CENTRO UNIVERSITÁRIO FEI
BRUNO CÉSAR PEREIRA DOS SANTOS CARVALHO

O impacto da inclusão de Índices Sociais e Macroeconômicos na Detecção Automática de *Fake News*

São Bernardo do Campo
2022

BRUNO CÉSAR PEREIRA DOS SANTOS CARVALHO

O impacto da inclusão de Índices Sociais e Macroeconômicos na Detecção Automática de *Fake News*

Dissertação de Mestrado apresentada ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de Mestre em Engenharia Elétrica. Orientado pelo Prof. Dr. Flavio Tonidandel.

São Bernardo do Campo

2022

Pereira dos Santos Carvalho, Bruno César.
O impacto da inclusão de Índices Sociais e Macroeconômicos na Detecção Automática de Fake News / Bruno César Pereira dos Santos Carvalho. São Bernardo do Campo, 2022.
90 f. : il.

Dissertação - Centro Universitário FEI.
Orientador: Prof. Dr. Flávio Tonidandel.

1. Fake news. 2. Bag of Words. 3. SVM. 4. Random Forest. 5. Naive Bayes. I. Tonidandel, Flávio , orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

APRESENTAÇÃO DE DISSERTAÇÃO
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Mestrado

PGE-10

Aluno(a): Bruno César Pereira dos Santos Carvalho **Matrícula:** 119304-4

Título do Trabalho: Avaliação do impacto da inclusão de índices sociais e macroeconômicos no desempenho da tarefa de detecção automática de fake news utilizando-se classificadores de aprendizado de máquina em conjunto do modelo bag of words

Área de Concentração: Inteligência Artificial Aplicada à Automação e Robótica

Orientador(a): Prof. Dr. Flavio Tonidandel

Data da realização da defesa: 29/09/2022 **ORIGINAL ASSINADA**

Avaliação da Banca Examinadora:

O aluno fez sua apresentação apropriadamente e respondeu satisfatoriamente todas as perguntas dos membros da banca. A banca sugere a alteração do título do trabalho para: "O impacto da inclusão de Índices Sociais e Macroeconômicos na Detecção Automática de Fake News". Sem outras considerações, o trabalho apresentado foi considerado aprovado por unanimidade.

A Banca Julgadora acima-assinada atribuiu ao aluno o seguinte resultado:

APROVADO ☒ REPROVADO ☐

MEMBROS DA BANCA EXAMINADORA

Prof. Dr. Flavio Tonidandel

Prof. Dr. Carlos Eduardo Thomaz

Prof. Dr. Julio Cesar dos Reis

Aprovação do Coordenador do Programa de Pós-graduação

Prof. Dr. Carlos Eduardo Thomaz

Dedico este trabalho a minha esposa, aos meus familiares e professores, que com sua presença o tornaram possível.

AGRADECIMENTOS

A minha esposa, Rafaela, por todo suporte e apoio durante esta etapa da minha vida.

Aos meus pais, Lourdes e Vital, que sempre buscaram me oferecer a melhor educação que puderam.

Ao meu orientador, Prof. Dr. Flávio Tonidandel, que sempre esteve presente da melhor maneira possível para as devidas orientações e conselhos no desenvolvimento deste trabalho.

“Uma vez que se privou o homem da verdade, é pura ilusão pretender torná-lo livre. Verdade e liberdade, com efeito, ou caminham juntas, ou juntas miseravelmente perecem.”

João Paulo II

RESUMO

As *fake news* são notícias falsas que possuem a intenção de se apresentar como verdadeiras, causando assim diversos impactos na sociedade, sejam econômicos e financeiros, políticos, ou até em saúde pública. A detecção automática de *fake news* teve um avanço importante nos últimos anos, porém existe a carência de trabalhos em língua portuguesa, principalmente devido à baixa oferta de bases nesta língua. Pode-se evidenciar, também, que os trabalhos existentes são focados na detecção automática utilizando-se as características linguísticas ou nos padrões de dispersão das mesmas em mídias sociais. Com isso, este trabalho traz uma abordagem complementar à tarefa de detecção automática desses tipos de textos, incluindo o estudo do possível impacto da inclusão de índices sociais e macroeconômicos, como taxa de inflação e desemprego, no desempenho dos modelos. O trabalho foca na utilização dos classificadores SVM, *Random Forest* e *Naive Bayes*, além do conhecido modelo *Bag of Words* para extração das características linguísticas.

Palavras-chave: *Fake news. Bag of Words.* SVM. *Random Forest. Naive Bayes.*

ABSTRACT

Fake news are known for being presented as real news, but in fact there is a clear intention to be harmful. Fake news can cause a diverse variety of impacts in the Society, as financial and economic, political, and even in public health. The automatic fake news detection task has improved in recent years, however there is a lack of Portuguese fake news databases, which impacts the number of such studies in Portuguese. Additionally, most of the studies are focused on linguistics features or news network dispersion patterns in social medias. Taking it into account, this study intends to bring a new approach to the automatic fake news detection task, including the assessment of the possible impact of social and macroeconomics indexes, such as inflation and unemployment rate, in the task of detecting fake news automatically. This work will consider the SVM, Random Forest and Naive Bayes classifiers. Beyond these models, it will be considered the known Bag of Words model for linguistic feature extraction.

Keywords: Fake news. Bag of Words. SVM. Random Forest. Naive Bayes.

ÍNDICE DE ILUSTRAÇÕES

ÍNDICE DE QUADROS

Quadro 1- Exemplo de notícias da base Fake.Br 37
Quadro 2 - Exemplo de notícias da base FakeRecogna 38

ÍNDICE DE TABELAS

Tabela 1 – Indicadores IPCA para o período de 2016 até 2021 45
Tabela 2 - Indicadores PNAD contínua para o período entre 2016 e 2021 46
Tabela 3 - Indicadores IBOVESPA contínua para o período entre 2016 e 2021 47
Tabela 4 - Cenários de teste considerados para base FakeRecogna completa 51
Tabela 5 - Cenários de teste considerados para base FakeRecogna considerando-se apenas as **cinquenta primeiras palavras de cada notícia** 52
Tabela 6 - Cenários de teste considerados para base Fake.Br completa 52
Tabela 7 - Cenários de teste considerados para base Fake.Br considerando-se apenas as **cinquenta primeiras palavras de cada notícia** 53
Tabela 8 - Análise FST 54
Tabela 9 - Valores Críticos do FST para 20 cenários com três classificadores e α=5% 55

ÍNDICE DE FIGURAS

Figura 1 – Popularidade do termo *fake news* nas buscas do Google 16
Figura 2 - Processo macroscópico de classificação de notícias 21
Figura 3 - Ilustração do treinamento de um modelo 22
Figura 4 - Algoritmo Random Forest de Breiman 24
Figura 5 - Exemplo de Support Vector Machine 25
Figura 6 - Exemplo de utilização de Kernel para linearização 26
Figura 7 - Workflow macroscópico da tarefa de detecção de *fake news* 29
Figura 8 - Workflow macroscópico da tarefa de detecção de fake news considerando variáveis adicionais de contexto social e econômico 31
Figura 9 - *Bag of Words* 32
Figura 10 - Vetor do *Bag of Words* por Frequência 32
Figura 11 - Vetor do *Bag of Words* por Presença (binário) 33
Figura 12 - Fase I de Pré-Processamento 33

Figura 13 - Exemplo em inglês de stemming 34
Figura 14 – Exemplos de stemming utilizando o algoritmo de Orengo e Huyck 34
Figura 15 - Algoritmo de Stemming para português 35
Figura 16 - Pré-Processamento completo (Fases I e II) 35
Figura 17 - Processo com a base Fake.Br Corpus 42
Figura 18 - Processo de deslocamento de datas com a base Fake.Br Corpus 44
Figura 19 - Exemplos de AUC 50
Figura 20 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias 56
Figura 21 - Curva AUC Cenário CSVMRecogna 57
Figura 22 - Curva AUC Cenário CSVMRecogna0d 57
Figura 23 - Curva AUC Cenário CSVMRecogna30d 57
Figura 24 - Curva AUC Cenário CSVMRecogna60d 57
Figura 25 - Curva AUC Cenário CSVMRecogna90d 57
Figura 26 - Precisão, Recall, F1-Score e Acurácia os Cenários CNBRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias 58
Figura 27 - Curva AUC Cenário CNBRecogna 59
Figura 28 - Curva AUC Cenário CNBRecogna0d 59
Figura 29 - Curva AUC Cenário CNBRecogna30d 59
Figura 30 - Curva AUC Cenário CNBRecogna60d 59
Figura 31 - Curva AUC Cenário CNBRecogna90d 59
Figura 32 - Precisão, Recall, F1-Score e Acurácia para os Cenários CRFRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias 60
Figura 33 - Curva AUC Cenário CRFRecogna 61
Figura 34 - Curva AUC Cenário CRFRecogna0d 61
Figura 35 - Curva AUC Cenário CRFRecogna30d 61
Figura 36 - Curva AUC Cenário CRFRecogna60d 61
Figura 37 - Curva AUC Cenário CRFRecogna90d 61
Figura 38 - Precisão, Recall, F1-Score e Acurácia para os Cenários CSVMRecogna50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias 62
Figura 39 - Curva AUC Cenário CSVMRecogna50W 63
Figura 40 - Curva AUC Cenário CSVMRecogna50W0d 63
Figura 41 - Curva AUC Cenário CSVMRecogna50W30d 63
Figura 42 - Curva AUC Cenário CSVMRecogna50W60d 63

Figura 43 - Curva AUC Cenário CSVMRecogna50W90d 63
Figura 44 - Precisão, Recall, F1-Score e Acurácia para os Cenários CNBRecogna50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias 64
Figura 45 - Curva AUC Cenário CNBRecogna50W 65
Figura 46 - Curva AUC Cenário CNBRecogna50W0d 65
Figura 47 - Curva AUC Cenário CNBRecogna50W30d 65
Figura 48 - Curva AUC Cenário CNBRecogna50W60d 65
Figura 49 - Curva AUC Cenário CNBRecogna50W90d 65
Figura 50 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias 66
Figura 51 - Curva AUC Cenário CRFRecogna50W 67
Figura 52 - Curva AUC Cenário CRFRecogna50W0d 67
Figura 53 - Curva AUC Cenário CRFRecogna50W30d 67
Figura 54 - Curva AUC Cenário CRFRecogna50W60d 67
Figura 55 - Curva AUC Cenário CRFRecogna50W90d 67
Figura 56 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias 68
Figura 57 - Curva AUC Cenário CSVMFake.BrCorpus 69
Figura 58 - Curva AUC Cenário CSVMFake.BrCorpus0d 69
Figura 59 - Curva AUC Cenário CSVMFake.BrCorpus30d 69
Figura 60 - Curva AUC Cenário CSVMFake.BrCorpus60d 69
Figura 61 - Curva AUC Cenário CSVMFake.BrCorpus90d 69
Figura 62 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CNBFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias 70
Figura 63 - Curva AUC Cenário CNBFake.BrCorpus 71
Figura 64 - Curva AUC Cenário CNBFake.BrCorpus0d 71
Figura 65 - Curva AUC Cenário CNBFake.BrCorpus30d 71
Figura 66 - Curva AUC Cenário CNBFake.BrCorpus60d 71
Figura 67 - Curva AUC Cenário CNBFake.BrCorpus90d 71
Figura 68 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias 72
Figura 69 - Curva AUC Cenário CRFFake.BrCorpus 73
Figura 70 - Curva AUC Cenário CRFFake.BrCorpus0d 73
Figura 71 - Curva AUC Cenário CRFFake.BrCorpus30d 73

Figura 72 - Curva AUC Cenário CRFFake.BrCorpus60d 73
Figura 73 - Curva AUC Cenário CRFFake.BrCorpus90d 73
Figura 74 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMFake.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias 74
Figura 75 - Curva AUC Cenário CSVMFake.BrCorpus50W 75
Figura 76 - Curva AUC Cenário CSVMFake.BrCorpus50W0d 75
Figura 77 - Curva AUC Cenário CSVMFake.BrCorpus50W30d 75
Figura 78 - Curva AUC Cenário CSVMFake.BrCorpus50W60d 75
Figura 79 - Curva AUC Cenário CSVMFake.BrCorpus50W90d 75
Figura 80 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CNBFake.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias 76
Figura 81 - Curva AUC Cenário CNBFake.BrCorpus50W 77
Figura 82 - Curva AUC Cenário CNBFake.BrCorpus50W0d 77
Figura 83 - Curva AUC Cenário CNBFake.BrCorpus50W30d 77
Figura 84 - Curva AUC Cenário CNBFake.BrCorpus50W60d 77
Figura 85 - Curva AUC Cenário CNBFake.BrCorpus50W90d 77
Figura 86 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFke.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias 78
Figura 87 - Curva AUC Cenário CSVMFake.BrCorpus50W 79
Figura 88 - Curva AUC Cenário CSVMFake.BrCorpus50W0d 79
Figura 89 - Curva AUC Cenário CSVMFake.BrCorpus50W30d 79
Figura 90 - Curva AUC Cenário CSVMFake.BrCorpus50W60d 79
Figura 91 - Curva AUC Cenário CSVMFake.BrCorpus50W90d 79
Figura 92 – Aumento incremental com o SVM e base Recogna 80
Figura 93 - Aumento incremental com o RF e base Recogna 81
Figura 94 - Aumento incremental com o SVM e base Recogna modificada 82
Figura 95 - Aumento incremental com o RF e base Recogna modificada 83
Figura 96 - Aumento incremental com o SVM e base Fake.Br 84
Figura 97 - Aumento incremental com o RF e base Fake.Br 84
Figura 98 - Aumento incremental com o SVM e base Fake.Br modificada 85
Figura 99 - Aumento incremental com o RF e base Fake.Br modificada 85
Figura 100 – Indicação da melhoria incremental ao se comparar os modelos sem e com os indicadores sociais e macroeconômicos 86

SUMÁRIO

1 INTRODUÇÃO **14**
1.1 OBJETIVO 15
1.2 JUSTIFICATIVA E MOTIVAÇÃO 16
2 TRABALHOS RELACIONADOS **18**
3 REVISÃO BIBLIOGRÁFICA E CONCEITOS FUNDAMENTAIS **21**
3.1 APRENDIZADO DE MÁQUINA 21
3.1.1 Modelo Naive Bayes (NB) **22**
3.1.2 Modelo Random Forest (RF) **23**
3.1.3 Modelo Support Vector Machine (SVM) **24**
3.1.1.1 Modelo Support Vector Machine (SVM) Lineares e Não-Lineares ***25***
3.1.1.2 Formulação Matemática Aplicada em Algoritmos ***26***
3.2 TÉCNICAS DE DETECÇÃO DE FAKE NEWS 29
3.2.1 Content - Based **29**
3.2.2 Context - Based **30**
3.2.3 Modelo Híbrido **30**
3.3 NATURAL LANGUAGE PROCESSING 31
3.3.1 Bag of Words **32**
3.3.1.1 Pré-processamento e Stemmming ***33***
3.3.1.2 Stemmming ***34***
3.4 BASES DE DADOS CONSIDERADAS 36
3.4.1 Fake.Br Corpus **36**
3.4.1 FakeRecogna Corpus **37**
3.4.3 Base de Dados Sociais e Macroeconômicos **38**
3.4.3.1 Normalização de Dados ***40***
4 METODOLOGIA **42**
4.1 RECURSOS COMPUTACIONAIS E FERRAMENTAS 43
4.2 BASES DE DADOS DE NOTÍCIAS 43
4.3 BASES DE DADOS DE ÍNDICES 44

4.4 MÉTRICAS DE VALIDAÇÃO 48
4.5 TESTES EXECUTADOS 50
5 RESULTADOS 54
5.1 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKERECOGOGNA 55
5.2 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKERECOGOGNA 58
5.3 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKERECOGOGNA 60
5.4 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKERECOGOGNA50W 62
5.5 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKERECOGOGNA50W 64
5.6 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKERECOGOGNA50W 66
5.7 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKEBR 68
5.8 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKEBR 70
5.9 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKEBR 72
5.10 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKEBR50W 74
5.11 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKEBR50W 76
5.12 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKEBR50W 78
5.13 ANÁLISE GERAL E COMPARATIVA 80
6 DISCUSSÕES SOBRE OS RESULTADOS 88
7 CONCLUSÃO E ESTUDOS FUTUROS 89
REFERÊNCIAS 91

1 INTRODUÇÃO

A maneira como nos comunicamos e consumimos informações e notícias atualmente vem mudando com o aumento da utilização da internet e a disponibilidade cada vez maior de canais de divulgação que fogem dos tradicionais. Pode-se dizer que uma edição de jornal atualmente tem mais disponibilidade de informação do que um cidadão do século 19 era exposto em sua vida inteira. Esse avanço na maneira como a informação é disseminada e criada tem afetado o modo que lidamos com as informações, fazendo com que informações de mídias não oficiais e de redes sociais sejam cada vez mais lidas e tomadas como verdade (PENTINA; TARAFDAR, 2014). Além disso, sabe-se que mais de 60% das informações consumidas pela população brasileira têm origem em plataformas digitais (MITCHELL, 2016).

As notícias falsas estão presentes há muito tempo. Há um episódio famoso que ocorreu em 1938, quando durante uma transmissão de rádio do drama do escritor H. G. Well em formato de radionovela levou diversos telespectadores a acreditarem de fato que havia uma invasão marciana ocorrendo (TANDOC; LIM; LING, 2017).

A digitalização das plataformas de notícias fornece agora um espaço para não jornalistas seguirem com o compartilhamento de informações e com grande potencial de alcance (TANDOC; LIM; LING, 2017). Esses usuários podem postar fotos, áudios, vídeos e outras informações em primeira mão, e a popularidade dessas mídias e seus usuários, portanto, facilita a propagação de notícias não verificadas nesse meio (TANDOC; LIM; LING, 2017).

Dado que a maneira indicada de se evitar a dispersão de *fake news* inclui a avaliação das fontes, a checagem dos autores, conferência das datas de publicação, entre outros pontos, conclui-se que a verificação da veracidade ou não de uma notícia pode ser uma tarefa não-trivial, principalmente com a enorme quantidade de informação que há transitando atualmente (QUINTANILHA; DA SILVA; LAPA, 2019).

O termo *fake news* se tornou bastante comentado nos últimos tempos, principalmente após as eleições presidenciais de 2016 nos Estados Unidos, que segundo Tandoc, Lim e Ling (2018, apud ALBRIGHT, 2016) sofreram negativamente com a intensa dispersão de *fake news*, e embora as *fake news* tenham diferentes definições, há de comum entendimento, segundo Tandoc, Lim e Ling (2018), o fato de as *fake news* terem a intenção de parecer notícias verdadeiras, de se apropriarem do estilo e aparência de uma informação verídica.

Além do impacto negativo durante eleições, pode-se acrescentar, conforme apontado por Tandoc, Lim e Ling (2018), Guderlei e Assenmacher (2020) e Trung Tin (2018), que as

fake news têm grande impacto social, podendo gerar influência também em políticas de saúde pública e em questões econômicas.

Dado todo o contexto macroscópico da facilidade de dispersão de *fake news* em nossa sociedade, e os impactos sociais já evidenciados, este estudo tem como hipótese a ideia de que pode haver uma correlação entre o aumento da dispersão de *fake news* e a mudança negativa em índices sociais e macroeconômicos. Por exemplo, tem-se como hipótese neste estudo que o aumento do IPCA, Índice de Preços ao Consumidor, que é um dos principais índices de preços calculados pelo IBGE (Instituto Brasileiro de Geografia e Estatística) e mede a chamada inflação (MARTINEZ; CERQUEIRA, 2013), geraria um aumento na dispersão de *fake news* na sociedade.

A hipótese de que o aumento do IPCA pode gerar um aumento na dispersão de *fake news* é extrapolada para outros importantes índices; o IBOVESPA e a taxa de desemprego do país, dado que segundo o Datafolha (pesquisa de 2021), a taxa de desemprego, inflação e a situação econômica figuram entre as principais preocupações dos brasileiros. Por isso, este estudo procura entender se há melhoria no desempenho de um modelo de detecção automática de *fake news* ao se considerar esses índices como variáveis de entrada.

1.1 OBJETIVO

A proposta de trabalho tem como objetivo a avaliação do desempenho dos classificadores *Support Vector Machine* (SVM), *Random Forest* (RF) e *Naive Bayes* (NB) na tarefa de detecção automática de *fake news* considerando as variáveis de entrada do modelo *Bag Of Words*, porém levando-se em conta a adição de novas variáveis quantitativas para verificar se o modelo exerce uma melhor performance. São essas variáveis os índices da Taxa de Desemprego da Pesquisa Nacional por Amostra de Domicílios (PNAD), o IBOVESPA e o IPCA.

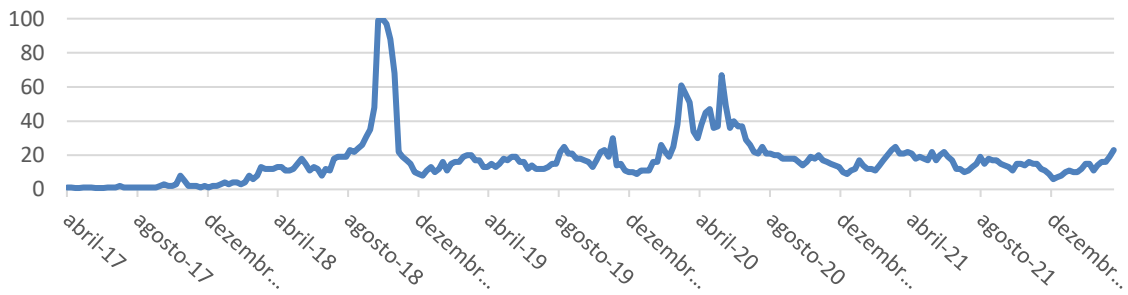
As notícias do estudo seguem em formato de texto, e os classificadores considerados se utilizam de Aprendizado de Máquina. O algoritmo de classificação considerado deve ter como resultado uma resposta binária objetiva: notícia falsa e notícia não-falsa, mesmo embora possam existir em bases de outros estudos outras classificações (notícias de tom satírico, não definidas etc.). Essa proposta tem como foco trabalhar especificamente com a língua portuguesa, mais especificamente o português brasileiro.

Resumidamente, pode se dizer que a hipótese principal que direciona o objetivo do estudo é que determinadas condições podem ter correlação com o surgimento de *fake news* e poderiam ser variáveis de contexto que auxiliariam na classificação correta de uma *fake news*.

1.2 JUSTIFICATIVA E MOTIVAÇÃO

As notícias falsas têm tido um foco grande na discussão pública nos anos recentes. Podemos inferir isso através da ferramenta Google Trends, que indica um aumento significativo na busca pelo termo *fake news* a partir de outubro de 2016, conforme a Figura 1 (o valor de 100 indica o pico de popularidade do termo na escala utilizada pelo Google Trends). Nota-se um pouco no último trimestre de 2018, que coincide com o período de eleições no país.

Figura 1 – Popularidade do termo *fake news* nas buscas do Google



Fonte: dados Google *Trends*

Adicionalmente, conforme já comentado anteriormente, pode-se evidenciar que, as *fake news* geram grande impacto social, podendo influenciar em eleições, políticas de saúde pública e gerar impactos em questões econômicas, daí a intenção de se avaliar se índices sociais e macroeconômicos podem servir como variáveis para auxiliar a detecção de *fake news*.

Acrescido aos claros impactos da dispersão de *fake news*, além da sua popularidade, leva-se em conta também o fato de termos poucos trabalhos avaliando o desempenho de modelos de detecção de *fake news* aplicados à realidade brasileira como fator importante para justificar o estudo do tema, conforme já evidenciado por Monteiro et al (2018). Além disso, vale ressaltar-se a dificuldade da tarefa de detecção automática de *fake news*¸ dado que se necessita da existência de bases de notícias classificadas e confiáveis, com uma quantidade suficientemente grande de registros. As notícias devem ter passado por uma curadoria especializada que garanta a confiabilidade da sua classificação em verdadeira ou falsa, daí a

necessidade de considerar-se bases disponibilidades por grupos jornalísticos especializados nesse tipo de atividade.

