afonso Miguel e Várzea, D. (2010). Contribuição para o estudo da leitura de folhetos informativos nas farmácias Portuguesas. *Revista Portuguesa de SaÃPÃ*, 28:179 – 186.

Fogg, B. J. e Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 80–87, New York, NY, USA. ACM.

Ginsca, Alexandru L. e Popescu, A. e. L. M. (2015). Credibility in information retrieval. *Found. Trends Inf. Retr.*, 9(5):355–475.

Horne, B. and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Horne, Benjamin D. e Dron, W. e. K. S. e. A. S. (2018). Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 235–238, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Jurafsky, Daniel e Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324. Relevance.

Liu, Xin e Nielek, R. e. W. A. e. A. K. (2013). Defending imitating attacks in web credibility evaluation systems. pages 1115–1122.

Olteanu, Alexandra e Peshterliev, S. e. L. X. e. A. K. (2013). Web credibility: Features exploration and credibility prediction. volume 7814, pages 557–568.

Papaioannou, Thanasis G. e Ranvier, J.-E. e. O. A. and Aberer, K. (2012). A decentralized recommender system for effective web credibility assessment. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 704–713, New York, NY, USA. ACM.

Pattanaphanchai, J., O'Hara, K., and Hall, W. (2013). Trustworthiness criteria for supporting users to assess the credibility of web information. In *the 22nd international conference on World Wide Web companion (WWW '13 Companion)*, pages 1123–1130.

Popat, K. (2017a). Assessing the credibility of claims on the web. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 735–739, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Popat, Kashyap e Mukherjee, S. e. S. J. e. W. G. (2017b). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1003–1012, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Popat, Kashyap e Mukherjee, S. e. S. J. e. W. G. (2018). Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 155–158, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Putri Ghaisani, A., Munajat, Q., and Handayani, P. W. (2017). Information credibility factors on information sharing activites in social media. In *2017 Second International Conference on Informatics and Computing (ICIC)*, pages 1–5.

Rejmund, E., Jaworski, W., and Wierzbicki, A. (2014). Exploratory study of relationships among statement credibility, context, and semantic similarity. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 459–462.

Schwarz, Julia e Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1245–1254, New York, NY, USA. ACM.

Scikit-learn (2020). Selectkbest.

Treeratpituk, Pucktada e Giles, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 39–48, New York, NY, USA. ACM.

Villela, S. M., Xavier, A. E., Fonseca Neto, R., and de Castro Leite, S. (2011). Seleção de caracteristicas utilizando busca ordenada e um classificador de larga margem. In Barreto, G. d. A. and Costa, J. A. F., editors, *Anais do 10 Congresso Brasileiro de Inteligência Computacional*, pages 1–8, Fortaleza, CE. SBIC.

Wahsheh, H. A., Alsmadi, I. M., and Al-Kabi, M. N. (2013). The evaluation of trust and credibility metrics: Websites of jordanian universities and e-government portals as a case study. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–6.

Wawer, Aleksander e Nielek, R. e. W. A. (2014). Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 1135–1140, New York, NY, USA. ACM.

Wierzbicki, A. (2018). *Web Content Credibility*. Springer International Publishing.

Witten, Ian H. e Frank, E. e. H. M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Xu, J., Yang, X., and Wang, L. (2011). Evaluation method of information credibility based on the trust features of web page. In *2011 Eighth Web Information Systems and Applications Conference*, pages 69–72.

Yamamoto, Y. (2017). Supporting credibility judgment in web search by yusuke yamamoto, with martin vesely as coordinator. *SIGWEB Newsl.*, (Spring):3:1–3:11.