O estudo presente segue dividido em 7 capítulos. Após o capítulo introdutório, no segundo capítulo, é apresentada uma revisão dos principais estudos relacionados às técnicas de detecção automática de *fake news*. Em seguida, no terceiro capítulo, há uma revisão teórica para suportar o entendimento da metodologia aplicada. No quarto capítulo tem-se a metodologia considerada, com suas definições para a construção dos testes e as métricas envolvidas. Os resultados são apresentados no quinto capítulo, e no sexto capítulo encontram-se discussões sobre os resultados e ameaças à validade do estudo. Por fim, no sexto capítulo, tem-se as conclusões e discussões sobre possíveis estudos futuros.

2 TRABALHOS RELACIONADOS

A tarefa de Detecção Automática de *fake news* pode ter como resultado a classificação binária de um trecho de texto, indicando que o mesmo se trata ou não de uma notícia falsa. É possível também existir uma classificação multidimensional, na qual teremos respostas adicionais, as quais podem incluir o assunto da notícia ou diferenciar, por exemplo, notícias falsas, notícias verdadeiras não-satíricas e notícias de tom satírico, conforme indicado por Guderlei e Assenmacher (2020) e Trung Tin (2018).

Os métodos de detecção automática de *fake news* se dividem basicamente em dois tipos: o *social context-based* e o *content-based*, conforme visto no trabalho de Shu et al (2017). O *social context-based* se utiliza de dados do contexto da informação, principalmente o padrão de difusão da informação em redes sociais e outras plataformas, e não do conteúdo em si, conforme encontrado nos trabalhos de Granskogen (2018). O método *content-based* se baseia no conteúdo da informação, podendo se subdividir no *knowledge-based*, que compara as informações analisadas com dados em bases de conhecimento, e o *style-based*, que extrai características linguísticas do texto em questão, conforme comentado por Kaliyar et al (2020).

Evidenciando-se o método *content-based*, nota-se a aplicação de modelos de aprendizado supervisionado (modelos BayesNet, JRip, OneR, ZeroR entre outros), como indicado na publicação de Ozbay e Alatas (2019), ou trabalhos que consideram classificadores clássicos, como KNN, *Naive Bayes*, SVM e *Random Forest*, notados na publicação de Reis et al (2019). Há também autores que propõem a utilização de modelos de *deep learning*, inclusive com a proposta de um novo modelo, o FDNet, como apontado por Kaliyiar e et al (2020).

Acerca da etapa de processamento dos textos, pode-se evidenciar as técnicas *Word2Vec* e *Bag of Words* como sendo bastante presentes, sendo utilizadas na etapa de processamento dos textos e obtenção dos parâmetros de entrada nos modelos utilizados por Faustini e Covões (2020). Para o processamento dos textos é usual também que ocorram etapas de remoção de *stopwords*, espaços em branco e pontuação, além da lematização de palavras, conforme demonstrado por Agarwal et al (2019).

Os textos não podem ser inseridos diretamente nos modelos de classificação, por isso devem passar pela etapa de processamento natural de linguagem, utilizando-se de métodos como o *Word2Vec* e *Bag of Words*, como evidenciado por Bokka et al (2019) em sua publicação.

Acerca das bases de dados consideradas para as pesquisas, há uma grande diversidade disponível em língua inglesa. No trabalho apresentado por Agarwal et al (2019, apud WANG,

2017), há a base de dados LIAR, uma base que contém treze colunas de dados (consideradas como variáveis) e já possui a separação entre três bases, sendo as bases de treino, validação e teste. Em seu estudo, Agarwal et al (2019) considera o (i) processamento dos textos da base, (ii) a extração de características, (iii) a passagem das características em modelos classificatórios (Naive Bayes, Regressão Logística, SVM, Gradiente Estocástico e Random Forest), e (iv) a resposta final (verdadeira ou falso).

No estudo apresentando por Orbay e Alatas (2020, apud HORNE; ADALI, 2017) as bases consideradas são utilizadas as bases (i) BuzzFeed Political News, (ii) Random Political News e (iii) ISO Fake News. Existem em (i) textos coletados de um artigo do BuzzFeed sobre notícias falsas e verdadeiras obtidas da rede social Facebook, totalizando 1627 artigos, já em (ii) há um número pouco relevante, de 75 notícias. Para a base (iii) são apresentadas 44898 notícias de diversos sites. As bases em questão são processadas em uma etapa inicial (tokenização, remoção de *stopwords* e *stemming*), depois passam por uma etapa de vetorização de textos para a criação dos parâmetros de entrada dos modelos, e aí existe a etapa de classificação, na qual tem-se diversos modelos (modelos BayesNet, JRip, OneR, ZeroR entre outros).

Em língua portuguesa evidencia-se o estudo de Monteiro et al (2018), no qual é citado que a base de dados proposta, até então, é a única base em língua portuguesa de referência para o estudo de detecção automática de *fake news*. A base de dados em questão, a Fake.Br Corpus, é composta de notícias falsas e verdadeiras exclusivamente em português brasileiro. Nessa base são encontradas 7200 notícias, sendo divididas igualmente entre notícias falsas e verdadeiras. As notícias falsas foram obtidas dos portais Diário do Brasil, A Folha do Brasil, The Jornal Brasil e Top Five. Para as verdadeiras, foram consultados os portais G1, Folha de São Paulo e Estadão.

Monteiro et al (2018) utiliza as características propostas por Pérez-Rosas e Mihalcea (2015) aplicadas em língua portuguesa, demonstrando, através dos resultados que as características utilizadas como entrada dos modelos podem ser válidas para o trabalho de detecção de *fake news* em língua portuguesa. As características propostas são, em inglês: *Part of speech (POS) tags, Semantic classes, Bag of words, Pausality, Emotiveness, Uncertainty* e *Non-immediacy*. Como modelos de classificação, foram considerados o SVM, *Naive Bayes*, *Random Forest* e *Multilayer Perceptron*.

Além da base Fake.Br proposta por Monteiro et al (2018), Garcia, Afonso e Papa (2022) apresenta uma base em português brasileiro que conta com 8821 notícias (FakeRecogna), sendo 4628 notícias falsas e 4193 notícias verdadeiras. Essa base conta com notícias falsas coletadas

de portais reconhecidos pela apuração de *fake news*; Boatos.Org, Fato ou Fake (grupo Globo), E-Farsas, UOL Confere, AFP Checamos e o Projeto Comprova. Já para as notícias verdadeiras, foram considerados os portais UOL, G1, Extra, além de páginas governamentais oficiais.

Ao serem analisados os trabalhos relacionados, principalmente em língua portuguesa, nota-se uma predileção pelo método *content-based*, no qual tem-se a análise do conteúdo da notícia. Para trabalhos em outras línguas, nota-se o uso de outros métodos, incluindo o método *social context-based* e até o híbrido, porém sempre considerando variáveis que são oriundas ou do conteúdo do texto, ou do contexto de dispersão de notícias nas mídias sociais (e.g.: perfil do usuário e engajamento da postagem).

Este trabalho tem como diferença o fato de ter-se como intenção ser levado em consideração outros tipos de variáveis de contexto, não exatamente do padrão de dispersão da notícia como o método *social context-based*, mas sim variáveis oriundas do contexto macroeconômico e social do país, tendo como hipótese a correlação da existência de um maior número de *fake news* sendo gerada em momentos mais críticos na sociedade. Desta maneira, o trabalho procura validar se realmente há sentido para se considerar variáveis de um contexto diferente do *social context-based* em um modelo de detecção automática de *fake news*.

Além disso, vale ser lembrado que existem diversos modelos classificatórios sendo considerados além do SVM, *Random Forest* (RF) e *Naive Bayes* (NB), porém nota-se que esses são os modelos mais amplamente aplicados.

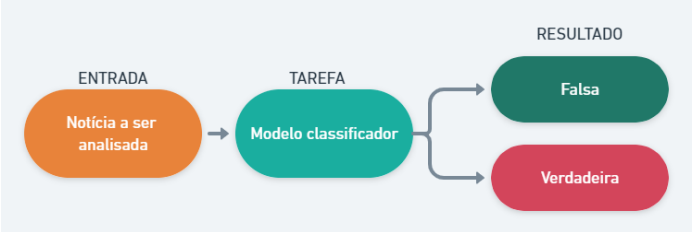
Adicionalmente, pode ser evidenciado, não somente no estudo de Monteito et al (2018), que as variáveis que imprimem um dos melhores resultados nesses estudos são oriundas do modelo *Bag Of Words*, por isso supõem-se que inclusão de novas variáveis linguísticas dificilmente geraria melhoria nos resultados em um modelo *content-based* de detecção de *fake news*, e assim não haveria necessidade de considerar-se um modelo mais complexo. Neste caso, o foco acaba direcionado na tentativa de se aprimorar os modelos SVM, *Random Forest* (RF) e *Naive Bayes* (NB), com variáveis *content-based* do modelo *Bag Of Words*, e posteriormente acrescentando-se as demais variáveis de contexto social e macroeconômico.

3 REVISÃO BIBLIOGRÁFICA E CONCEITOS FUNDAMENTAIS

Para a tarefa de detecção automática de *fake news*, além da definição do conceito de *fake news* conforme a seção introdutória, há também de se destacar a necessidade de entendimento de importantes conceitos utilizados como meio para a obtenção do resultado, que é a resposta binária entre uma notícia, classificando-a em falsa ou verdadeira.

De maneira macroscópica, o objetivo final dar-se-á conforme a Figura 2.

Figura 2 - Processo macroscópico de classificação de notícias



Fonte: autor

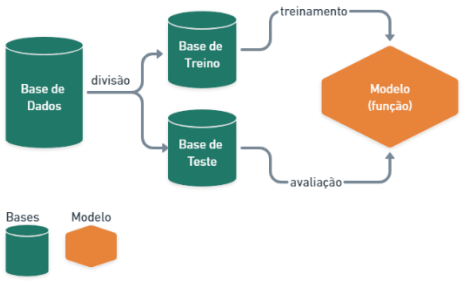
Para a obtenção do resultado, deve-se existir a etapa de treinamento do modelo classificatório a ser considerado, além da etapa de obtenção de dados para o treinamento, validação e teste.

3.1 APRENDIZADO DE MÁQUINA

Conforme apontado por Mahesh (2016), o propósito do aprendizado de máquina é aprender através de conjuntos de dados, e com a abundância de dados tem-se uma maior facilidade de executar essa tarefa. Há diversos algoritmos e a sua aplicação depende do tipo de problema a ser resolvido. Há algoritmos de aprendizado supervisionado, aprendizado não-supervisionado, entre outros tipos.

A tarefa de detecção de *fake news* é tida como uma tarefa de classificação, entre notícia falsa ou verdadeira, sendo um tipo de aprendizado supervisionado, o qual é executado, conforme o estudo de Mahesh (2016), com o aprendizado de uma função através dos exemplos vindos das bases de dados. As bases de dados são divididas em bases de treino e teste, e com as bases de treino as funções são inseridas, resultando em um modelo preditivo ou classificatório, conforme indicado na Figura 3.

Figura 3 - Ilustração do treinamento de um modelo



Fonte: Baseado em Mahesh (2016).

Para a tarefa de classificação das *fake news*, nota-se dos estudos relacionados elaborados por Abonizio et al (2020) e Faustini e Covões (2020), que os modelos de classificação clássicos como o SVM, *Naive Bayes* e *Random Forest* trazem bons resultados em língua portuguesa, mesmo embora possam ter um desempenho diferente em outras línguas. Dado que o interesse atual do trabalho é ser focado em língua portuguesa, entende-se que o estudo com esses modelos pode ser satisfatório. Além disso, eles já se mostraram razoavelmente eficientes em língua portuguesa, como apontado por Monteiro et al (2018), o qual demonstra o atingimento de uma acurácia de 89% em alguns casos, daí a intenção de utilizá-los como modelos desse estudo para fins de comparação

3.1.1 Modelo Naive Bayes (NB)

Conforme apontado por Zhang (2004), o *Naive Bayes* é um algoritmo de classificação por aprendizado que se utiliza da Regra de Bayes. Tendo-se *C* representando a variável classe, *c* sendo o seu valor, que para esse exemplo pode assumir duas classes (+ para a classe positiva, e – para a classe negativa) e que a probabilidade de um exemplo $E = (x_1, x_2, x_3,..., x_n)$ ser de uma classe *c*, segundo a Regra de Bayes, é:

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)} \quad (1)$$

E é classificada com a classe *C*=+, se e somente se tivermos a função $f_b(E)$, um classificador Bayesiano, conforme abaixo:

$$f_b(E) = \frac{p(C = +|E)}{p(C = -|E)} \geq 1 \quad (2)$$

Assume-se que todos os atributos são independentes, dado o valor da variável classe, portanto:

$$p(E|c) = \prod_{i=1}^{n} \quad p(x_i|c) \quad (3)$$

O classificador resultante, chamado de Classificador Bayesiano, ou simplesmente Naive Bayes (NB), será:

$$f_{nb}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{n} \quad \frac{p(xi|C = +)}{p(xi|C = -)} \quad (4)$$

O Naive Bayes é a forma mais simples de rede Bayesiana, na qual considera-se que há independência condicional, um atributo que dificilmente ocorre em aplicações reais, porém tem um funcionamento satisfatório (MITCHELL, 2016).

3.1.2 MODELO RANDOM FOREST (RF)

O *Random Forest* (RF), conforme apontado por Schonlau e Zou (2020), é um algoritmo de aprendizado que é construído por blocos funcionais que se organizam em forma de árvores de decisão. Um modelo baseado em árvore trabalha dividindo uma base de dados em duas partes seguindo um determinado critério de divisão, e isso é feito de maneira recursiva até que seja atingido o final da árvore, definido por um critério de parada, que é conhecido como as folhas da árvore.

Dependendo de como os critérios de parada de divisão de dados são concebidos, a árvore pode ser utilizada para classificação ou regressão. Um critério comum para divisão de dados é a entropia. Em cada nó a entropia é dada por:

$$E = -\sum_{i=1}^{c} \quad p_i \times log\ (p_i) \quad (5)$$

Na equação 5, c é o número de classes únicas e p_i é a probabilidade a priori de cada classe e a intenção é maximizar o valor de E em cada nó.

As árvores de decisão são propensas ao *overfitting*, ponto que é mitigado utilizando-se a seleção aleatória de subespaços, algo que se tornou, conforme apontado por Breiman (2001), o *Random Forest*. O *Random Forest* pondera as predições utilizando-se de diversas árvores individuais concebidas por amostras numa estratégia de *bootstrapping*.

Abaixo, na Figura 4, tem-se a implementação do algoritmo do *Random Forest*.

Figura 4 - Algoritmo Random Forest de Breiman

```
for i ← 1 to B do
        Draw a bootstrap sample of size N from the training data;
        while node size != minimum node size do
                randomly select a subset of m predictor variables from total p;
                for j ← 1 to m do
                        if jth predictor optimizes splitting criterion then
                                split internal node into two child nodes;
                                break;
                        end
                end
        end
end
return the ensemble tree of all B subtrees generated in the outer for loop;
```

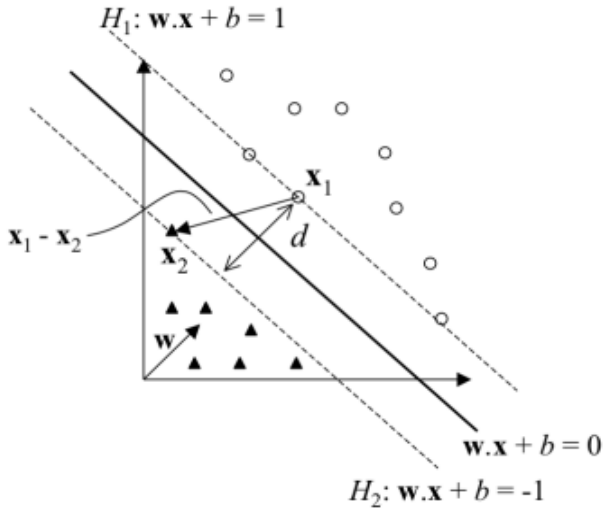
Fonte: Schonlau e Zou (2020)

3.1.3 MODELO SUPPORT VECTOR MACHINE (SVM)

O SVM, por extenso *Support Vector Machines*, é um algoritmo de classificação que se utiliza de aprendizado supervisionado. Mais especificamente, o algoritmo é fundamentado em um problema de otimização matemática e na teoria de aprendizado estatístico.

No SVM tem-se uma representação de exemplos como pontos, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço, o qual deve ser maximizado e deve garantir a divisão entre as classes. Novas entradas são mapeadas e devem ser inseridas em uma das classes mapeadas, conforme apontado por Lorena e Carvalho (2007). Há também uma explicação adicional, a qual diz que "o SVM basicamente desenha margens entre as classes. As margens são feitas de uma maneira que a distância entre as margens e as classes é a máxima e, portanto, reduz o erro de classificação" (MAHESH, 2016). A ilustração na Figura 5 nos auxilia no entendimento.

Figura 5 - Exemplo de Support Vector Machine



Fonte: Lorena e Carvalho (2007)

Na Figura 5, tem-se um hiperplano formado pelas linhas diagonais. Em linhas gerais, o SVM tem a intenção de encontrar o hiperplano que melhor separa as classes e maximiza a distância entre os pontos mais próximos das classes distintas, representados pelas figuras geométricas (triângulos e círculos).

3.1.1.1 Modelo Support Vector Machine (SVM) Lineares e Não-Lineares

As SVM's inicialmente foram concebidas para a separação de classes através de uma função linear, porém nem sempre isso é possível em cenários reais. Para o caso de separação de notícias falsas, não-falsas e satíricas, segundo Abonizio et al (2020), em seu estudo de detecção de *fake news* em múltiplos idiomas, o modelo que forneceu o melhor resultado dentre os utilizados para a língua portuguesa foi o SVM com função linear utilizando-se os parâmetros padrões do Sci-Kit Learn.

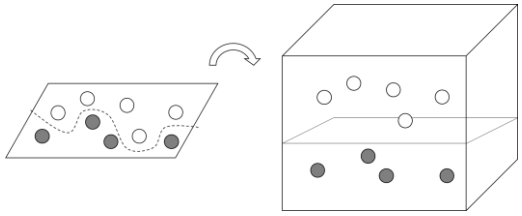
Novamente segundo o estudo de SVM's de Lorena e Carvalho (2007), as SVM's não lineares seguem o mesmo raciocínio dos SVM's lineares, porém há um ajuste com a utilização das funções em dimensões maiores. Por exemplo, uma função em R^2, um espaço bidimensional, que é não-linear, pode se tornar uma função linear em um espaço em R^3, tridimensional. Esse

mapeamento pode ser complexo, porém através da utilização de Kernels, há uma simplificação do cálculo.

Kernels são funções auxiliares utilizadas para realizar o mapeamento de uma função em uma dimensão para uma dimensão diferente. Para o entendimento, supõe-se que existe um mapeamento φ tal que $R^n \rightarrow R^m$, assim o produto escalar de dois vetores em R^m, **x** e **y** por exemplo, será $\varphi(x)^T \varphi(y)$. Assim, a função Kernel é uma função *k* que corresponde ao produto escalar $k(x, y) = \varphi(x)^T \varphi(y)$ (HOFMANN; SCHOLKOPF; SMOLA, 2008).

O melhor entendimento do processo de linearização auxiliado pela utilização de Kernels pode ser visto na Figura 6 abaixo. Nela, temos a transformação de uma função não linear de separação de duas classes em um espaço bidimensional, para uma função linear de separação de duas classes em um espaço tridimensional. Essa função utilizada é o chamado Kernel (φ).

Figura 6 - Exemplo de utilização de Kernel para linearização



Fonte: adaptado de Lorena e Carvalho (2007)

Conforme já apontado por Lorena e Carvalho (2007), para o SVM, entende-se que a intenção é encontrar uma dimensão superior na qual exista uma separação linear entre as diferentes classes, e comumente são utilizadas funções polinomiais, gaussianas e de sigmoides como Kernel. Adicionalmente, segundo Mahesh (2016), essa tarefa pode ser chamada de "truque de Kernel".

Seguindo o raciocínio em Monteiro et al (2018), o qual considera o SVM para a tarefa de detecção de *fake news*, este estudo seguirá considerando a aplicação com um Kernel linear para o SVM.

3.1.1.2 FORMULAÇÃO MATEMÁTICA APLICADA EM ALGORITMOS

Conforme apontado em Lorena e Carvalho (2007), tem-se a seguinte formulação matemática aplicada em algoritmos para o SVM linear, a qual pode ser melhor compreendida seguindo a ilustração na Figura 5.

Tem-se um conjunto de treinamento x_i com $\forall i$=*1,2,3, ..., D* em um problema de separação de duas classes linearmente separáveis, w_1 e w_2, no qual cada amostra associa-se $y_i = 1$ quando $x_i \in w_1$ e $y_i = -1$ quando $x_i \in w_2$ (LORENA; CARVALHO, 2007).

Portanto:

$$D(x) = \sum_{i=1}^{D} w_i x_i + b \tag{6}$$

E se for levado em conta a equação vetorial do produto interno tem-se:

$$D(x) = w^T . x + b \tag{7}$$

Na equação 7 (LORENA; CARVALHO, 2007), observa-se o produto interno entre a transposta do vetor $\boldsymbol{w}$ e o vetor $\boldsymbol{x}$, o que é equivalente à equação 1 (LORENA; CARVALHO, 2007). Além disso, considera-se que *w* é um vetor *d*-dimensional (pesos) e *b* um termo independente em ambas as equações. Os exemplos de treinamento devem ser considerados linearmente separáveis, ou seja:

$$x_i w + b \geq 1 \, para \, y_i = 1 \tag{8}$$
$$x_i w + b \leq 1 \, para \, y_i = -1 \tag{9}$$

Combinando as igualdades 8 e 9 (LORENA; CARVALHO, 2007), pode-se obter:

$$y_i(x_i w + b) - 1 \geq 0, \qquad com \, \forall i = 1,2,3, \ldots, D \tag{10}$$

Também das equações 8 e 9, e analisando a Figura 5, tem-se que a distância entre os hiperplanos *H1* e *H2*, que nos fornece a margem, resulta em $\frac{2}{||w||}$, e para maximizar os dois hiperplanos que geram essa margem, minimiza-se $||w||^2$, e assim tem-se:

$$min_{w,b} = \frac{1}{2}||w||^2 \tag{11}$$

O problema originado na equação 6 (LORENA; CARVALHO, 2007) pode ser resolvido com a utilização da técnica de multiplicadores de Lagrange, acrescentando-se os multiplicadores α_i, obtendo-se assim:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{D} \alpha_i [y_i(w. x_i + b) - 1] \tag{12}$$
$$Com \, \alpha_i \, , (\forall \text{i} = 1,2,3, \ldots, \text{D})$$

É necessário minimizar a equação 12 (LORENA; CARVALHO, 2007) com relação a **w** e b e com o resultado restante maximizar com relação a $\alpha \geq 0$, conhecido como um problema de otimização dual.

Em um ponto ótimo, as seguintes equações de ponto de cela são:

$$\frac{\partial L}{\partial b} = 0 \quad e \quad \frac{\partial L}{\partial w} = 0 \tag{13}$$

Gerando respectivamente,

$$\sum_{i=1}^{D} \alpha_i y_i = 0 \quad e \quad w = \sum_{i=1}^{D} \alpha_i y_i x_i = 0 \tag{14}$$

Substituindo esses resultados na formulação lagrangeana, maximiza-se:

$$max_\alpha \ W(\alpha) = \sum_{i=1}^{D} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{D} \alpha_i \alpha_j y_i y_j (x_i . x_j) \tag{15}$$

$$Sujeito\ a\ \alpha_i \geq 0, \forall i = 1,2,3,\dots,D \ e \sum_{i=1}^{D} \alpha_i y_i$$

Agora deve-se tomar a derivada da equação 15 (LORENA; CARVALHO, 2007) em relação a α e igualá-la a zero, tendo-se em vista que o produto escalar entre todos os x_i e x_j é conhecido.

$$\sum_{i=1}^{D} \alpha_i y_i \quad e \quad 0 \leq \alpha_i < C \tag{16}$$

Sabendo-se α_i é possível encontrar os pesos **w** que maximizam a separação entre os hiperplanos.

$$w = \sum_{i=1}^{D} \alpha_i y_i x_i \tag{17}$$

Após a obtenção dos pesos **w**, através da equação 17 (LORENA; CARVALHO, 2007), pelo treinamento deste método, dado um **ponto *u* desconhecido**, medido com *features* x_i, é possível classifica-lo através da observação do sinal da equação 18 (LORENA; CARVALHO, 2007). Destaca-se que o ponto **u** é um ponto de interesse a ser classificado após o treinamento do modelo.

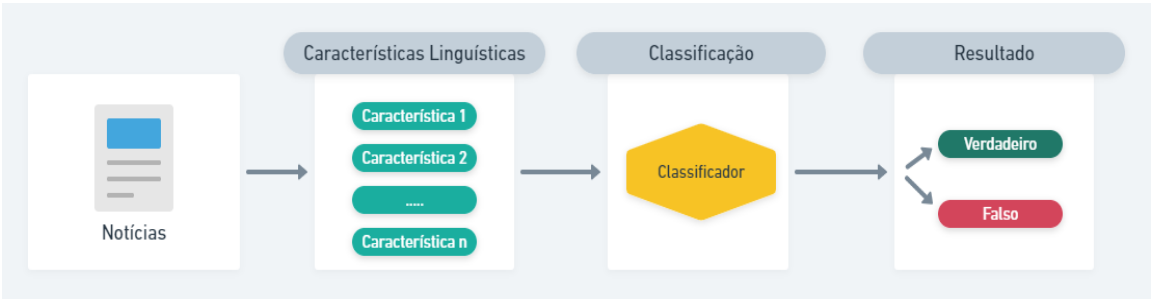
$$f(x) = w.u + b = (\sum_{i=1}^{D} a_i y_i x_i . u) + b \quad (18)$$

3.2 TÉCNICAS DE DETECÇÃO DE FAKE NEWS

Os métodos de detecção automática de *fake news* se dividem basicamente em dois tipos: o *social context-based* e *content-based*. Na mídia tradicional, normalmente o método se baseia no conteúdo da notícia, enquanto em mídias sociais podem ser utilizadas informações adicionais, do contexto de dispersão da notícia, a fim de auxiliar a detecção. Vale ressaltar que o foco deste trabalho é discutir modelos computacionais automatizados, mesmo embora existam outras possibilidades (SHU et al, 2017).

Para a detecção de *fake news* é necessária a extração das características da notícia, seja em um método *social context-based* ou em um *content-base*, conforme indicado na Figura 7.

Figura 7 - Workflow macroscópico da tarefa de detecção de *fake news*



Fonte: Baseado em Shu et al (2017)

3.2.1 CONTENT - BASED

O método *content-based* se baseia no conteúdo da informação, podendo se subdividir no baseado em linguística e no visual. A intenção é utilizar informações como a fonte e/ou autor, título do texto, texto principal, imagens ou vídeos. Ao analisar esses atributos, são criadas estratégias de extração de características discriminativas das notícias que podem ser utilizadas para essa tarefa (SHU et al, 2017).

3.2.2 CONTEXT - BASED

O *social context-based* se utiliza de dados do contexto da informação, principalmente o padrão de difusão da informação em redes sociais e outras plataformas, e não do conteúdo em si, conforme encontrado nos trabalhos de Granskogen (2018).

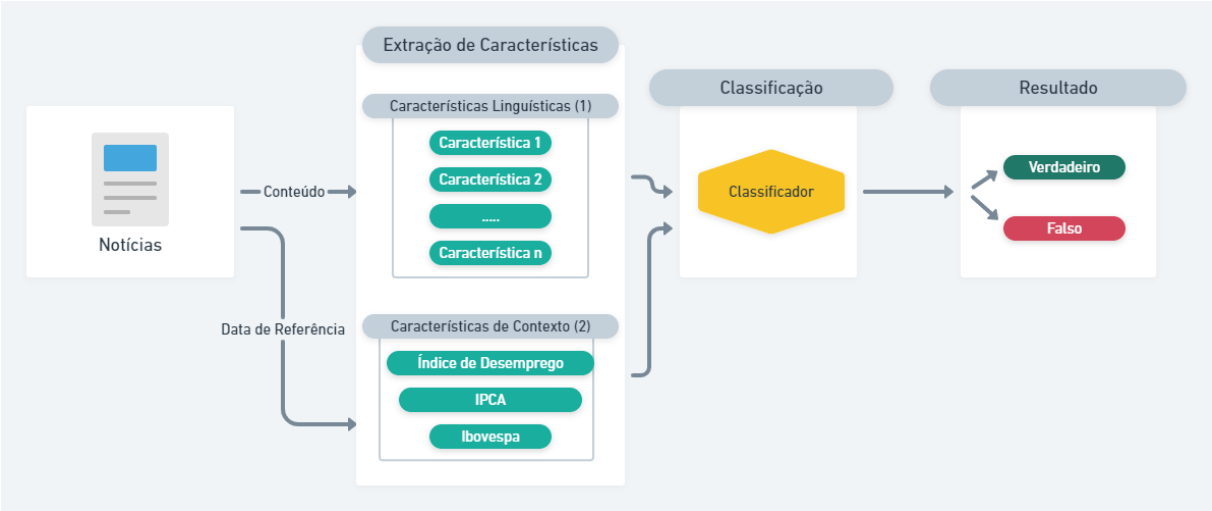
Existem três aspectos a serem considerados: usuários, posts gerados e redes. Para os usuários, é interessante capturar o padrão dos perfis, já que eventualmente notícias falsas são geradas e espalhadas por robôs, e por esse motivo é interessante entender a credibilidade dos perfis que originam e dispersam tais notícias. Além disso, é interessante coletar as reações às postagens geradas (como as "curtidas" no Facebook), já que comportamentos não-orgânicos podem insinuar a presença de robôs que viriam a gerar engajamento e visibilidade da notícia. Outro aspecto mais complexo é o padrão de rede da notícia, que indica o comportamento e difusão entre perfis e redes sociais diferentes (SHU et al, 2017).

3.2.3 MODELO HÍBRIDO

Há também a possibilidade de utilização de modelos híbridos, conforme apontado por Bin Guo et al (2020). Nesse modelo há a combinação do método *social context-based* com o *content-based*. Conforme indicado no estudo de Shu et al (2017), o método *social context-based* se utiliza de informações bem definidas do padrão de difusão das notícias em redes sociais, já o *content-based* se utiliza do conteúdo da notícia em si para uma abordagem linguística.

Neste estudo tem-se a intenção de considerar o modelo de características já utilizado por Monteiro et al (2018), o qual já validou diversas características linguísticas que devem ser utilizadas, e serão exploradas em sessões posteriores. Adicionalmente, este estudo pretende incluir novas variáveis de contexto, não sendo oriundas de redes sociais como no método clássico (*social context-based*), mas sim um contexto de indicadores macroeconômicos e sociais do país, os quais serão explorados em sessões posteriores. A hipótese que se tem aqui é que determinadas condições podem ter correlação com o surgimento de *fake news* e poderiam ser variáveis de contexto que auxiliariam na classificação correta de uma *fake news*, sendo essas consideradas variáveis quantitativas adicionadas ao modelo.

Figura 8 - Workflow macroscópico da tarefa de detecção de fake news considerando variáveis adicionais de contexto social e econômico



Fonte: Adaptado de Shu et al (2017)

3.3 NATURAL LANGUAGE PROCESSING

O processamento de linguagem natural, na sigla em inglês NLP (*Natural Language Processing)*, surgiu no começo no começo dos anos 50, sendo uma intersecção entre as áreas de Inteligência Artificial e Linguística, conforme indicado por Nadkarni, Ohno-Machado e Chapman (2011). Para a retirada das características dos textos que servem como entrada nos modelos e classificadores, atualmente são utilizadas técnicas de vetorização de palavras, as quais transformam textos em vetores de números, segundo explicado por Bokka et al (2019).

Conforme validado por Monteiro et al (2018), embora houve a introdução de variáveis adicionais em seus testes, o que trouxe resultados satisfatórios foi a utilização da conhecida técnica *Bag of Words*. O *Bag of Word* trouxe resultados interessantes nas quatro métricas de avaliação de desempenho (Precisão, Acurácia, *F-Score* e *Recall*) adotadas por Monteiro et al (2018).

Já que a adição de outras variáveis linguísticas não acrescenta mudanças significativas nos resultados (MONTEIRO et al, 2018), tem-se como hipótese para o estudo que a utilização do *Bag of Words* nos modelos, acrescido das variáveis adicionais propostas, é um objeto de estudo relevante para avaliar possíveis melhorias na detecção de *fake news* com os algoritmos SVM, *Random Forest* e *Naive Bayes.*

Para a correta aplicação do *Bag of Words* são considerados os conceitos que ocorrem no pré-processamento do texto, como remoção de *stopwords*, pontuação, letras maiúsculas e *stemming*.

Nota-se que há um desempenho bastante satisfatório do *Bag of Words*, mas além disso, Sebastiani (2002) em seu estudo comenta que implementações mais sofisticadas não acrescentaram uma efetividade melhor.

3.3.1 BAG OF WORDS

O *Bag of Words*, conforme apresentado por Bokka et al (2019), pode ser executado contando-se a quantidade de cada palavra no texto, e assim a saída seria uma matriz onde cada linha representa um documento e as colunas as palavras. Sendo assim, teremos uma matriz M [m,n], com *m* textos ou documentos e *n* palavras consideradas, conforme indicado na Figura 9.

Figura 9 - *Bag of Words*

	Palavra 1	**Palavra 2**	**Palavra 3**	**....**	**Palavra *n***
Documento 1	Contagem	Contagem	Contagem	Contagem	Contagem
Documento 2	Contagem	Contagem	Contagem	Contagem	Contagem
Documento 2	Contagem	Contagem	Contagem	Contagem	Contagem
....	....	....	....	....	....
Documento *m*	Contagem	Contagem	Contagem	Contagem	Contagem

Fonte: Adaptado de Bokka et al (2019)

Para o trecho de texto abaixo, tem-se o seguinte resultado para os vetores do *Bag of Words*:

Texto: "A avaliação dos alunos segue os critérios impostos no início do semestre. Cada semestre será composto por uma avaliação prática e uma teórica.

Figura 10 - Vetor do *Bag of Words* por Frequência

Palavra	A	avaliação	dos	alunos	segue	os	critérios	impostos	no	início
#	1	3	1	1	1	1	1	1	1	1

Palavra	do	semestre	Cada	será	composto	por	uma	prática	e	teórica
#	1	2	1	1	1	1	2	1	1	1

Fonte: autor

Figura 11 - Vetor do *Bag of Words* por Presença (binário)

Palavra	A	avaliação	dos	alunos	segue	os	critérios	impostos	no	início
#	1	1	1	1	1	1	1	1	1	1

Palavra	do	semestre	Cada	será	composto	por	uma	prática	e	teórica
#	1	1	1	1	1	1	1	1	1	1

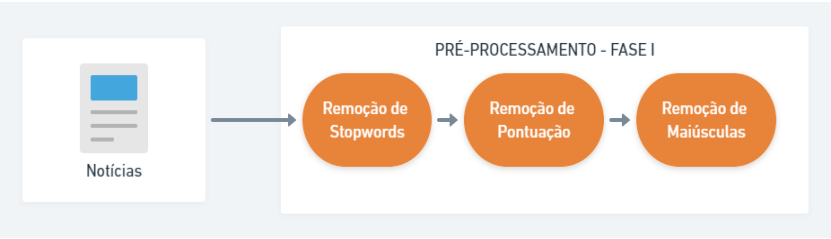
Fonte: autor

As Figuras 10 e 11 demonstram as duas possibilidades para o *Bag of Words*, sendo que em uma delas observa-se a contagem com a frequência de cada palavra, já na outra tem-se apenas uma indicação binária; 1 se a palavra está presente, 0 se a palavra não está presente.

3.3.1.1 PRÉ-PROCESSAMENTO E STEMMMING

Anteriormente ao *Bag Of Words* é importante, conforme comentado por Agarwal et al (2019), executar a limpeza de ruídos dos textos. Isso inclui palavras e termos que expressam pouco sentido relevante e ainda há a redução da dimensionalidade do vetor. Nesse caso, pode-se executar o framework abaixo, Figura 12, para seguir com essa etapa de pré-processamento, chamada neste trabalho como Pré-Processamento Fase I.

Figura 12 - Fase I de Pré-Processamento



Fonte: autor

Para a remoção de *stopwords*, há a possibilidade de se usar dicionários de *stopwords* prontos em língua portuguesa. Exemplos de *stopwords* são: o, a, de e para. Conforme já mencionado, em Argawal et al (2019) tem-se claro que as *stopwords* trazem pouca relevância e expressam pouco sentido nos textos.

Após a execução da Fase I, existe a necessidade de execução da fase de *stemming*.

3.3.1.2 STEMMMING

Conforme apontado em Porter (1998), a remoção automática de sufixos é bastante importante no campo de recuperação de informações. Porter (1998) demonstra em seu estudo um algoritmo bastante eficiente para a tarefa conhecida como *stemming*, que é uma tarefa de remoção de sufixos.

O *stemming*, conforme exemplo abaixo presente na Figura 13, procura retirar sufixos de palavras distintas, mas que possuem um prefixo comum e mantêm sentidos similares (PORTER, 1998).

Figura 13 - Exemplo em inglês de stemming

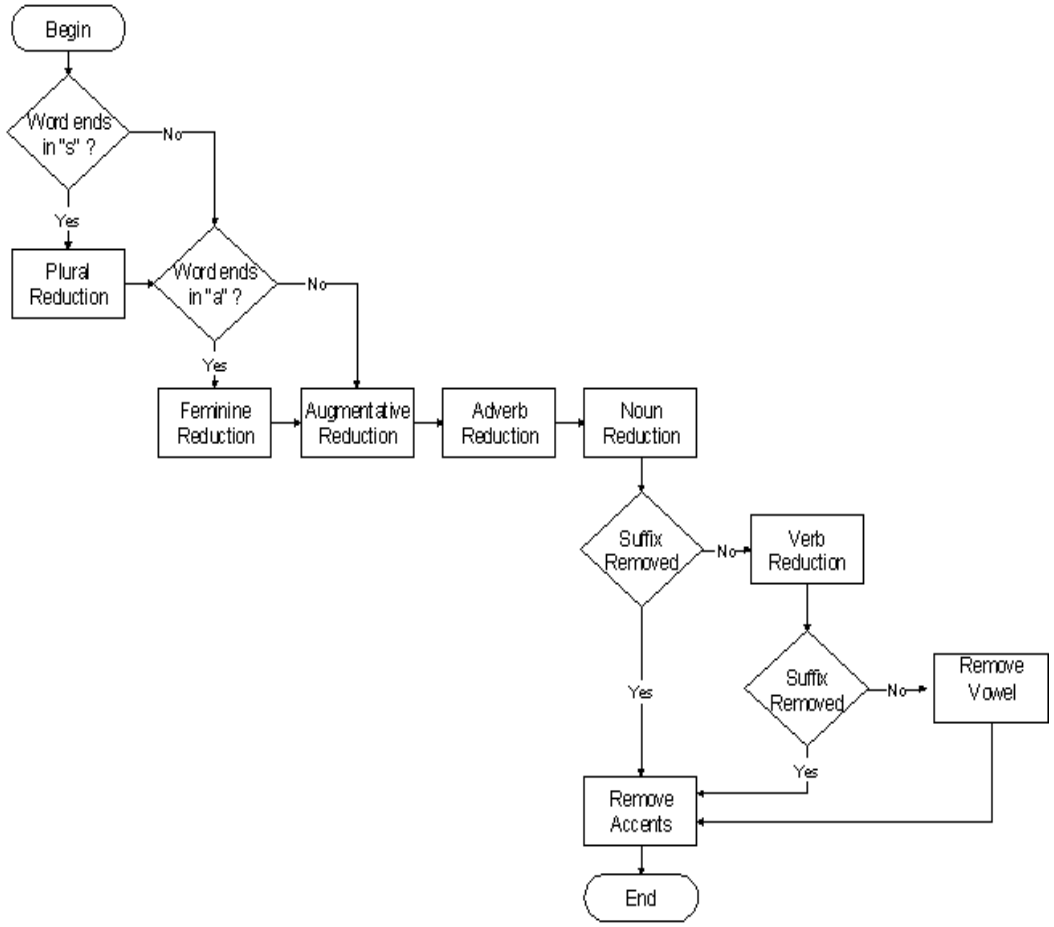
Fonte: Adaptador de Porter (1998)

O algoritmo de Porter (1998) se tornou bastante conhecido e eficiente para a língua inglesa. Após isso, Porter (1998) elabora um novo algoritmo para língua portuguesa, porém com pouca especificidade para o português, já que foi inspirado no algoritmo específico para a língua inglesa. Por isso, Orengo e Huyck (2001) elaboraram um algoritmo específico para o Português que apresenta um resultado mais satisfatório, o qual está representado na Figura 15. Além disso, na Figura 14 temos um exemplo da aplicação do algoritmo de Orengo e Huyck.

Figura 14 – Exemplos de stemming utilizando o algoritmo de Orengo e Huyck

Fonte: autor

Figura 15 - Algoritmo de Stemming para português



Fonte: Orengo e Huyck (2001)

Para este estudo, nomeia-se essa fase de Fase 2 do Pré-Processamento, conforme Figura 16.

Figura 16 - Pré-Processamento completo (Fases I e II)

Fonte: autor

O algoritmo de Orengo e Hyuck (2001) passa por 8 etapas, são elas:

- Redução de plurais

- Redução de femininos
- Redução de advérbios
- Redução de aumentativos/diminutivos
- Redução de substantivos e adjetivos
- Redução de verbos
- Remoção de vogais
- Remoção de acentos

Baseando-se no algoritmo de Orengo e Huyck (2001), Coelho (2007) prepara um novo algoritmo que acrescenta melhorias, com a disponibilização de documentação específica para facilitar a sua implementação por outros pesquisadores.

3.4 BASES DE DADOS CONSIDERADAS

Para o correto treinamento dos modelos utilizados, obtenção das características linguísticas dos textos e obtenção dos dados sociais e macroeconômicos há a necessidade de obtenção de dados de algumas bases. Para esse estudo serão consideradas as bases notícias Fake.Br Corpus (MONTEIRO et al, 2018), FakeRecogna (GARCIA; AFONSO; PAPA, 2022) além de dados do IPCA e Índice de Desemprego, obtidos através do site oficial do IBGE[1], e o Índice Bovespa, obtido do site oficial da B3[2].

3.4.1 FAKE.BR CORPUS

Monteiro et al (2018) preparam em seu estudo o Fake.Br Corpus, o qual é composto de notícias falsas e verdadeiras, sendo 3.600 notícias de cada tipo, as quais foram coletadas e classificadas manualmente. As notícias falsas foram coletadas dos portais Diário do Brasil, Folha do Brasil e Top Five TV, já as notícias verdadeiras foram coletadas dos portais G1, Folha de São Paulo e Estadão. Majoritariamente, as notícias coletadas datam do intervalo de janeiro de 2016 à janeiro de 2018.

[1] https://www.ibge.gov.br/indicadores
[2] https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-estatisticas-historicas.htm

Abaixo, no Quadro 1, seguem exemplos de notícias do Fake.Br Corpus, os quais foram extraídos do estudo de Monteiro et al (2018). Sendo duas notícias falsas e duas notícias verdadeiras.

Quadro 1- Exemplo de notícias da base Fake.Br

Notícia	Classificação
Michel Temer propõe fim do carnaval por 20 anos, “PEC dos gastos”. Michel Temer afirmou que não deve haver gastos com aparatos supérfluos sem pensar primeiramente na educação do Brasil. A medida pretende calcelar o carnaval de 2018.	FALSA
Acabou a mordomia! Ingresso mais barato pra mulher é legal. Baladas que davam meia entrada para mulher, ou até mesmo gratuidade, esto na ilegalidade agora. Acabou o preconceito com os homens nas casas de show de todo o Brasil.	FALSA
Michel Temer não quer o fim do Carnaval por 20 anos. Notícias falsas misturam proximidade dos festejos, crise econômica e medidas impopulares do governo do peemedebista.	VERDADEIRA
Ingresso feminino barato como marketing ‘não inferioriza mulher’, diz juíza do DF. Afirmação consta em decisão sobre preços diferentes para homens e mulheres em festa no Lago Paranoa. ‘Prática permite que mulher possa optar por participar de tais eventos sociais’, diz texto	VERDADEIRA

Fonte: Adaptado de Monteiro et al (2018)

3.4.1 FAKERECOGNA CORPUS

Garcia, Afonso e Papa (2022) preparam em seu estudo o FakeRecogna Corpus, o qual é composto de notícias falsas e verdadeiras, sendo 4628 falsas, coletadas de portais reconhecidos pela apuração de *fake news*. Os portais são o Boatos.Org, o Fato ou Fake (grupo Globo), o E-Farsas, o UOL Confere, o AFP Checamos e o Projeto Comprova. Já para as notícias verdadeiras, as quais totalizam 4193 notícias, foram considerados os portais UOL, G1, Extra, além de páginas governamentais oficiais.

Abaixo, no Quadro 2, seguem quatro exemplos de notícias do FakeRecogna Corpus. Sendo duas falsas e duas verdadeiras.

Quadro 2 - Exemplo de notícias da base FakeRecogna

Notícia	Classificação
Enquanto o ex lutador Maguila, ganhava muito dinheiro, tinha saúde, andava cercado de amigos e familiares. Agora, passando por dificuldades, doente, quase sem nenhum dinheiro, foi abandonado pelos filhos, mulher e amigos e se encontra internado em um asilo. Adílson Maguila foi um dos maiores lutadores mundiais de boxe, ganhando vários títulos nacionais e internacionais, levando o esporte brasileiro aos principais noticiários do mundo. Maguila gravou um vídeo na clínica, pedindo socorro…	FALSA
Prefeito de Lisboa Portugal fala sobre chinêses que espalho covid-19 no mundo. Ouçam o que o prefeito de Lisboa afirma . Quando o filho de Bolsonaro culpou a China por causa do coronavirus ouve repercussão no mundo todo, agora o prefeito de Lisboa em Portugal fala a mesma coisa e explica com detalhe como a China espalhou o vírus para o mundo.	FALSA
COPENHAGEN, 19 NOV (ANSA) - A mutação do coronavírus Sars-CoV-2 detectada em visons na Dinamarca foi "provavelmente extinta", informou o Ministério da Saúde hoje. "Não há novos casos da versão da mutação Cluster 5 desde o dia 15 de setembro, o que induz as autoridades sanitárias dinamarquesas a concluir que essa versão foi, com toda a probabilidade, extinta", informa a nota oficial...	VERDADEIRA
Após o governador do estado de Minas Gerais, Romeu Zema (Novo), se posicionar contra o processo de impeachment do presidente Jair Bolsonaro (sem partido) ao dizer ser "contrário à posição do partido", o Novo foi às redes sociais esclarecer o "mal-entendido" gerado pelo dirigente e disse que "não há decisão formada sobre o impeachment". Durante entrevista à Rádio 98FM na quarta-feira (27), Zema afirmou que, apesar de não concordar com tudo o que o presidente faz —citando a condução de Bolsonaro diante dos problemas gerados pela pandemia e a questão da vacinação no Brasil—, ele não era a favor da abertura de um processo de impeachment. "Eu vi a posição do partido, e sou contrário à ela", afirmou Zema.	VERDADEIRA

Fonte: Adaptado de Garcia, Afonso e Papa (2022)

3.4.3 Base de Dados Sociais e Macroeconômicos

Conforme citado anteriormente, serão considerados dados macroeconômicos e sociais. Os dados de IPCA e Índice de Desemprego foram obtidos no site oficial do IBGE.

De acordo com o seu site oficial, o IBGE, Instituto Brasileiro de Geografia e Estatística, se define como o principal provedor de dados e informações do País (IBGE). Duas das variáveis de interesse podem ter os seus dados oficiais obtidos no site oficial do IBGE; IPCA e Índice de Desemprego, conforme indicado abaixo:

- IPCA: O IPCA - Índice Nacional de Preços ao Consumidor Amplo - aponta a variação do custo de vida médio de famílias com renda mensal de 1 e 40 salários-mínimos (IBGE). Os dados utilizados foram obtidos da base histórica mensal do site do IBGE[3] e foram tratados para verificar a possível correlação e possível aplicação na classificação de uma *fake news*.

- Índice de Desemprego: O desemprego se refere às pessoas com idade para trabalhar que não estão trabalhando, mas estão disponíveis e tentam encontrar trabalho. É calculado pela Pesquisa Nacional por Amostra de Domicílios Contínua, conhecida como PNAD Contínua (IBGE). Os dados utilizados foram obtidos da base histórica mensal do site do IBGE [4]e foram tratados para verificar a possível correlação e possível aplicação na classificação de uma *fake news*.

Além dos dados de IPCA e Índice de Desemprego, foram considerados os dados do Índice BOVESPA, os quais foram obtidos no site oficial do IBOVESPA B3. Conforme apontado no site oficial da IBOVESPA B3 (IBOVESPA B3): "O Ibovespa é o principal indicador de desempenho das ações negociadas na B3 e reúne as empresas mais importantes do mercado de capitais brasileiro. Foi criado em 1968 e, ao longo desses 50 anos, consolidou-se como referência para investidores ao redor do mundo." É também dito no site oficial que "Reavaliado a cada quatro meses, o índice é resultado de uma carteira teórica de ativos. É composto pelas ações e *units* de companhias listadas na B3 que atendem aos critérios descritos na sua metodologia, correspondendo a cerca de **80% do número de negócios e do volume financeiro do nosso mercado de capitais**." (IBOVESPA B3)

O Ibovespa possui índices mínimos e máximos em suas medições mensais, portanto considerou-se o valor médio entre esses pontos, os quais foram tratados para verificar a possível correlação para a aplicação na classificação de uma *fake news*.

Os dados do Ibovespa, PNAD Contínua e IPCA seguem nas Tabelas 1, 2 e 3 contidas na Seção 4.

[3] https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=series-historicas
[4] https://www.ibge.gov.br/estatisticas/sociais/trabalho/9173-pesquisa-nacional-por-amostra-de-domicilios-continua-trimestral.html?=&t=series-historicas

3.4.3.1 NORMALIZAÇÃO DE DADOS

Um modelo de aprendizado realiza o mapeamento de variáveis de entrada em variáveis de saída, no entanto as variáveis podem trabalhar em escalas das mais diversas. Por exemplo, é possível se ter um modelo meteorológico preditivo que trata variáveis de temperatura e velocidade em unidades diferentes (graus celsius e km/h), e essa diferença de unidades pode afetar o modelo preditivo (KUHN; JOHNSON, 2019).

Problemas de classificação se beneficiam da normalização dos dados (HALL; WITTEN, 2011), por isso os dados de índices sociais e macroeconômicos seguirão a normalização conforme abaixo:

$$y = (x - min) / (max - min) \quad (19)$$

Na equação 19 acima tem-se que *y* será o dado final normalizado, *x* é o dado na escala original, e os valores *max* e *min* serão os maiores e menores valores em cada escala considerada (HALL; WITTEN, 2011).

3.5 FRIEDMAN STATISTICAL TEST (FST)

Conforme demonstrado por Demsar (2006), para a comparação do desempenho de diferentes algoritmos em diferentes bases de dados, há diversas abordagens e métodos utilizados. Usualmente, os pesquisadores consideram a média dos desempenhos dos diferentes algoritmos em diferentes bases de dados, porém é objeto de debate se as taxas de erro em domínios diferentes podem ser consideradas comparáveis, tornando assim a média uma medida com pouco valor. Sendo assim, o *Friedman Statistical Test* (FST) traz valor para análise de comparação do desempenho de diferentes algoritmos em diferentes bases de dados.

No FST, a hipótese nula é de que todos os modelos têm desempenho semelhante. Primeiramente deve-se ranquear os algoritmos de acordo com a performance, sendo o primeiro o de melhor performance e o último o de pior performance.

Tem-se r_i^j como o ranque do *j-ésimo* de *k* algoritmos na *i-ésima* base de dados de *N* possíveis e assim o FST compara a média (R_j) dos ranques dos algoritmos, que diz que na hipótese nula todos terão o mesmo ranque. Os valores para a distribuição de Friedman são calculados nas equações 21 e 20 abaixo, sendo que o valor F_F deve ser comparado com os valores críticos da tabela presente na Seção 5, na qual teremos os resultados.

$$\chi_F{}^2 = \frac{12N}{k(k+1)}\left[\sum_j \quad R_j{}^2 - \frac{k(k+1)^2}{4}\right] \tag{20}$$

$$F_F = \frac{(N-1)\chi_F{}^2}{N(k-1) - \chi_F{}^2} \tag{21}$$

Considerando-se este trabalho, teremos a utilização de três algoritmos diferentes (SVM, *Naive Bayes* e *Random Forest*), portanto o valor de *k* será 3. Além disso, teremos 20 conjuntos de bases de testes, tendo bases com e sem os indicadores sociais e macroeconômicos, além de modificações nas bases de testes. Todo esse contexto de cenários será mais bem detalhado na seção 4.3, porém o daqui tira-se que o valor de N será 20. A ideia geral é gerar um ranqueamento do desempenho de cada algoritmo em cada cenário, e então aplicar o método do FST, para garantir que há comparabilidade entre os diversos cenários propostos. Assim, garantindo-se que a hipótese nula, a de que os cenários não são comparáveis, seja rejeitada.

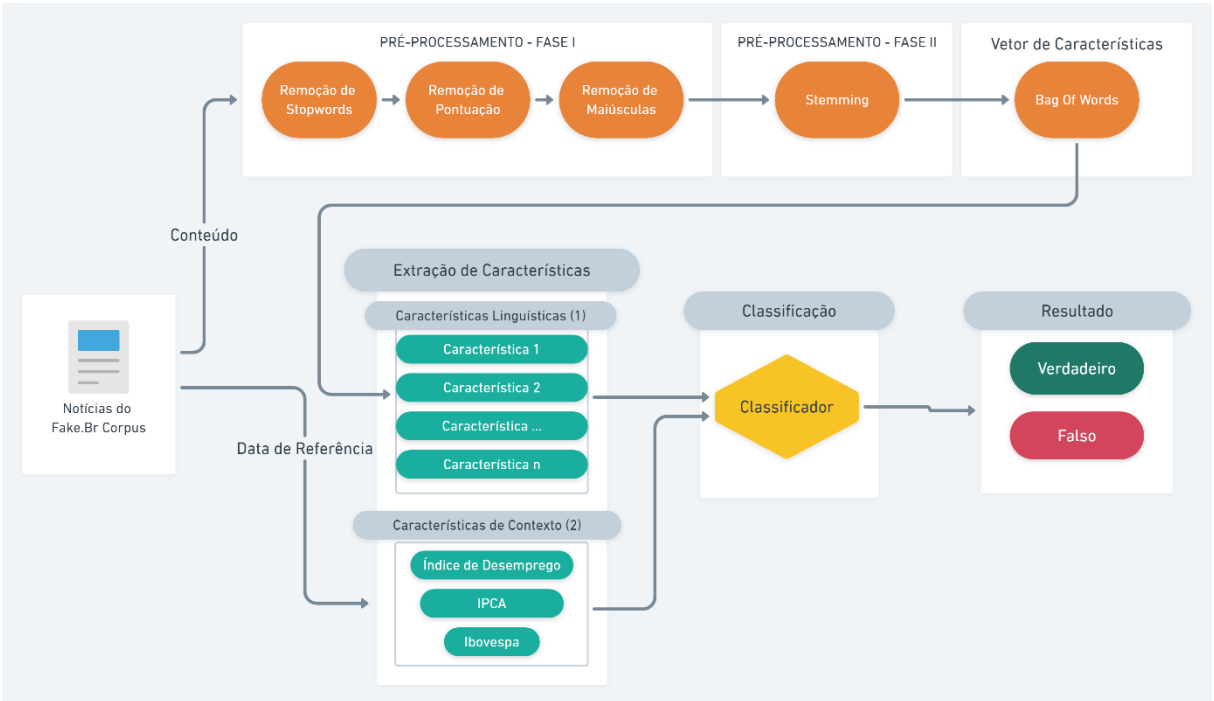
4 METODOLOGIA

Para o treinamento dos modelos a serem considerados, foram utilizadas as bases de dados Fake.Br Corpus, preparada por Monteiro et al (2018), além da base FakeRecogna preparada por Garcia, Afonso e Papa (2022).

Primeiramente, conforme indicado na Figura 17, foram executadas as etapas de Pré-Processamento dos textos das notícias do Fake.Br Corpus, o que inclui a retirada de *stopwords*, remoção de pontuação, remoção de maiúsculas e *stemming* seguindo o algoritmo de Orengo e Hyuck (2001). O modelo *Bag of Words* foi aplicado para a obtenção do vetor de característica a ser inserido no modelo. Após isso, foram utilizados os classificadores propostos (SVM, *Random Forest* e *Naive Bayes*).

Após essa etapa, foram consideradas a inclusão de novas variáveis quantitativas numéricas de contexto social e macroeconômico, na intenção de trabalhar em um modelo de detecção de *fake news* híbrido, semelhante ao que é apontado por Bin Guo et al (2020), porém sem características do padrão de dispersão da notícia, e sim com a inclusão de índices sociais e macroeconômicos.

Figura 17 - Processo com a base Fake.Br Corpus



Fonte: elaboração própria

Após isso, foram criados cenários adicionais que consideram apenas as cinquenta primeiras palavras das bases de notícias, sendo essa uma etapa que ocorre antes da remoção de *stopwords* e, a fim de levar mais robustez aos testes realizados.

Todas as etapas anteriores foram também realizadas com a base FakeRecogna (GARCIA; AFONSO; PAPA, 2022).

4.1 RECURSOS COMPUTACIONAIS E FERRAMENTAS

Para as tarefas de pré-processamento dos textos e das bases de dados, além do treinamento dos modelos e coleta de resultados, foi utilizada a linguagem Python com o auxílio de bibliotecas específicas, como Pandas[5], NumPy[6], Natural Language Toolkit (NLTK) [7] e o Sci Kit Learn[8]. Para o SVM, foi considerado um *kernel* linear e o *Naive Bayes* considerado foi o gaussiano.

Para o treinamento dos modelos foi utilizado o ambiente do Google Colaboratory Pro, com processamento de GPU. Dado que se trata de um ambiente compartilhado, o Google Colaboratory Pro aloca para os seus assinantes, de maneira arbitrária, uma GPU T4 ou P100.

Os códigos implementados e bases consideradas podem ser encontradas no diretório do GitHub mantido pelo autor[9].

4.2 BASES DE DADOS DE NOTÍCIAS

Foram consideradas as bases Fake.Br (MONTEIRO et al, 2018) e FakeRecogna (GARCIA; AFONSO; PAPA, 2022).

A base Fake.Br Corpus possui os textos das notícias e os dados de data de publicação, os quais são importantes para a correlação dos dados de índices sociais e macroeconômicos. A base Fake.Br Corpus encontra-se disponível no repositório GitHub mantido pelos pesquisadores [10].

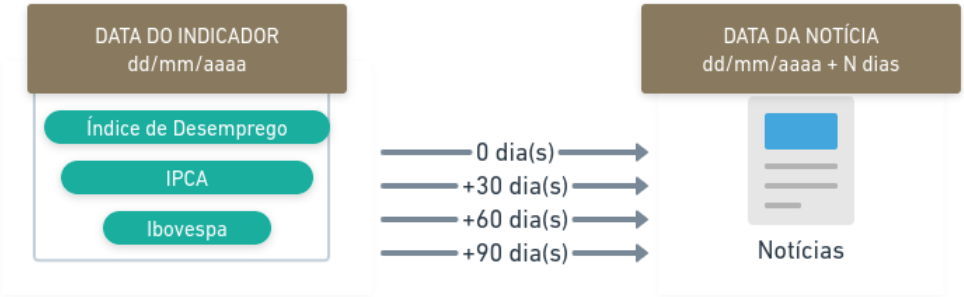
---

[5] Pandas: https://pandas.pydata.org/
[6] Numpy: https://numpy.org/
[7] NLTK: https://www.nltk.org/
[8] Sci Kit Learn: https://scikit-learn.org/stable/
[9] https://github.com/brunocpsc/dissertacao_2022
[10] https://github.com/roneysco/Fake.br-Corpus

Assim como o Fake.Br Corpus, a base FakeRecogna contém, além das notícias, os dados de data de publicação, os quais são importantes para a correlação dos dados de índices sociais e macroeconômicos. A mesma se encontra no repositório do GitHub mantido pelos autores[11].

Vale ressaltar, conforme apontado na Figura 18 abaixo, que se tem como uma hipótese que os efeitos a serem observados no padrão de dispersão de notícias falsas devem ocorrer um tempo após a divulgação dos índices sociais e macroeconômicos, assim será realizado um deslocamento entre a data de referência da divulgação da notícia e a data de referência do índice econômico, variando-se entre 0, 30, 60 e 90 dias. Esse padrão foi escolhido pois alguns dos índices são atualizados em períodos mensais ou trimestrais, e com isso teremos alguns cenários adicionais.

Figura 18 - Processo de deslocamento de datas com a base Fake.Br Corpus



Fonte: autor

4.3 BASES DE DADOS DE ÍNDICES

Os dados de IPCA e índice de desemprego podem ser obtidos do site oficial do IBGE, e a hipótese que se leva em conta é que quanto maior o IPCA do mês de referência, maior se teria uma circulação geral de notícias falsas, algo que pode ser extrapolado para os índices de desempregado coletados pela PNAD Contínua, também disponível no site oficial do IBGE.

Quanto ao IBOVESPA, tem se como hipótese que quanto maior o valor do índice, menor seria a dispersão de notícias falsas na sociedade.

Para o caso dessas variáveis, foi executada a normalização dos valores no período de todas as notícias, obtendo-se um valor entre 0 e 1 como variável de entrada no modelo, conforme indicado na seção 3.4.3.1. Sendo assim, para o IPCA, considerando-se o período entre os anos de 2016 e 2021, tem-se como resultado os valores na Tabela 1.

[11] https://github.com/Gabriel-Lino-Garcia/FakeRecogna

Para o os valores de taxa de desemprego da PNAD contínua, considera-se que os valores para os trimestres podem ser extrapolados e serem considerados para cada mês, conforme os valores contidos na Tabela 2.

Para os valores de taxa IBOVESPA, considera-se os valores mensais presentes na Tabela 3. Nota-se que o IBOVESPA é fornecido mensalmente com valores mínimos e máximos apurados, sendo assim é natural inferir-se a utilização da média desses valores.

Todos os valores das tabelas 1, 2 e 3 foram normalizados conforme a equação 19.

Tabela 1 – Indicadores IPCA para o período de 2016 até 2021

ANO	MÊS	VARIAÇÃO %	NORMALIZADO	ANO	MÊS	VARIAÇÃO %	NORMALIZADO
2016	JAN	1,27	0,95	2019	JAN	0,32	0,4
2016	FEV	0,90	0,74	2019	FEV	0,43	0,47
2016	MAR	0,43	0,47	2019	MAR	0,75	0,65
2016	ABR	0,61	0,57	2019	ABR	0,57	0,55
2016	MAI	0,78	0,67	2019	MAI	0,13	0,29
2016	JUN	0,35	0,42	2019	JUN	0,01	0,23
2016	JUL	0,52	0,52	2019	JUL	0,19	0,33
2016	AGO	0,44	0,47	2019	AGO	0,11	0,28
2016	SET	0,08	0,27	2019	SET	-0,04	0,2
2016	OUT	0,26	0,37	2019	OUT	0,10	0,28
2016	NOV	0,18	0,32	2019	NOV	0,51	0,51
2016	DEZ	0,30	0,39	2019	DEZ	1,15	0,88
2017	JAN	0,38	0,44	2020	JAN	0,21	0,34
2017	FEV	0,33	0,41	2020	FEV	0,25	0,36
2017	MAR	0,25	0,36	2020	MAR	0,07	0,26
2017	ABR	0,14	0,3	2020	ABR	-0,31	0,04
2017	MAI	0,31	0,4	2020	MAI	-0,38	0
2017	JUN	-0,23	0,09	2020	JUN	0,26	0,37
2017	JUL	0,24	0,36	2020	JUL	0,36	0,43
2017	AGO	0,19	0,33	2020	AGO	0,24	0,36
2017	SET	0,16	0,31	2020	SET	0,64	0,59
2017	OUT	0,42	0,46	2020	OUT	0,86	0,72
2017	NOV	0,28	0,38	2020	NOV	0,89	0,73
2017	DEZ	0,44	0,47	2020	DEZ	1,35	1
2018	JAN	0,29	0,39	2021	JAN	0,25	0,36
2018	FEV	0,32	0,4	2021	FEV	0,86	0,72
2018	MAR	0,09	0,27	2021	MAR	0,93	0,76
2018	ABR	0,22	0,35	2021	ABR	0,31	0,4
2018	MAI	0,40	0,45	2021	MAI	0,83	0,7
2018	JUN	1,26	0,95	2021	JUN	0,53	0,53
2018	JUL	0,33	0,41	2021	JUL	0,96	0,77
2018	AGO	-0,09	0,17	2021	AGO	0,87	0,72
2018	SET	0,48	0,5	2021	SET	1,16	0,89
2018	OUT	0,45	0,48	2021	OUT	1,25	0,94
2018	NOV	-0,21	0,1	2021	NOV	0,95	0,77
2018	DEZ	0,15	0,31	2021	DEZ	0,73	0,64

Fonte: Adaptado de IBGE

Tabela 2 - Indicadores PNAD contínua para o período entre 2016 e 2021

ANO	TRIMESTRE	MÊS	VARIAÇÃO %	NORMALIZADO
2016	jan-fev-mar	JAN	11,1	0,00
2016	fev-mar-abr	FEV	11,3	0,05
2016	mar-abr-mai	MAR	11,3	0,05
2016	abr-mai-jun	ABR	11,4	0,08
2016	mai-jun-jul	MAI	11,7	0,16
2016	jun-jul-ago	JUN	11,9	0,21
2016	jul-ago-set	JUL	11,9	0,21
2016	ago-set-out	AGO	11,9	0,21
2016	set-out-nov	SET	12,0	0,24
2016	out-nov-dez	OUT	12,2	0,29
2016	nov-dez-jan	NOV	12,7	0,42
2016	dez-jan-fev	DEZ	13,3	0,58
2017	jan-fev-mar	JAN	13,9	0,74
2017	fev-mar-abr	FEV	13,7	0,68
2017	mar-abr-mai	MAR	13,4	0,61
2017	abr-mai-jun	ABR	13,1	0,53
2017	mai-jun-jul	MAI	12,9	0,47
2017	jun-jul-ago	JUN	12,7	0,42
2017	jul-ago-set	JUL	12,5	0,37
2017	ago-set-out	AGO	12,3	0,32
2017	set-out-nov	SET	12,1	0,26
2017	out-nov-dez	OUT	11,9	0,21
2017	nov-dez-jan	NOV	12,3	0,32
2017	dez-jan-fev	DEZ	12,7	0,42
2018	jan-fev-mar	JAN	13,2	0,55
2018	fev-mar-abr	FEV	13,0	0,50
2018	mar-abr-mai	MAR	12,8	0,45
2018	abr-mai-jun	ABR	12,6	0,39
2018	mai-jun-jul	MAI	12,4	0,34
2018	jun-jul-ago	JUN	12,3	0,32
2018	jul-ago-set	JUL	12,0	0,24
2018	ago-set-out	AGO	11,9	0,21
2018	set-out-nov	SET	11,7	0,16
2018	out-nov-dez	OUT	11,7	0,16
2018	nov-dez-jan	NOV	12,2	0,29
2018	dez-jan-fev	DEZ	12,6	0,39
2019	jan-fev-mar	JAN	12,8	0,45
2019	fev-mar-abr	FEV	12,6	0,39
2019	mar-abr-mai	MAR	12,4	0,34
2019	abr-mai-jun	ABR	12,1	0,26
2019	mai-jun-jul	MAI	12,0	0,24
2019	jun-jul-ago	JUN	11,9	0,21
2019	jul-ago-set	JUL	11,9	0,21
2019	ago-set-out	AGO	11,8	0,18
2019	set-out-nov	SET	11,3	0,05
2019	out-nov-dez	OUT	11,1	0,00
2019	nov-dez-jan	NOV	11,4	0,08
2019	dez-jan-fev	DEZ	11,8	0,18
2020	jan-fev-mar	JAN	12,4	0,34
2020	fev-mar-abr	FEV	12,7	0,42
2020	mar-abr-mai	MAR	13,1	0,53
2020	abr-mai-jun	ABR	13,6	0,66
2020	mai-jun-jul	MAI	14,1	0,79
2020	jun-jul-ago	JUN	14,8	0,97
2020	jul-ago-set	JUL	14,9	1,00
2020	ago-set-out	AGO	14,6	0,92
2020	set-out-nov	SET	14,4	0,87
2020	out-nov-dez	OUT	14,2	0,82
2020	nov-dez-jan	NOV	14,5	0,89
2020	dez-jan-fev	DEZ	14,6	0,92
2021	jan-fev-mar	JAN	14,9	1,00
2021	fev-mar-abr	FEV	14,8	0,97
2021	mar-abr-mai	MAR	14,7	0,95
2021	abr-mai-jun	ABR	14,2	0,82
2021	mai-jun-jul	MAI	13,7	0,68
2021	jun-jul-ago	JUN	13,1	0,53
2021	jul-ago-set	JUL	12,6	0,39
2021	ago-set-out	AGO	12,1	0,26
2021	set-out-nov	SET	11,6	0,13
2021	out-nov-dez	OUT	11,1	0,00
2021	nov-dez-jan	NOV	11,2	0,03
2021	dez-jan-fev	DEZ	11,2	0,03

Fonte: Adaptado de IBGE

Tabela 3 - Indicadores IBOVESPA contínua para o período entre 2016 e 2021

ANO	MÊS	MÍNIMO	MÁXIMO	MÉDIA	NORMALIZADO
2016	Jan	37497,47	42419,32	39958,4	0,00
2016	Fev	38596,16	43234,85	40915,51	0,01
2016	Mar	44121,79	51248,92	47685,36	0,09
2016	Abr	48096,24	54477,78	51287,01	0,13
2016	Mai	48471,7	53561,53	51016,62	0,12
2016	Jun	48648,29	51629,29	50138,79	0,11
2016	Jul	51842,27	57308,2	54575,24	0,16
2016	Ago	56162,38	59323,83	57743,11	0,20
2016	Set	56820,77	60231,65	58526,21	0,21
2016	Out	59339,22	64924,51	62131,87	0,25
2016	Nov	59183,5	64157,67	61670,59	0,24
2016	Dez	57110,99	61414,4	59262,7	0,22
2017	Jan	59588,7	66190,62	62889,66	0,26
2017	Fev	63992,93	69052,02	66522,48	0,30
2017	Mar	62980,37	66988,87	64984,62	0,28
2017	Abr	62826,28	65768,91	64297,6	0,27
2017	Mai	61597,05	68684,49	65140,77	0,28
2017	Jun	60761,74	63170,73	61966,24	0,25
2017	Jul	62322,4	65920,36	64121,38	0,27
2017	Ago	66516,23	71329,85	68923,04	0,33
2017	Set	71923,11	76004,15	73963,63	0,38
2017	Out	74308,49	76989,78	75649,14	0,40
2017	Nov	70826,59	74594,61	72710,6	0,37
2017	Dez	72264,45	76402,08	74333,27	0,39
2018	Jan	77891,04	85530,84	81710,94	0,47
2018	Fev	80898,7	87652,64	84275,67	0,50
2018	Mar	83808,06	86900,43	85354,25	0,51
2018	Abr	82861,58	86444,66	84653,12	0,50
2018	Mai	75355,84	86536,97	80946,41	0,46
2018	Jun	69814,74	78596,06	74205,4	0,39
2018	Jul	72839,74	80275,6	76557,67	0,41
2018	Ago	75180,4	81434,98	78307,69	0,43
2018	Set	74656,51	80000,09	77328,3	0,42
2018	Out	78623,66	87423,55	83023,61	0,48
2018	Nov	84914,11	89709,57	87311,84	0,53
2018	Dez	85136,11	89820,09	87478,1	0,53
2019	Jan	91012,32	97677,2	94344,76	0,61
2019	Fev	94405,6	98588,64	96497,12	0,64
2019	Mar	91903,41	99993,93	95948,67	0,63
2019	Abr	92875	97369,29	95122,15	0,62
2019	Mai	89992,73	97457,36	93725,05	0,61
2019	Jun	95998,75	102062,33	99030,54	0,66
2019	Jul	100605,2	105817,06	103211,1	0,71
2019	Ago	96429,6	104115,23	100272,4	0,68
2019	Set	99680,83	105319,4	102500,1	0,70
2019	Out	99981,4	108407,54	104194,5	0,72
2019	Nov	105864,2	109580,57	107722,4	0,76
2019	Dez	108927,8	117203,2	113065,5	0,82
2020	Jan	113760,6	119527,63	116644,1	0,86
2020	Fev	102983,5	116674,13	109828,8	0,79
2020	Mar	63569,62	107224,22	85396,92	0,51
2020	Abr	69537,56	83170,8	76354,18	0,41
2020	Mai	77556,62	87946,25	82751,44	0,48
2020	Jun	88620,1	97644,67	93132,39	0,60
2020	Jul	96203,2	105605,17	100904,2	0,69
2020	Ago	99369,15	104125,64	101747,4	0,70
2020	Set	93580,35	102167,65	97874	0,65
2020	Out	93952,4	101917,73	97935,07	0,65
2020	Nov	95979,71	110575,47	103277,6	0,71
2020	Dez	111399,9	119409,15	115404,5	0,85
2021	Jan	115067,6	125076,63	120072,1	0,90
2021	Fev	110035,2	120355,79	115195,5	0,85
2021	Mar	110334,8	116849,67	113592,3	0,83
2021	Abr	115253,3	121113,93	118183,6	0,88
2021	Mai	117712	126215,73	121963,9	0,92
2021	Jun	126801,7	130776,27	128789	1,00
2021	Jul	121800,8	128406,51	125103,7	0,96
2021	Ago	116642,6	123576,56	120109,6	0,90
2021	Set	108843,7	119395,6	114119,7	0,83
2021	Out	103500,7	114647,99	109074,4	0,78
2021	Nov	101915,5	107594,67	104755,1	0,73
2021	Dez	100774,6	108326,33	104550,5	0,73

Fonte: Adaptado de B3

Para a introdução dos valores dos índices sociais e macroeconômicos no modelo de classificação como variáveis de entrada se dá o seguinte processo:

I. A coleta dos dados, normalizados, do IBOVESPA, taxa de desemprego da PNAD contínua e IPCA para os períodos das notícias coletadas;
II. Cada notícia, seja na base Fake.Br Corpus ou na FakeRecogna, segue acompanhada da sua data de referência, a qual é utilizada para considerar os dados do item (I) acima como variável de entrada para o mesmo mês e ano de referência.

4.4 MÉTRICAS DE VALIDAÇÃO

Como métricas de avaliação dos resultados dos testes realizados, foram analisadas as medidas de AUC (*area under the curve*), *F1-Score*, Acurácia, *Recall* e *Precision* para as bases de dados, com e sem a aplicação das variáveis adicionais de índices sociais e macroeconômicos.

Conforme apontado por Fawcett (2006), o AUC é especialmente importante para a detecção de *fake news* já que pode ser usado para controlar o *trade off* entre as taxas de falso positivo e falso negativo. Já o *F1-Score* indica a performance geral do classificador. A medida de acurácia é uma medida simples para avaliação de classificadores, porém é um bom indicativo inicial de desempenho em problemas de classificação com exemplos de treinamento balanceados entre as classes, que é o caso do trabalho em questão. A métrica de *Precision* busca indicar dentre todas as classificações de uma classe específica que o modelo executou, quantas estão corretas e realmente pertencem a determinada classe. Já a métrica de *Recall* busca indicar a proporção de previsões corretas de uma determinada classe entre o total de exemplos dessa classe específica.

Para a execução dos testes, foram separadas amostras de treino e teste, com a utilização de reamostragem para a execução de um processo de testes mais robusto.

Abaixo seguem as equações das métricas a serem consideradas.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$F_{measure} = \frac{2.Recall.Precisão}{Recall + Precisão} \tag{23}$$

Para o *F1-Score* consideramos as medidas de *Recall* e Precisão, as quais são obtidas das equações 24 e 25.

$$Precisão = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

As taxas *True Positive (TP), True Negative (TN), False Positive (FP)* e *False Negative (FN)* são consideradas nas medidas de Precisão e *Recall*.

$$FPR = \frac{FP}{TN + FP} \tag{26}$$

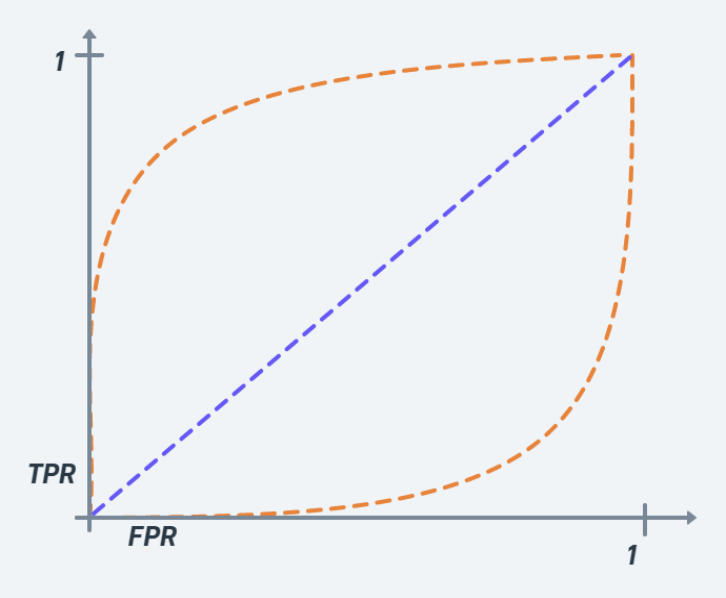
$$Recall = TPR = \frac{TP}{TP + FN} \tag{27}$$

É válido ressaltar também os conceitos de *True Positive, True Negative, False Positive* e *False Negative*, os quais seguem de base para a elaboração de uma Matriz de Confusão, que contém as quatro métricas abaixo:

- *True Positive (TP)* = casos classificados como **notícia verdadeira** pelo modelo e que **de fato são notícias verdadeiras**;
- *True Negative (TN)*= casos classificados como **notícia falsa** pelo modelo e que **de fato são notícias falsas**;
- *False Positive (FP)*= casos classificados como **notícia verdadeira** pelo modelo e que **na verdade são notificas falsas**;
- *False Negative (FN)*= casos classificados como **notícia falsa** pelo modelo e que **não verdade são notificas verdadeiras**.

As medidas de *FPR* (*false positive rate*, em português seria a taxa de falso positivos) e *TPR (true positive rate*, em português seria taxa de verdadeiros positivos) são consideradas para a obtenção da curva AUC, a qual segue com a *TPR* no eixo vertical e a *FPR* no eixo horizontal (modelos com bom desempenho de separação entre classes tendem a ter uma AUC que se aproxima de 1), conforme indicado na Figura 19 abaixo. Nota-se que as curvas laranjas demonstram um bom classificador, já a curva azul demonstra um classificador ineficaz, já que não realiza (AUC=0,5) a diferenciação entre classes (HANLEY; MCNEIL, 1982).

Figura 19 - Exemplos de AUC



Fonte: Adaptado de Hanley e McNeil (1982).

4.5 TESTES EXECUTADOS

Os testes executados seguiram as seguintes etapas:

i) Para ambas as bases (FakeBr e FakeRecogna), completas **e** com somente as cinquenta primeiras palavras de cada notícia, SEM as variáveis de índices sociais e macroeconômicos:

   a. Separação entre as amostras de treino e teste, utilizando-se do método *k-fold cross validation*, com um $k = 5$. Assim tem-se a separação da base de dados em k grupos aleatórios, sendo 80% para treino e 20% para teste.

   b. Para cada vez teremos a coleta dos dados de Acurácia, *Recall*, Precisão e *F1-Score*, além da análise da curva AUC para o valor médio.

   c. As etapas são executadas para os três modelos de classificação selecionados (SVM, *Random Forest* e *Naive Bayes*).

ii) Para ambas as bases (FakeBr e FakeRecogna), completas **e** com somente as cinquenta primeiras palavras de cada notícia, COM as variáveis de índices sociais e macroeconômicos:

a. Separação entre as amostras de treino e teste, utilizando-se do método *k-fold cross validation*, com um $k = 5$. Assim tem-se a separação da base de dados em k grupos aleatórios, sendo 80% para treino e 20% para teste.
b. Para cada vez teremos a coleta dos dados de Acurácia, *Recall*, Precisão e *F1-Score*, além da análise da curva AUC para o valor médio.
c. As etapas são executadas para os três modelos de classificação selecionados (SVM, *Random Forest* e *Naive Bayes*).

Considerando-se as etapas i) e ii) acima, tem-se os seguintes cenários para avaliação:

Tabela 4 - Cenários de teste considerados para base FakeRecogna completa

Cenário	Modelo	Base	Índices
CSVMRecogna	SVM	Recogna	-
CSVMRecogna0d	SVM	Recogna	0d
CSVMRecogna30d	SVM	Recogna	30d
CSVMRecogna60d	SVM	Recogna	60d
CSVMRecogna90d	SVM	Recogna	90d
CNBRecogna	NB	Recogna	-
CNBRecogna0d	NB	Recogna	0d
CNBRecogna30d	NB	Recogna	30d
CNBRecogna60d	NB	Recogna	60d
CNBRecogna90d	NB	Recogna	90d
CRFRecogna	RF	Recogna	-
CRFRecogna0d	RF	Recogna	0d
CRFRecogna30d	RF	Recogna	30d
CRFRecogna60d	RF	Recogna	60d
CRFRecogna90d	RF	Recogna	90d

Fonte: autor

Na Tabela 4, acima, tem-se os cenários que contemplam os testes com os modelos SVM, *Naive Bayes* e *Random Forest* com a base FakeRecogna. Há cenários de testes que não contemplam a inserção de índices sociais e macroeconômicos, e cenário que contempla os testes com índices sociais e macroeconômicos com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (indicados por 0d, 30d, 60d e 90d na última coluna da tabela).

Tabela 5 - Cenários de teste considerados para base FakeRecogna considerando-se apenas as **cinquenta primeiras palavras de cada notícia**

Cenário	Modelo	Base	Índices
CSVMRecogna50W	SVM	Recogna 50W	-
CSVMRecogna50W0d	SVM	Recogna 50W	0d
CSVMRecogna50W30d	SVM	Recogna 50W	30d
CSVMRecogna50W60d	SVM	Recogna 50W	60d
CSVMRecogna50W90d	SVM	Recogna 50W	90d
CNBRecogna50W	NB	Recogna 50W	-
CNBRecogna50W0d	NB	Recogna 50W	0d
CNBRecogna50W30d	NB	Recogna 50W	30d
CNBRecogna50W60d	NB	Recogna 50W	60d
CNBRecogna50W90d	NB	Recogna 50W	90d
CRFRecogna50W	RF	Recogna 50W	-
CRFRecogna50W0d	RF	Recogna 50W	0d
CRFRecogna50W30d	RF	Recogna 50W	30d
CRFRecogna50W60d	RF	Recogna 50W	60d
CRFRecogna50W90d	RF	Recogna 50W	90d

Fonte: autor

Na Tabela 5, acima, tem-se os cenários que contemplam os testes com os modelos SVM, *Naive Bayes* e *Random Forest* com a base FakeRecogna modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Ressalta-se aqui que foi determinada de maneira arbitrária a quantidade de palavras a ser considerada. Há cenários de testes que não contemplam a inserção de índices sociais e macroeconômicos, e cenário que contempla os testes com índices sociais e macroeconômicos com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (indicados por 0d, 30d, 60d e 90d na última coluna da tabela).

Tabela 6 - Cenários de teste considerados para base Fake.Br completa

Cenário	Modelo	Base	Índices
CSVMFake.BrCorpus	SVM	Fake.Br Corpus	-
CSVMFake.BrCorpus0d	SVM	Fake.Br Corpus	0d
CSVMFake.BrCorpus30d	SVM	Fake.Br Corpus	30d
CSVMFake.BrCorpus60d	SVM	Fake.Br Corpus	60d
CSVMFake.BrCorpus90d	SVM	Fake.Br Corpus	90d
CNBFake.BrCorpus	NB	Fake.Br Corpus	-
CNBFake.BrCorpus0d	NB	Fake.Br Corpus	0d
CNBFake.BrCorpus30d	NB	Fake.Br Corpus	30d
CNBFake.BrCorpus60d	NB	Fake.Br Corpus	60d
CNBFake.BrCorpus90d	NB	Fake.Br Corpus	90d
CRFFake.BrCorpus	RF	Fake.Br Corpus	-
CRFFake.BrCorpus0d	RF	Fake.Br Corpus	0d
CRFFake.BrCorpus30d	RF	Fake.Br Corpus	30d
CRFFake.BrCorpus60d	RF	Fake.Br Corpus	60d
CRFFake.BrCorpus90d	RF	Fake.Br Corpus	90d

Fonte: autor

Na Tabela 6, acima, tem-se os cenários que contemplam os testes com os modelos SVM, *Naive Bayes* e *Random Forest* com a base Fake.Br. Há cenários de testes que não contemplam a inserção de índices sociais e macroeconômicos, e cenário que contempla os testes com índices sociais e macroeconômicos com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (indicados por 0d, 30d, 60d e 90d na última coluna da tabela).

Tabela 7 - Cenários de teste considerados para base Fake.Br considerando-se apenas as **cinquenta primeiras palavras de cada notícia**

Cenário	Modelo	Base	Índices
CSVMFake.BrCorpus50W	SVM	Fake.Br Corpus 50W	-
CSVMFake.BrCorpus50W0d	SVM	Fake.Br Corpus 50W	0d
CSVMFake.BrCorpus50W30d	SVM	Fake.Br Corpus 50W	30d
CSVMFake.BrCorpus50W60d	SVM	Fake.Br Corpus 50W	60d
CSVMFake.BrCorpus50W90d	SVM	Fake.Br Corpus 50W	90d
CNBFake.BrCorpus50W	NB	Fake.Br Corpus 50W	-
CNBFake.BrCorpus50W0d	NB	Fake.Br Corpus 50W	0d
CNBFake.BrCorpus50W30d	NB	Fake.Br Corpus 50W	30d
CNBFake.BrCorpus50W60d	NB	Fake.Br Corpus 50W	60d
CNBFake.BrCorpus50W90d	NB	Fake.Br Corpus 50W	90d
CRFFake.BrCorpus50W	RF	Fake.Br Corpus 50W	-
CRFFake.BrCorpus50W0d	RF	Fake.Br Corpus 50W	0d
CRFFake.BrCorpus50W30d	RF	Fake.Br Corpus 50W	30d
CRFFake.BrCorpus50W60d	RF	Fake.Br Corpus 50W	60d
CRFFake.BrCorpus50W90d	RF	Fake.Br Corpus 50W	90d

Fonte: autor

Na Tabela 7, acima, tem-se os cenários que contemplam os testes com os modelos SVM, *Naive Bayes* e *Random Forest* com a base Fake.Br modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Há cenários de testes que não contemplam a inserção de índices sociais e macroeconômicos, e cenário que contempla os testes com índices sociais e macroeconômicos com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (indicados por 0d, 30d, 60d e 90d na última coluna da tabela).

Com os cenários indicados, foi executado o *Friedman Statistical Test* (FST) para os valores de acurácia obtidos nos testes. Caso os valores de acurácia possam ser comparados, poderemos realizar a comparação para todas as outras métricas.

Com isso, é possível verificar se a hipótese nula pode ser rejeitada, e assim garantir que é possível realizar a comparação entre diferentes bases de dados com diferentes modelos classificatórios. Nesse caso, tem-se que a hipótese nula é de que os modelos têm desempenhos semelhantes nas bases testadas.

5 RESULTADOS

Conforme indicado na seção anterior, o FST foi executado para que seja avaliado se a hipótese nula pode ser rejeitada. Os resultados seguem na Tabela 8 abaixo:

Tabela 8 - Análise FST

Base de Texto	Índices	SVM	NB	RF	RANKS		
Recogna	-	0,9019386	0,7402778	0,917923	2	3	1
Recogna	0d	0,9044324	0,7402778	0,9273322	2	3	1
Recogna	30d	0,9063594	0,7402778	0,9251786	2	3	1
Recogna	60d	0,9030718	0,7402778	0,9249518	2	3	1
Recogna	90d	0,903299	0,7402778	0,9255186	2	3	1
Recogna 50W	-	0,8879942	0,715791	0,9068132	2	3	1
Recogna 50W	0d	0,895703	0,715791	0,9189436	2	3	1
Recogna 50W	30d	0,8949096	0,715791	0,9165626	2	3	1
Recogna 50W	60d	0,8902616	0,715791	0,9199636	2	3	1
Recogna 50W	90d	0,8872008	0,715791	0,9133886	2	3	1
Fake.Br Corpus	-	0,8881944	0,7204166	0,881528	1	3	2
Fake.Br Corpus	0d	0,8884722	0,7204166	0,8788886	1	3	2
Fake.Br Corpus	30d	0,8894444	0,7204166	0,8819444	1	3	2
Fake.Br Corpus	60d	0,8881946	0,7204166	0,8830556	1	3	2
Fake.Br Corpus	90d	0,8877778	0,7204166	0,881528	1	3	2
Fake.Br Corpus 50W	-	0,8308332	0,7202778	0,8429168	2	3	1
Fake.Br Corpus 50W	0d	0,8311114	0,7202778	0,8365276	2	3	1
Fake.Br Corpus 50W	30d	0,8312502	0,7202778	0,8409722	2	3	1
Fake.Br Corpus 50W	60d	0,8338888	0,7202778	0,8408332	2	3	1
Fake.Br Corpus 50W	90d	0,8318056	0,7202778	0,8419446	2	3	1

Fonte: autor

Do FST, têm-se que k=3, já que são considerados três algoritmos sendo utilizados (SVM, NB e RF), e N=20, já que são considerados 20 cenários possíveis para cada algoritmo, conforme indicado nas tabelas 4, 5, 6 e 7. Daí, utilizando-se as equações 21 e 22, encontra-se o valor de F_F será 82,33, que é maior que valor crítico indicado na Tabela 9, rejeitando-se assim a hipótese nula e, dado isso, é possível prosseguir-se com a análise dos demais resultados. Vale ressaltar que as métricas de Acurácia, Precisão, *Recall* e *F1-Score* foram obtidas através do *k-fold cross validation*, e os resultados apresentados são a média entre os valores obtidos. Para a curva AUC é seguido o mesmo racional.

Uma informação importante que já é possível obter da Tabela 5 é que (destaques coloridos), dentre 20 cenários possíveis com diferentes classificadores e bases, o algoritmo *Random Forest* é que o detém o melhor desempenho. O SVM oscila entre o segundo e o melhor desempenho, já o *Naive Bayes* é o que detém, sempre, o pior desempenho.

Tabela 9 - Valores Críticos do FST para 20 cenários com três classificadores e α=5%

n	Valor Crítico
2	-
3	6,000
4	6,500
5	6,400
6	7,000
7	7,143
8	6,250
9	6,222
10	6,200
11	6,545
12	6,500
13	6,615
14	6,143
15	6,400
16	6,500
17	6,118
18	6,333
19	6,421
20	6,300

Fonte: Demsar (2006)

Abaixo seguem subseções com os resultados para os cenários apresentados nas tabelas 4, 5, 6 e 7. Há uma subseção para cada algoritmo e cada base, portanto têm-se subseções para os resultados com os algoritmos SVM, *Naive Bayes* e *Random Forest*, e cada uma delas com as bases FakeRecogna e Fake.Br, além de serem consideradas as bases modificadas, com apenas as cinquenta primeiras palavras de cada notícia.

5.1 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKERECOGOGNA

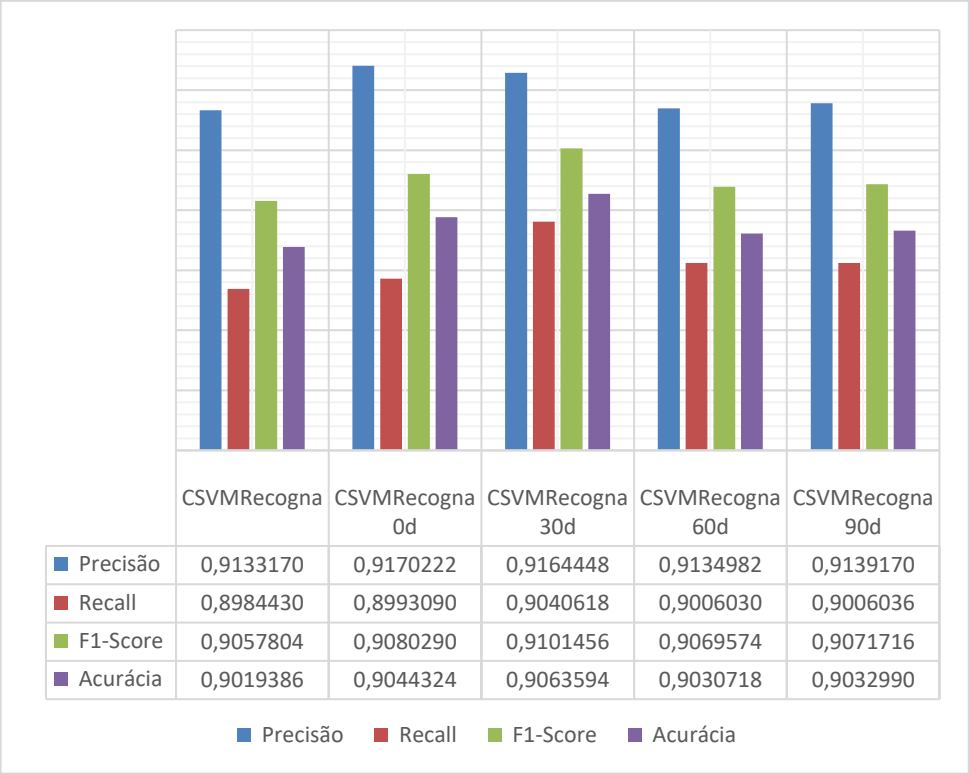
Na Figura 20 observam-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo SVM** e a **base FakeRecogna** sem indicadores sociais e macroeconômicos, e com indicadores sociais e macroeconômicos com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme a Tabela 4).

Vale já ressaltar que é possível observar na Figura 20 uma melhoria visível em todos as métricas para os cenários com indicadores sociais e macroeconômicos, com deslocamentos principalmente entre 0 e 30 dias.

Além disso, nas Figuras 21, 22, 23, 24 e 25 há as curvas AUC para os cenários indicados, porém dado o aumento singelo, não se nota uma melhoria visual chamativa para as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 20, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

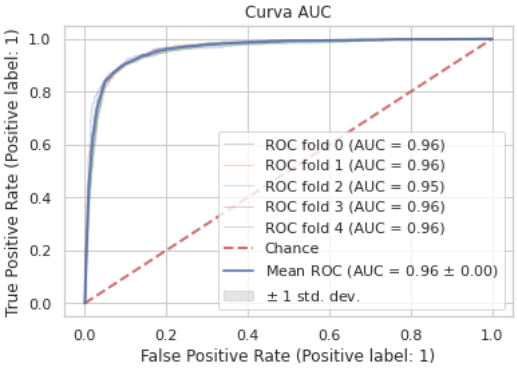
Figura 20 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CSVMRecogna	CSVMRecogna 0d	CSVMRecogna 30d	CSVMRecogna 60d	CSVMRecogna 90d
Precisão	0,9133170	0,9170222	0,9164448	0,9134982	0,9139170
Recall	0,8984430	0,8993090	0,9040618	0,9006030	0,9006036
F1-Score	0,9057804	0,9080290	0,9101456	0,9069574	0,9071716
Acurácia	0,9019386	0,9044324	0,9063594	0,9030718	0,9032990

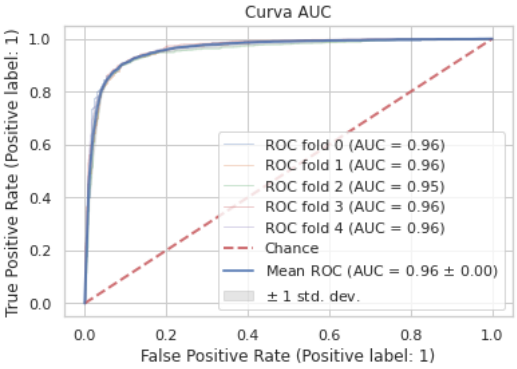
Fonte: autor

Figura 21 - Curva AUC Cenário CSVMRecogna



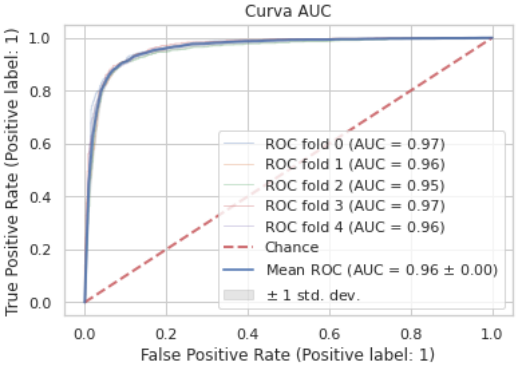
Fonte: autor

Figura 22 - Curva AUC Cenário CSVMRecogna0d



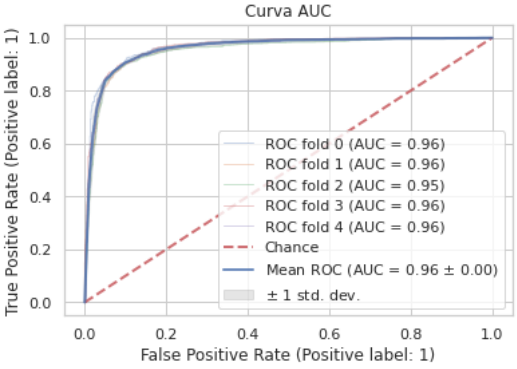
Fonte: autor

Figura 23 - Curva AUC Cenário CSVMRecogna30d



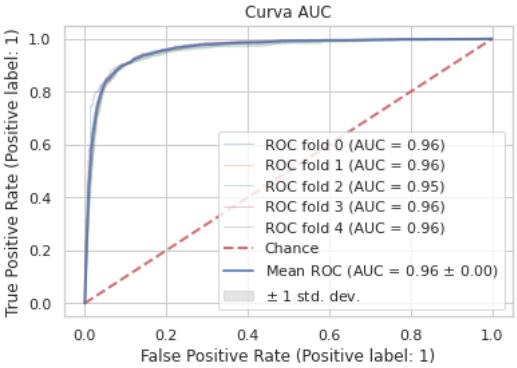
Fonte: autor

Figura 24 - Curva AUC Cenário CSVMRecogna60d



Fonte: autor

Figura 25 - Curva AUC Cenário CSVMRecogna90d



Fonte: autor

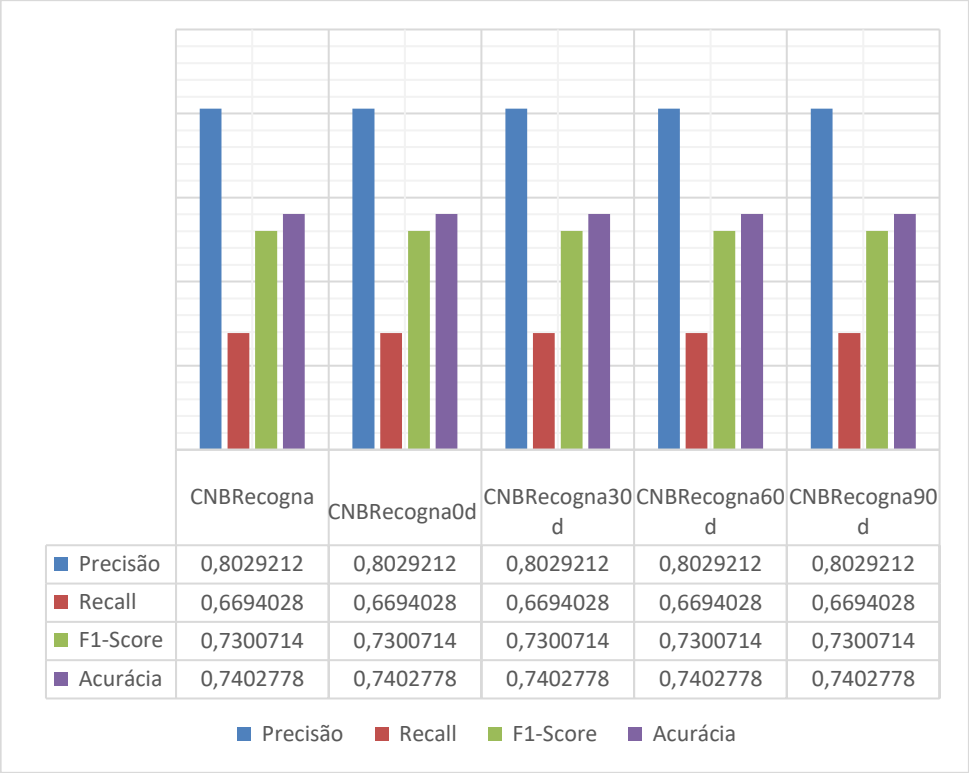
5.2 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKERECOGOGNA

Na Figura 26 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo *Naive Bayes*** e a **base FakeRecogna** sem indicadores sociais e macroeconômicos e com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 4).

Para o modelo *Naive Bayes*, diferente do que ocorre com os modelos SVM e *Random Forest*, não se nota qualquer alteração nos cenários propostos, algo que se reflete nas curvas AUC indicadas nas Figuras 27, 28, 29, 30 e 31.

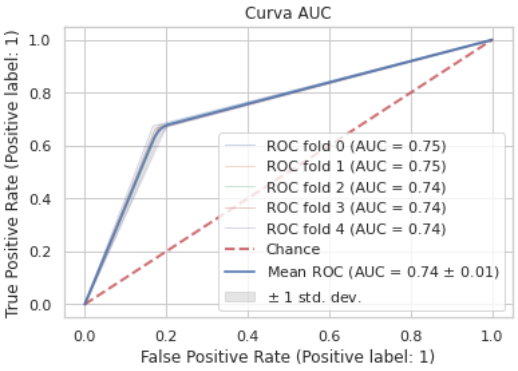
Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 26, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

Figura 26 - Precisão, Recall, F1-Score e Acurácia os Cenários CNBRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias



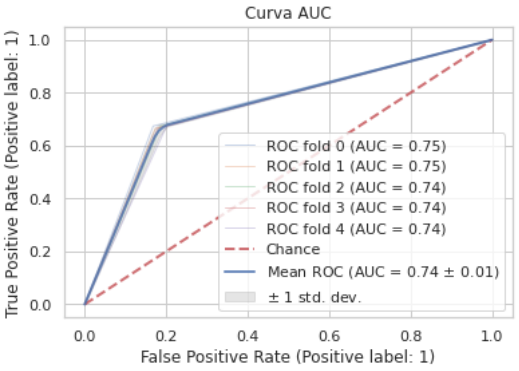
Fonte: autor

Figura 27 - Curva AUC Cenário CNBRecogna



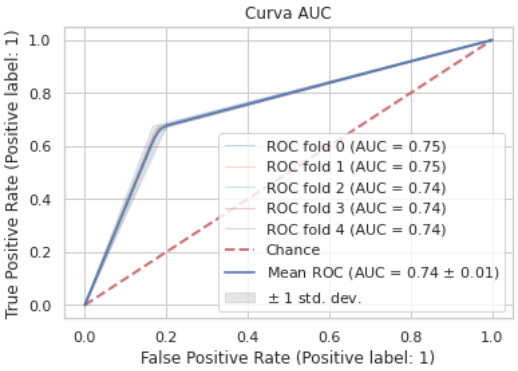
Fonte: autor

Figura 28 - Curva AUC Cenário CNBRecogna0d



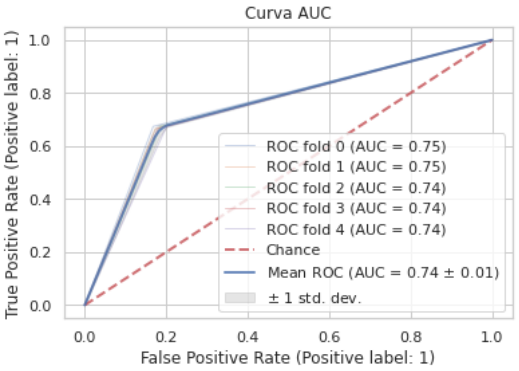
Fonte: autor

Figura 29 - Curva AUC Cenário CNBRecogna30d



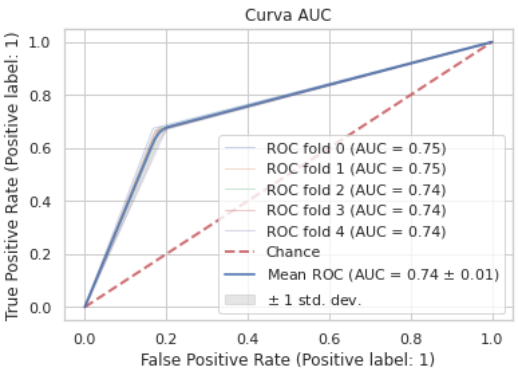
Fonte: autor

Figura 30 - Curva AUC Cenário CNBRecogna60d



Fonte: autor

Figura 31 - Curva AUC Cenário CNBRecogna90d



Fonte: autor

5.3 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKERECOGOGNA

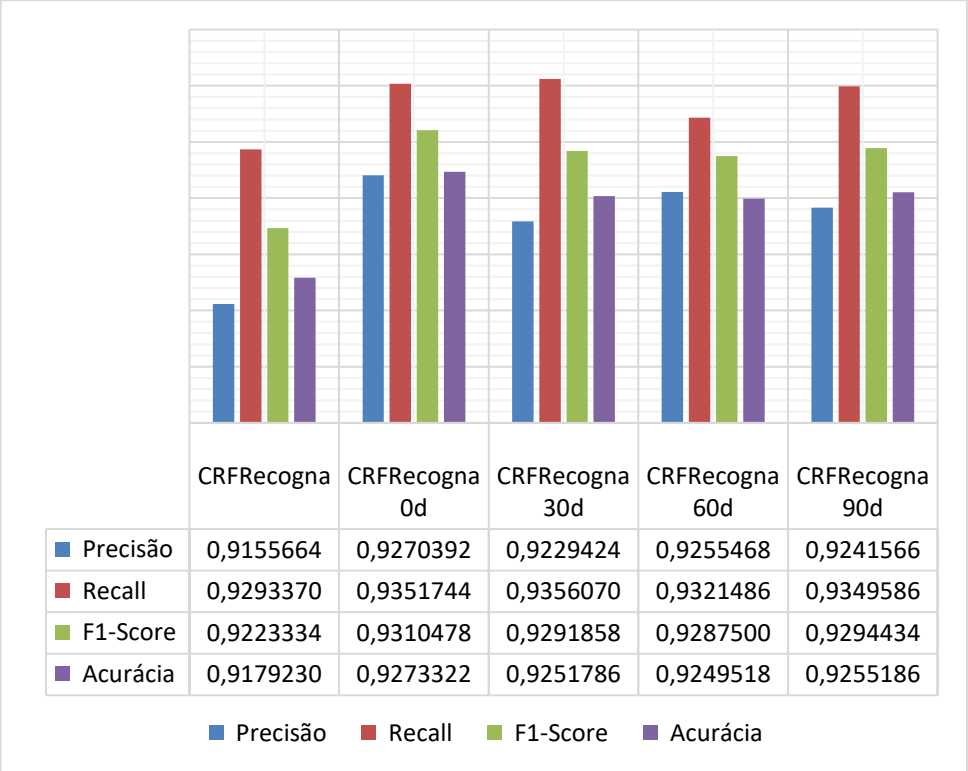
Na Figura 32 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários **com o modelo *Random Forest*** e a **base FakeRecogna** sem indicadores sociais e macroeconômicos e com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 4).

Vale já ressaltar que é possível observar na Figura 30 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos principalmente entre 0 e 30 dias.

Além disso, nas Figuras 33, 34, 35, 36 e 37 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 20, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

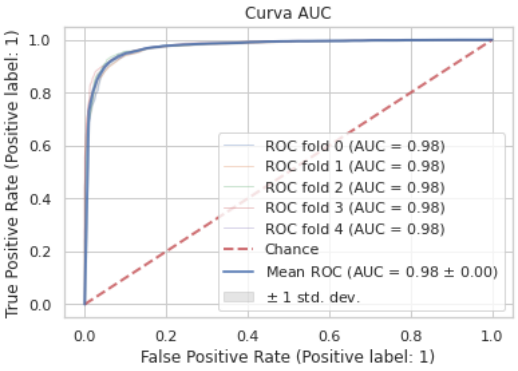
Figura 32 - Precisão, Recall, F1-Score e Acurácia para os Cenários CRFRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CRFRecogna	CRFRecogna 0d	CRFRecogna 30d	CRFRecogna 60d	CRFRecogna 90d
Precisão	0,9155664	0,9270392	0,9229424	0,9255468	0,9241566
Recall	0,9293370	0,9351744	0,9356070	0,9321486	0,9349586
F1-Score	0,9223334	0,9310478	0,9291858	0,9287500	0,9294434
Acurácia	0,9179230	0,9273322	0,9251786	0,9249518	0,9255186

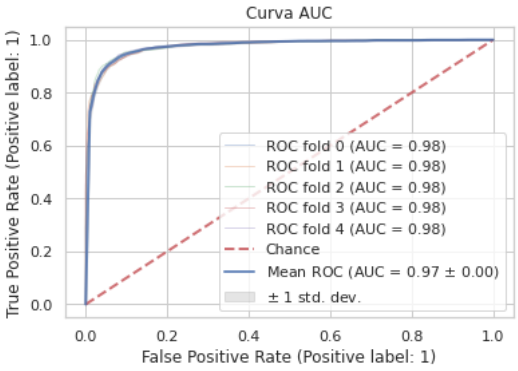
Fonte: autor

Figura 33 - Curva AUC Cenário CRFRecogna



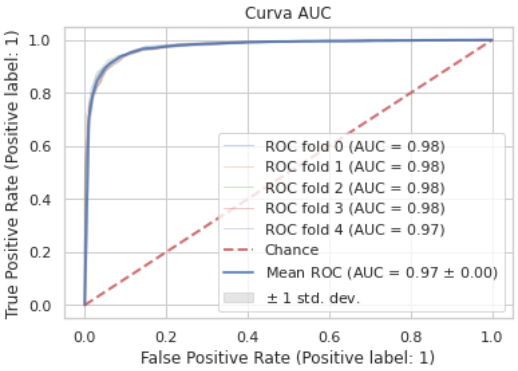
Fonte: autor

Figura 34 - Curva AUC Cenário CRFRecogna0d



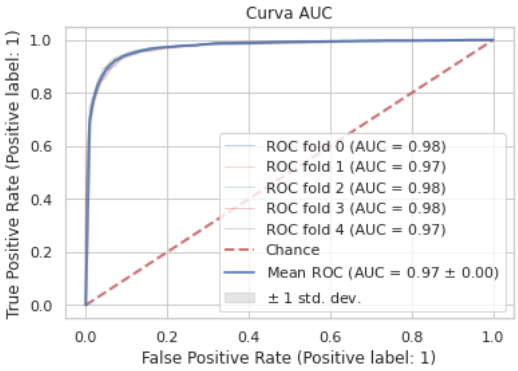
Fonte: autor

Figura 35 - Curva AUC Cenário CRFRecogna30d



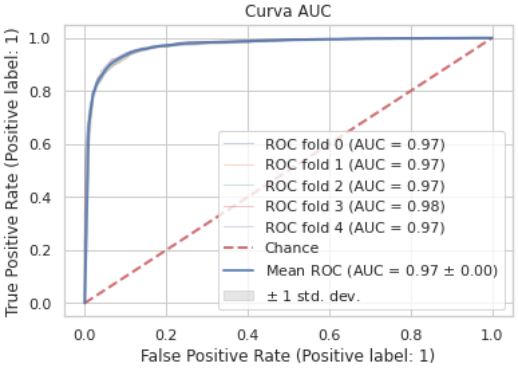
Fonte: autor

Figura 36 - Curva AUC Cenário CRFRecogna60d



Fonte: autor

Figura 37 - Curva AUC Cenário CRFRecogna90d



Fonte: autor

5.4 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKERECOGOGNA50W

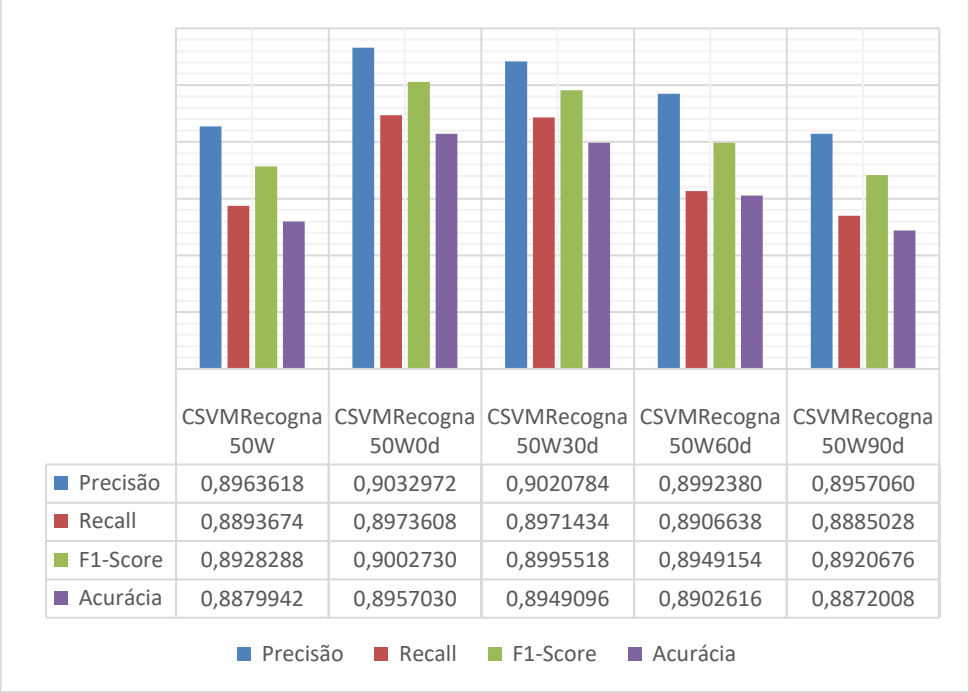
Na Figura 38 tem-se os resultados de Precisão, Recall, F1-Score e Acurácia para os cenários com o **modelo SVM** e a **base FakeRecogna modificada**. Essa modificação considera apenas as cinquenta primeiras palavras de cada notícia da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 5).

Vale já ressaltar que é possível observar na Figura 36 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos principalmente entre 0 e 30 dias.

Além disso, nas Figuras 39, 40, 41, 42 e 43 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 38, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

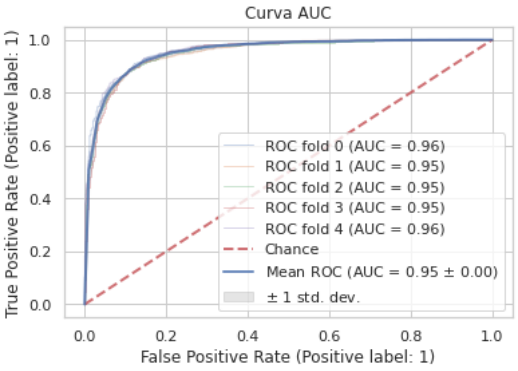
Figura 38 - Precisão, Recall, F1-Score e Acurácia para os Cenários CSVMRecogna50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CSVMRecogna 50W	CSVMRecogna 50W0d	CSVMRecogna 50W30d	CSVMRecogna 50W60d	CSVMRecogna 50W90d
Precisão	0,8963618	0,9032972	0,9020784	0,8992380	0,8957060
Recall	0,8893674	0,8973608	0,8971434	0,8906638	0,8885028
F1-Score	0,8928288	0,9002730	0,8995518	0,8949154	0,8920676
Acurácia	0,8879942	0,8957030	0,8949096	0,8902616	0,8872008

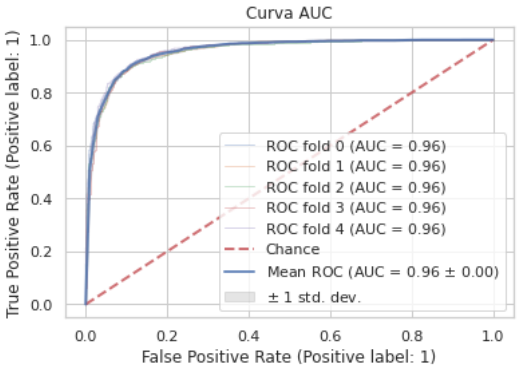
Fonte: autor

Figura 39 - Curva AUC Cenário CSVMRecogna50W



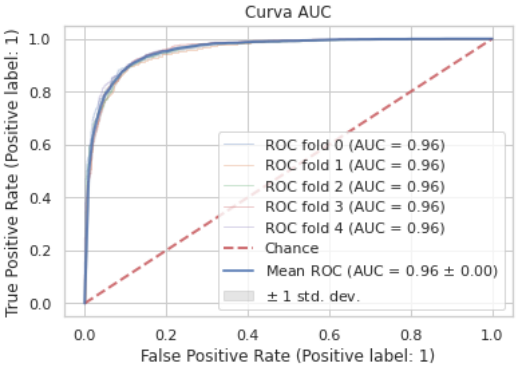
Fonte: autor

Figura 40 - Curva AUC Cenário CSVMRecogna50W0d



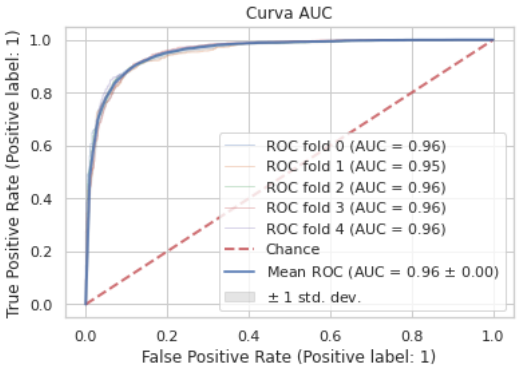
Fonte: autor

Figura 41 - Curva AUC Cenário CSVMRecogna50W30d



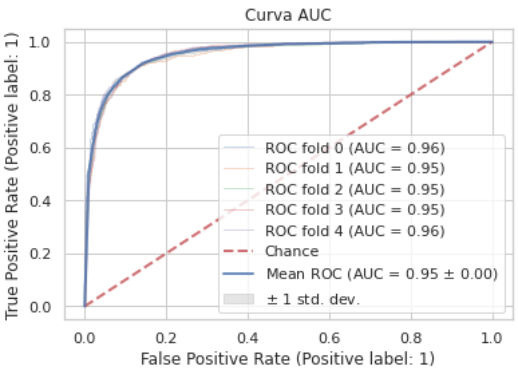
Fonte: autor

Figura 42 - Curva AUC Cenário CSVMRecogna50W60d



Fonte: autor

Figura 43 - Curva AUC Cenário CSVMRecogna50W90d



Fonte: autor

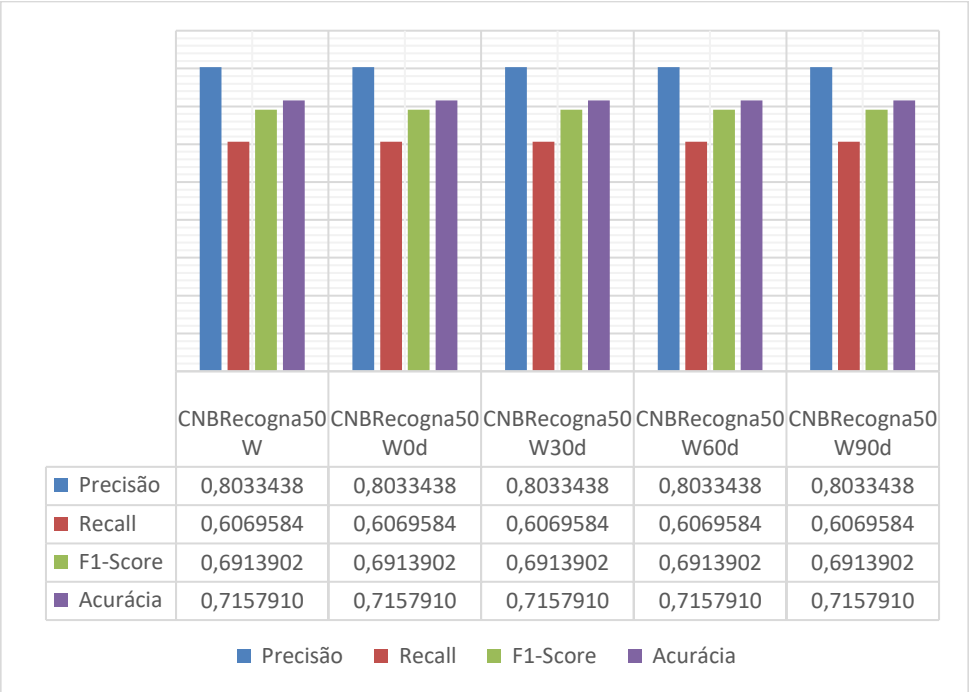
5.5 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKERECOGOGNA50W

Na Figura 44 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo *Naive Bayesx*** e a **base FakeRecogna** modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 5).

Para o modelo NB, diferente do que ocorre com os modelos SVM e RF, não se nota qualquer alteração nos cenários propostos, algo que se reflete nas curvas AUC indicadas nas Figuras 45, 46, 47, 48 e 49.

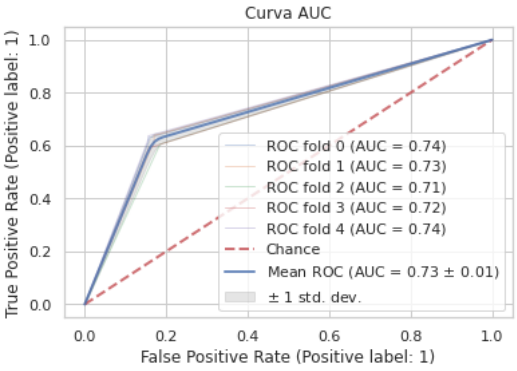
Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 44, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

Figura 44 - Precisão, Recall, F1-Score e Acurácia para os Cenários CNBRecogna50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias



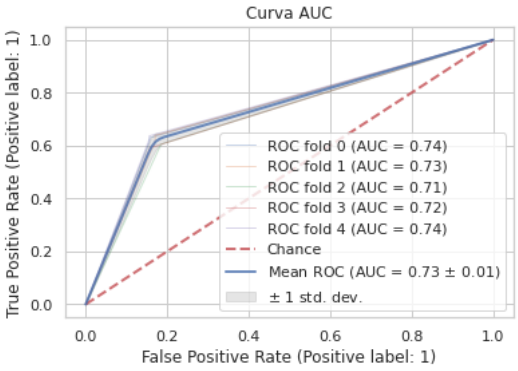
Fonte: autor

Figura 45 - Curva AUC Cenário CNBRecogna50W



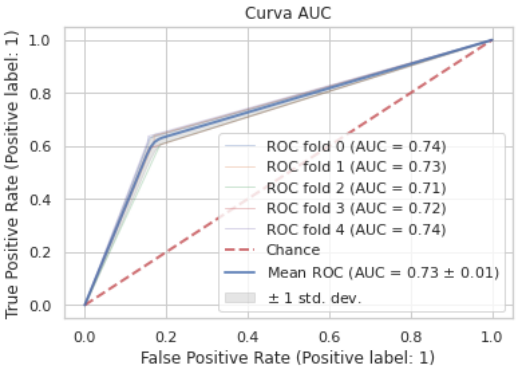
Fonte: autor

Figura 46 - Curva AUC Cenário CNBRecogna50W0d



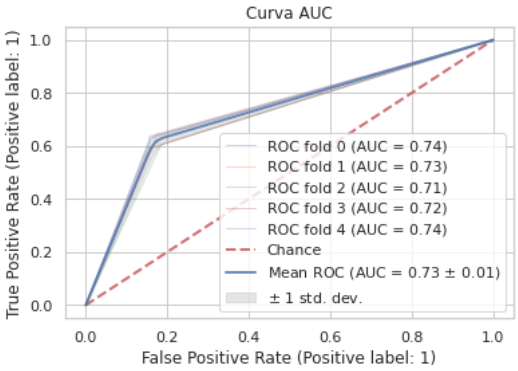
Fonte: autor

Figura 47 - Curva AUC Cenário CNBRecogna50W30d



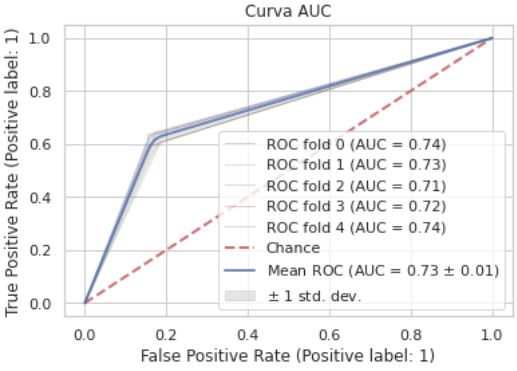
Fonte: autor

Figura 48 - Curva AUC Cenário CNBRecogna50W60d



Fonte: autor

Figura 49 - Curva AUC Cenário CNBRecogna50W90d



Fonte: autor

5.6 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKERECOGOGNA50W

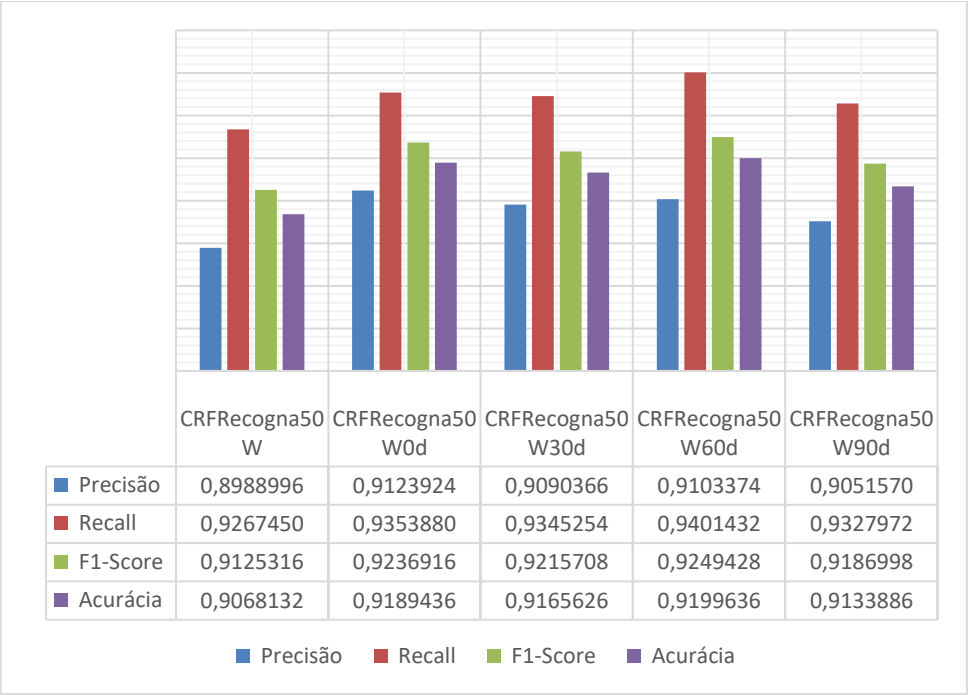
Na Figura 50 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo RF** e a **base FakeRecogna** modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 5).

Vale já ressaltar que é possível observar na Figura 36 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos principalmente entre 0 e 30 dias.

Além disso, nas Figuras 51, 52, 53, 54 e 55 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 50, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

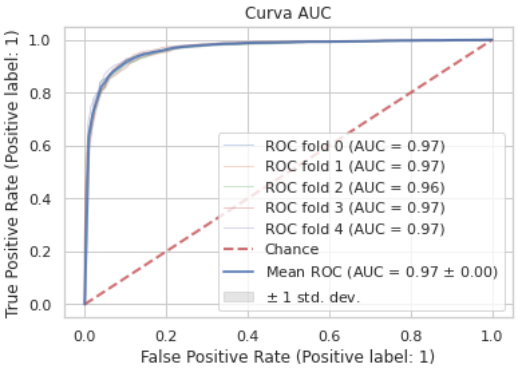
Figura 50 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFRecogna sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CRFRecogna50W	CRFRecogna50W0d	CRFRecogna50W30d	CRFRecogna50W60d	CRFRecogna50W90d
Precisão	0,8988996	0,9123924	0,9090366	0,9103374	0,9051570
Recall	0,9267450	0,9353880	0,9345254	0,9401432	0,9327972
F1-Score	0,9125316	0,9236916	0,9215708	0,9249428	0,9186998
Acurácia	0,9068132	0,9189436	0,9165626	0,9199636	0,9133886

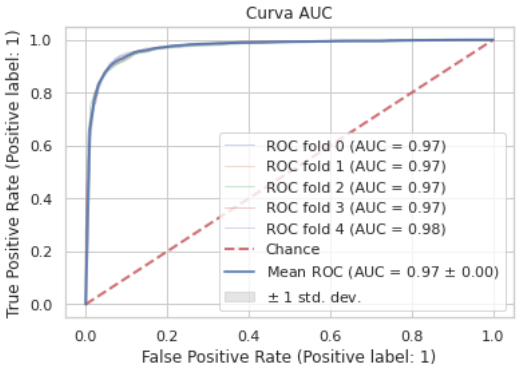
Fonte: autor

Figura 51 - Curva AUC Cenário CRFRecogna50W



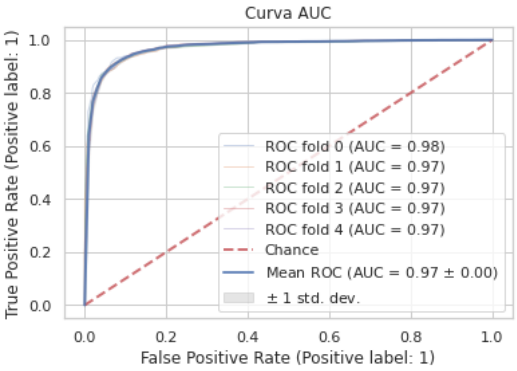
Fonte: autor

Figura 52 - Curva AUC Cenário CRFRecogna50W0d



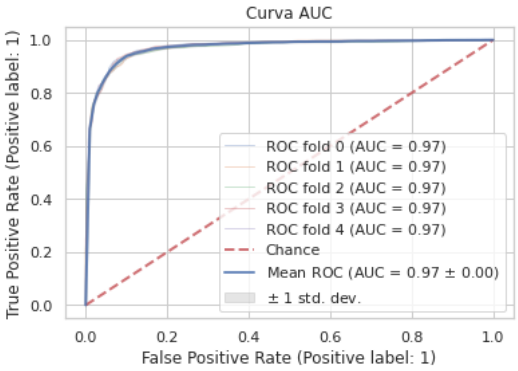
Fonte: autor

Figura 53 - Curva AUC Cenário CRFRecogna50W30d



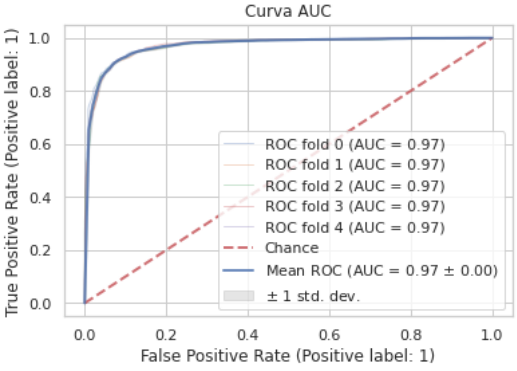
Fonte: autor

Figura 54 - Curva AUC Cenário CRFRecogna50W60d



Fonte: autor

Figura 55 - Curva AUC Cenário CRFRecogna50W90d



Fonte: autor

5.7 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKEBR

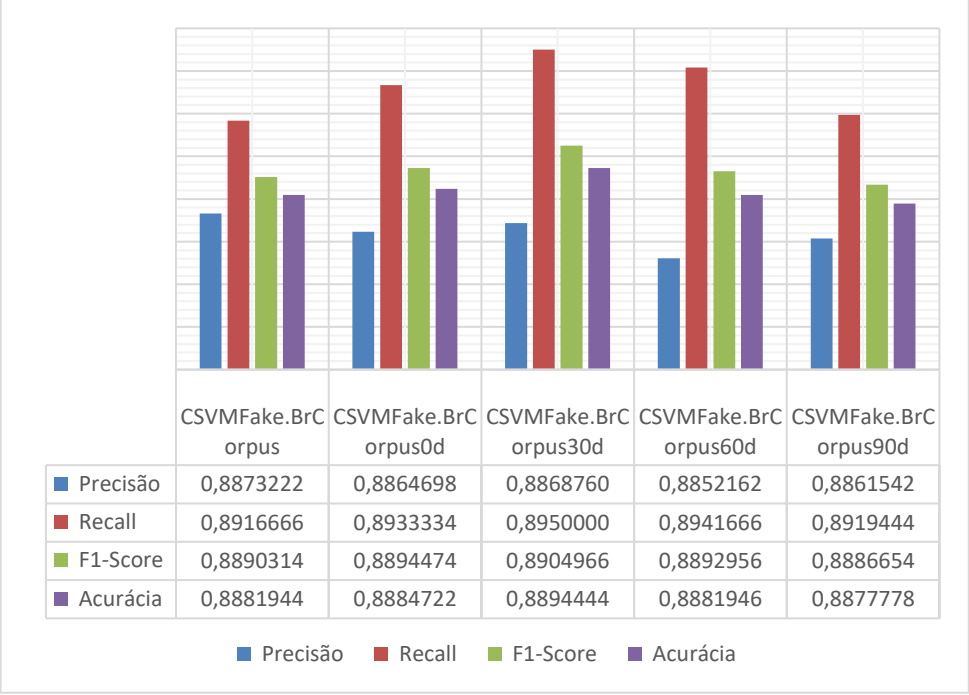
Na Figura 56 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo SVM** e a **base Fake.Br** sem e com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 6).

Vale já ressaltar que é possível observar na Figura 54 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos principalmente entre 30 e 60 dias.

Além disso, nas Figuras 57, 58, 59, 60 e 61 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

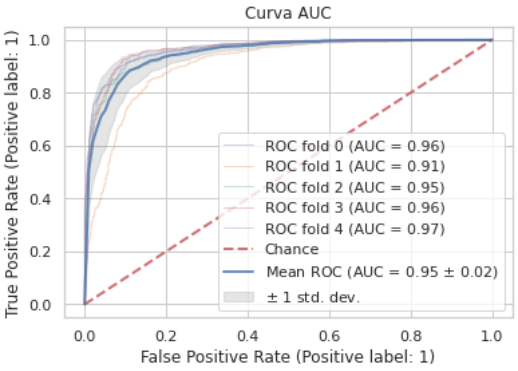
Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 56, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

Figura 56 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias



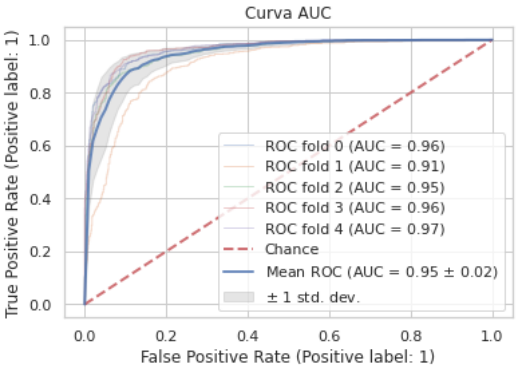
Fonte: autor

Figura 57 - Curva AUC Cenário CSVMFake.BrCorpus



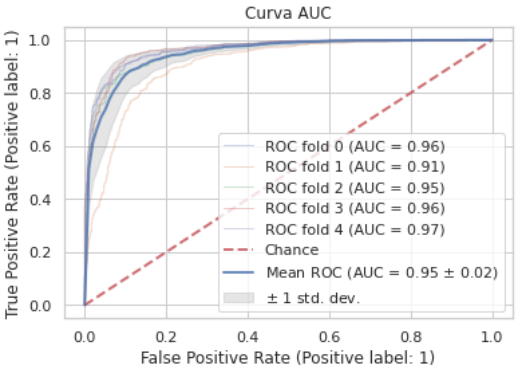
Fonte: autor

Figura 58 - Curva AUC Cenário CSVMFake.BrCorpus0d



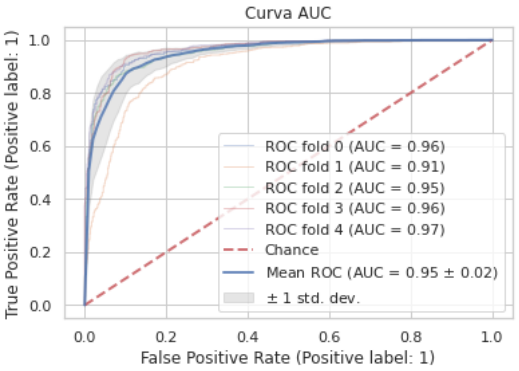
Fonte: autor

Figura 59 - Curva AUC Cenário CSVMFake.BrCorpus30d



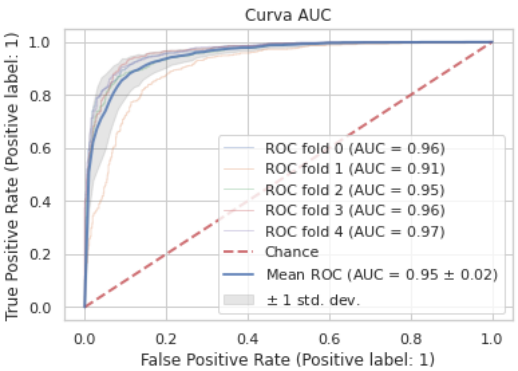
Fonte: autor

Figura 60 - Curva AUC Cenário CSVMFake.BrCorpus60d



Fonte: autor

Figura 61 - Curva AUC Cenário CSVMFake.BrCorpus90d



Fonte: autor

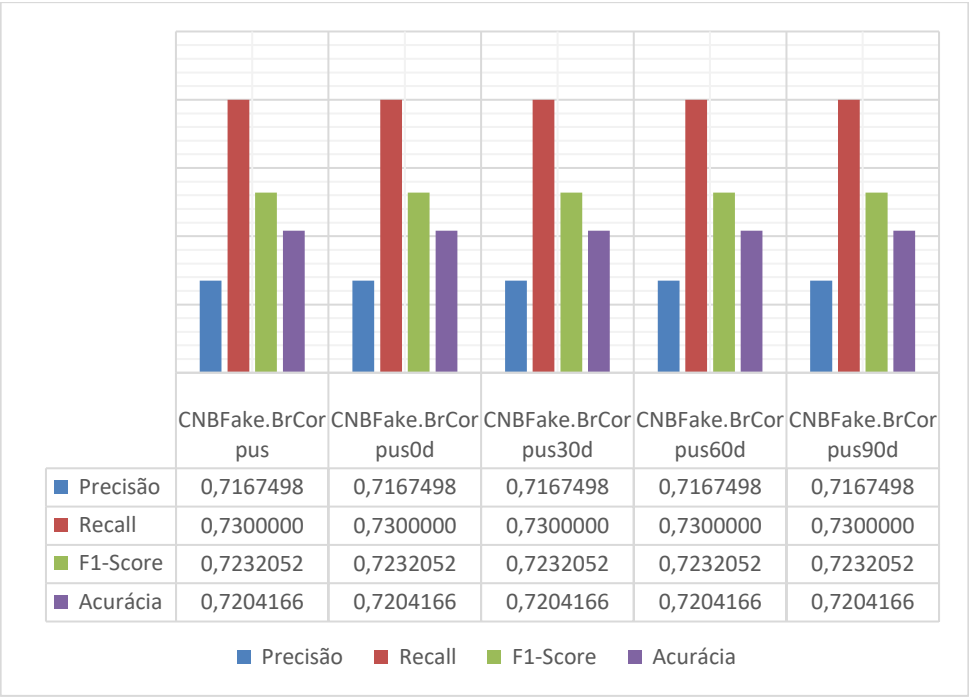
5.8 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKEBR

Na Figura 62 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo NB** e a **base Fake.Br** sem e com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 6).

Para o modelo NB, diferente do que ocorre com os modelos SVM e RF, não se nota qualquer alteração nos cenários propostos, algo que se reflete nas curvas AUC indicadas nas Figuras 63, 64, 65, 66 e 67.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 62, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

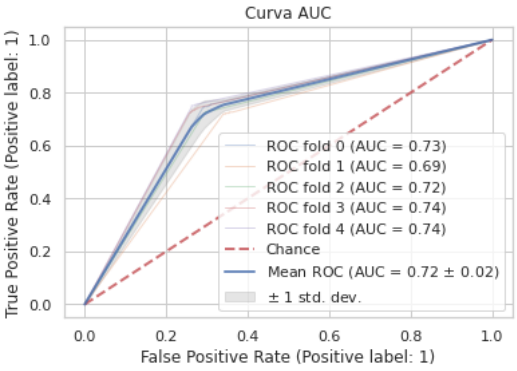
Figura 62 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CNBFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CNBFake.BrCorpus	CNBFake.BrCorpus0d	CNBFake.BrCorpus30d	CNBFake.BrCorpus60d	CNBFake.BrCorpus90d
Precisão	0,7167498	0,7167498	0,7167498	0,7167498	0,7167498
Recall	0,7300000	0,7300000	0,7300000	0,7300000	0,7300000
F1-Score	0,7232052	0,7232052	0,7232052	0,7232052	0,7232052
Acurácia	0,7204166	0,7204166	0,7204166	0,7204166	0,7204166

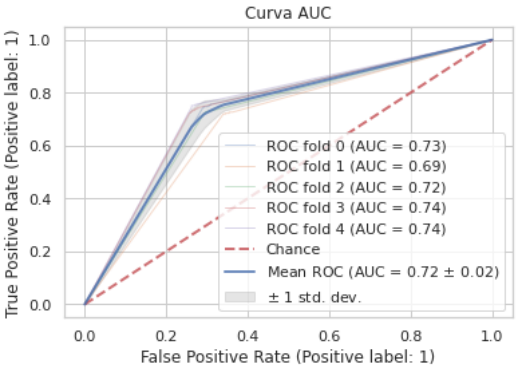
Fonte: autor

Figura 63 - Curva AUC Cenário CNBFake.BrCorpus



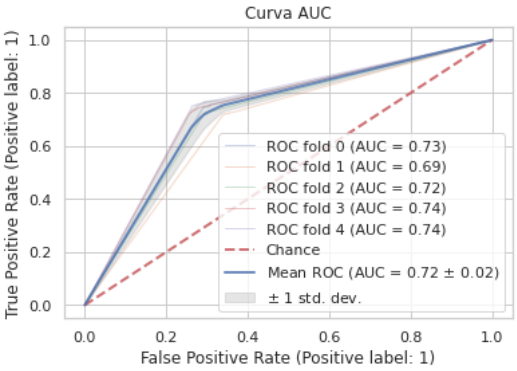
Fonte: autor

Figura 64 - Curva AUC Cenário CNBFake.BrCorpus0d



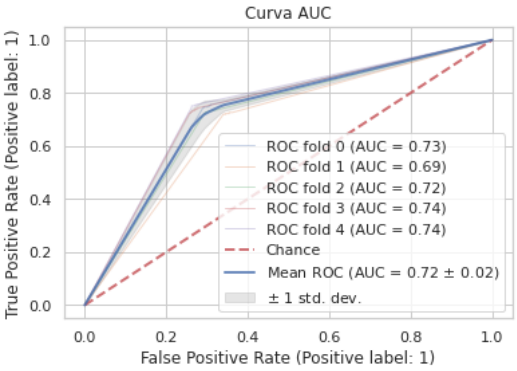
Fonte: autor

Figura 65 - Curva AUC Cenário CNBFake.BrCorpus30d



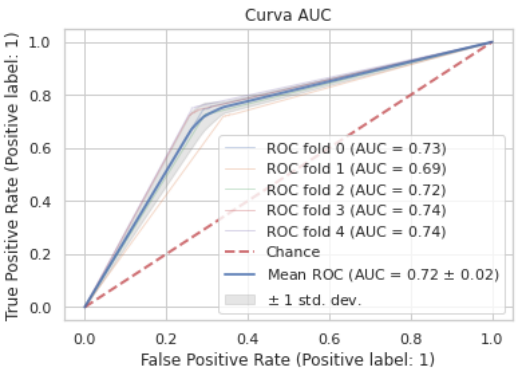
Fonte: autor

Figura 66 - Curva AUC Cenário CNBFake.BrCorpus60d



Fonte: autor

Figura 67 - Curva AUC Cenário CNBFake.BrCorpus90d



Fonte: autor

5.9 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKEBR

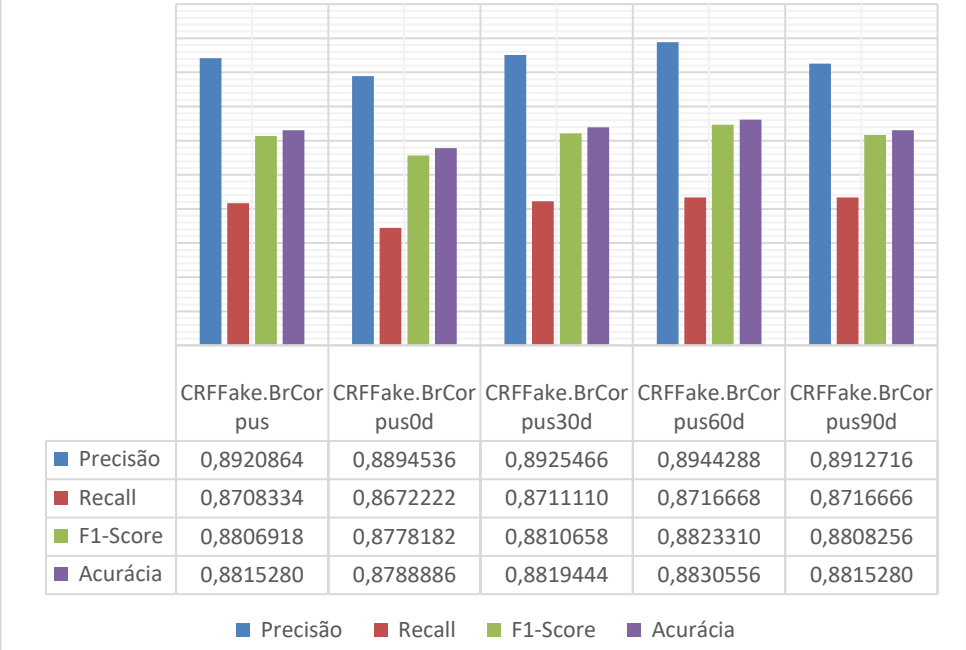
Na Figura 68 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo RF** e a **base Fake.Br** sem e com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 6).

Vale já ressaltar que é possível observar na Figura 54 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos principalmente entre 30 e 60 dias.

Além disso, nas Figuras 69, 70, 71, 72 e 73 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 68, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

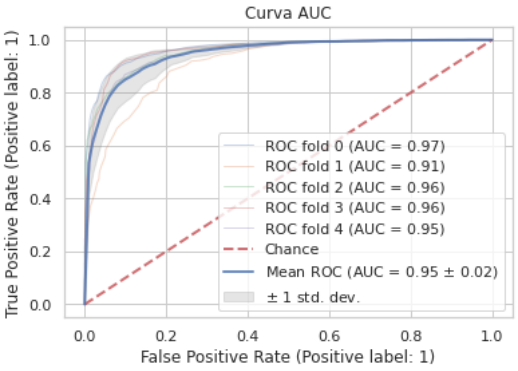
Figura 68 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFFake.BrCorpus sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CRFFake.BrCorpus	CRFFake.BrCorpus0d	CRFFake.BrCorpus30d	CRFFake.BrCorpus60d	CRFFake.BrCorpus90d
Precisão	0,8920864	0,8894536	0,8925466	0,8944288	0,8912716
Recall	0,8708334	0,8672222	0,8711110	0,8716668	0,8716666
F1-Score	0,8806918	0,8778182	0,8810658	0,8823310	0,8808256
Acurácia	0,8815280	0,8788886	0,8819444	0,8830556	0,8815280

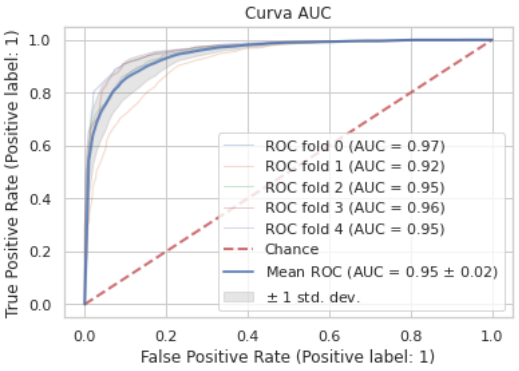
Fonte: autor

Figura 69 - Curva AUC Cenário CRFFake.BrCorpus



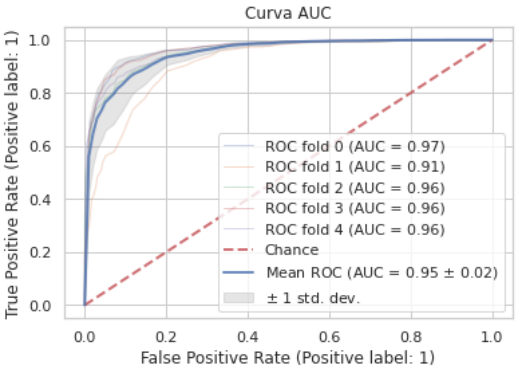
Fonte: autor

Figura 70 - Curva AUC Cenário CRFFake.BrCorpus0d



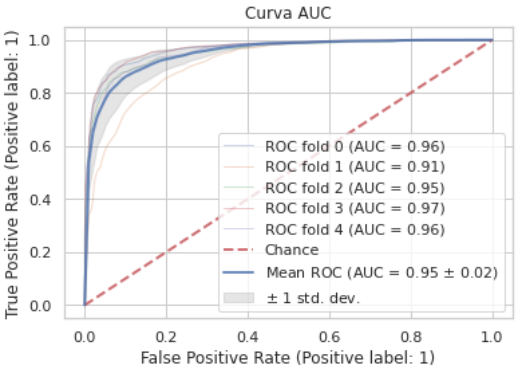
Fonte: autor

Figura 71 - Curva AUC Cenário CRFFake.BrCorpus30d



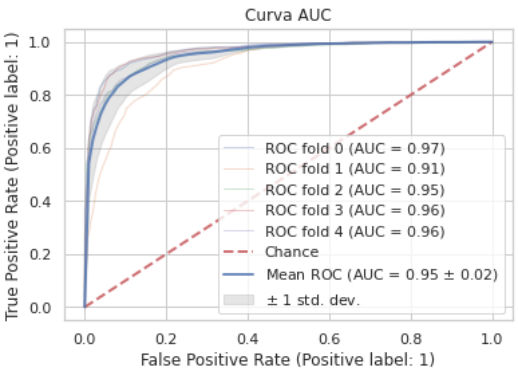
Fonte: autor

Figura 72 - Curva AUC Cenário CRFFake.BrCorpus60d



Fonte: autor

Figura 73 - Curva AUC Cenário CRFFake.BrCorpus90d



Fonte: autor

5.10 RESULTADOS PARA OS CENÁRIOS COM SVM E A BASE FAKEBR50W

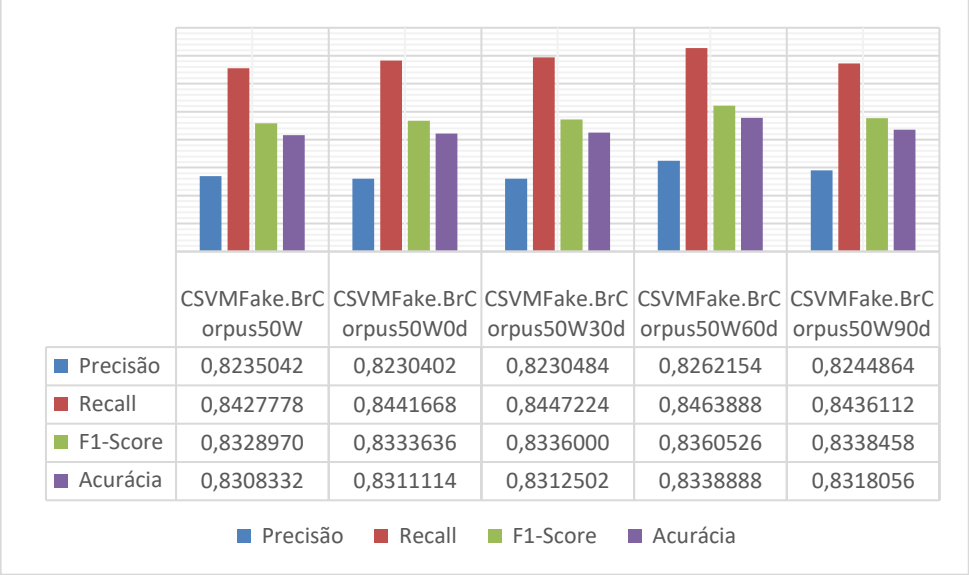
Na Figura 74 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo SVM** e a **base Fake.Br** modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 7).

Vale já ressaltar que é possível observar na Figura 72 uma melhoria visível em todos as métricas para os cenários indicadores sociais e macroeconômicos com deslocamentos entre 60 e 90 dias.

Além disso, nas Figuras 75, 76, 77, 78 e 79 observam-se as curvas AUC para os cenários indicados. Quanto ao indicado nas curvas AUC, não se nota uma melhoria visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 74, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

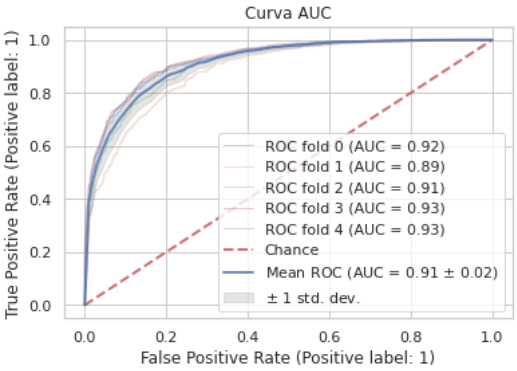
Figura 74 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CSVMFake.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CSVMFake.BrCorpus50W	CSVMFake.BrCorpus50W0d	CSVMFake.BrCorpus50W30d	CSVMFake.BrCorpus50W60d	CSVMFake.BrCorpus50W90d
Precisão	0,8235042	0,8230402	0,8230484	0,8262154	0,8244864
Recall	0,8427778	0,8441668	0,8447224	0,8463888	0,8436112
F1-Score	0,8328970	0,8333636	0,8336000	0,8360526	0,8338458
Acurácia	0,8308332	0,8311114	0,8312502	0,8338888	0,8318056

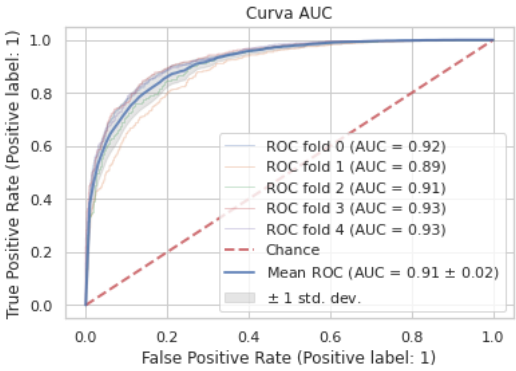
Fonte: autor

Figura 75 - Curva AUC Cenário CSVMFake.BrCorpus50W



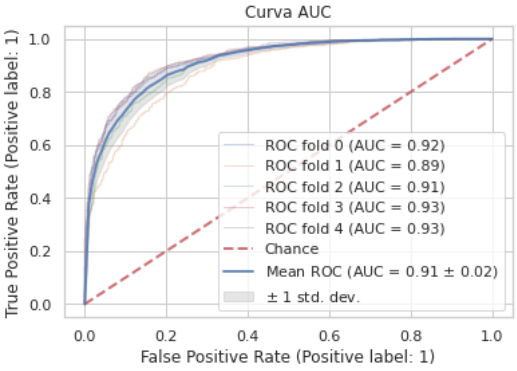
Fonte: autor

Figura 76 - Curva AUC Cenário CSVMFake.BrCorpus50W0d



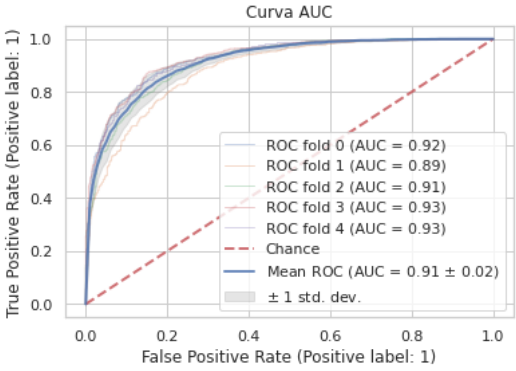
Fonte: autor

Figura 77 - Curva AUC Cenário CSVMFake.BrCorpus50W30d



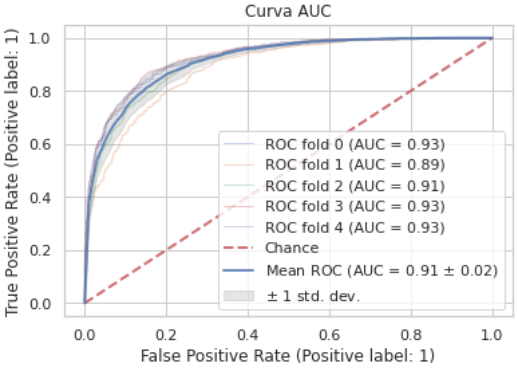
Fonte: autor

Figura 78 - Curva AUC Cenário CSVMFake.BrCorpus50W60d



Fonte: autor

Figura 79 - Curva AUC Cenário CSVMFake.BrCorpus50W90d



Fonte: autor

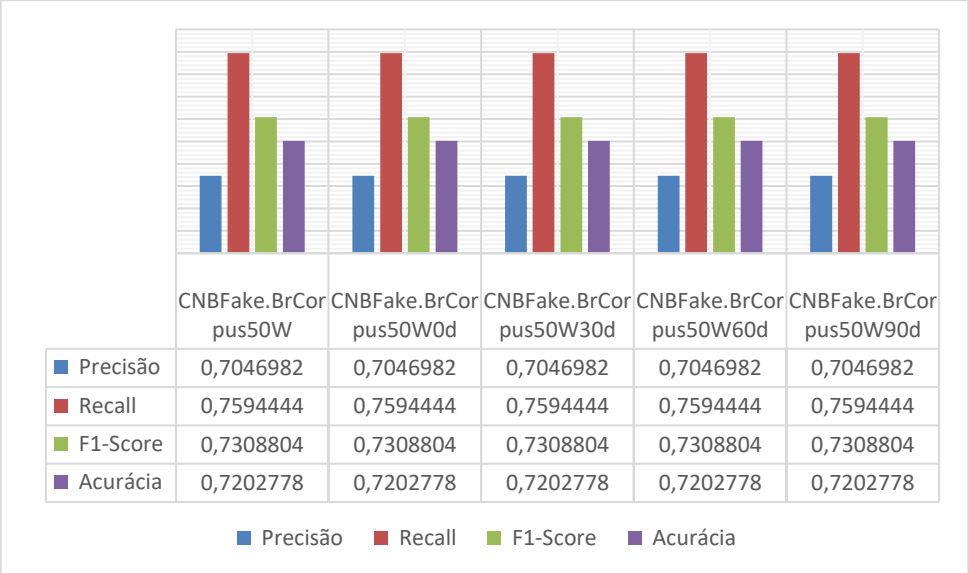
5.11 RESULTADOS PARA OS CENÁRIOS COM NB E A BASE FAKEBR50W

Na Figura 80 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo NB**, a **base Fake.Br** modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 7).

Para o modelo NB, diferente do que ocorre com os modelos SVM e RF, não se nota qualquer alteração nos cenários propostos, algo que se reflete nas curvas AUC indicadas nas Figuras 81, 82, 83, 84 e 85.

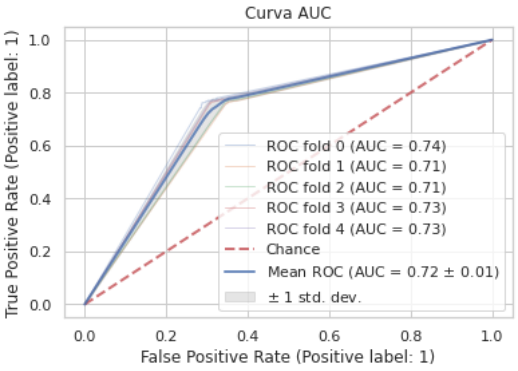
Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 80, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

Figura 80 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CNBFake.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias



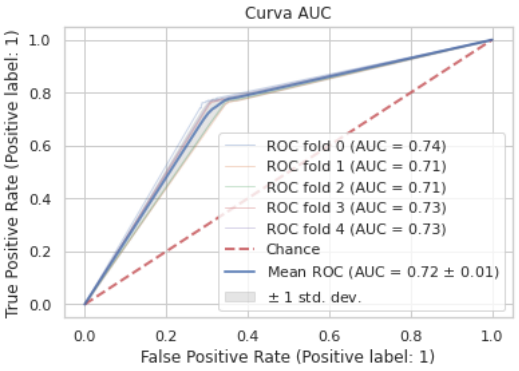
Fonte: autor

Figura 81 - Curva AUC Cenário CNBFake.BrCorpus50W



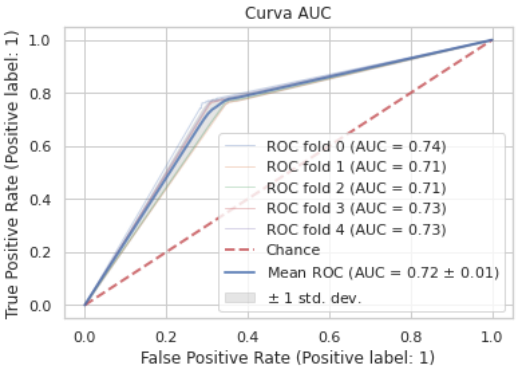
Fonte: autor

Figura 82 - Curva AUC Cenário CNBFake.BrCorpus50W0d



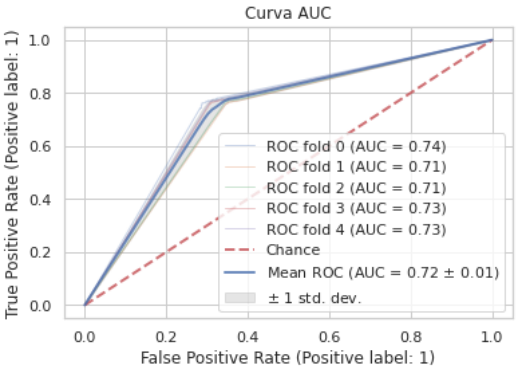
Fonte: autor

Figura 83 - Curva AUC Cenário CNBFake.BrCorpus50W30d



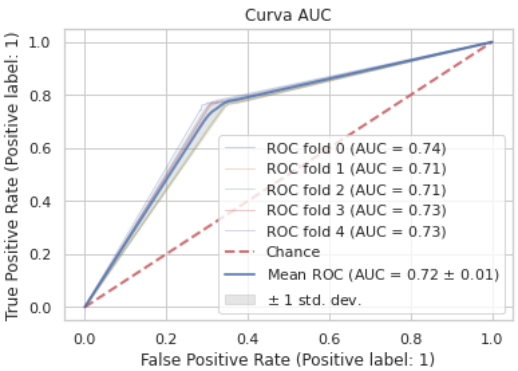
Fonte: autor

Figura 84 - Curva AUC Cenário CNBFake.BrCorpus50W60d



Fonte: autor

Figura 85 - Curva AUC Cenário CNBFake.BrCorpus50W90d



Fonte: autor

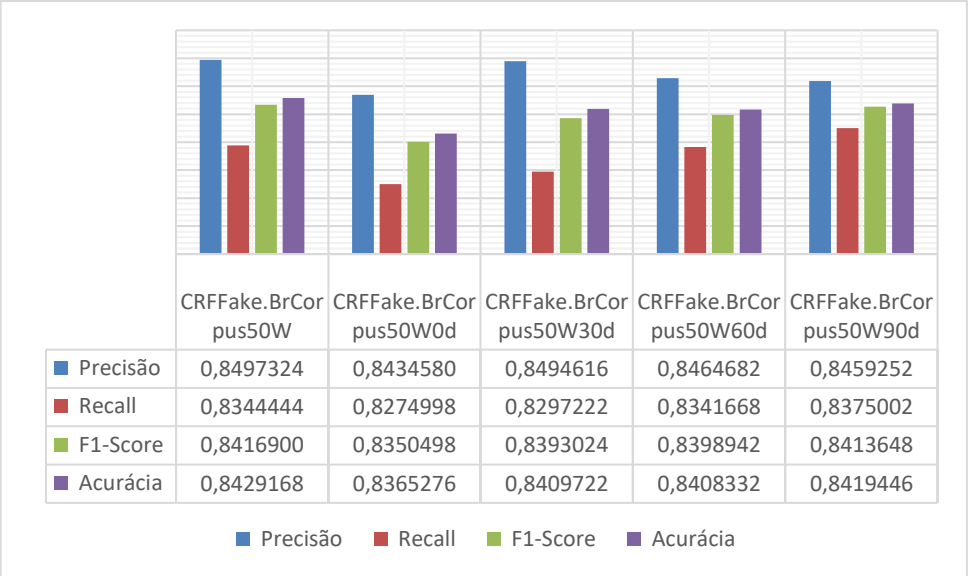
5.12 RESULTADOS PARA OS CENÁRIOS COM RF E A BASE FAKEBR50W

Na Figura 86 tem-se os resultados de Precisão, *Recall*, *F1-Score* e Acurácia para os cenários com o **modelo RF** e a **base Fake.Br** modificada. Essa modificação **considera apenas as cinquenta primeiras palavras de cada notícia** da base. Além disso, há testes com e sem indicadores sociais e macroeconômicos, sendo a base com indicadores sociais com deslocamentos de 0 dias, 30 dias, 60 dias e 90 dias (conforme Tabela 7).

Vale já ressaltar que é possível observar na Figura 84 uma piora visível em todos as métricas para os cenários indicadores sociais e macroeconômicos, sendo esses os únicos cenários nos quais esse evento é observado. Mesmo embora isso ocorra, observam-se nas Figuras 87, 88, 89, 90 e 91, onde tem-se as curvas AUC, que não há nota uma piora visual chamativa entre as taxas de classificação entre as taxas de falsos positivos e verdadeiros positivos.

Cada cenário de testes resultou em cinco grupo de colunas de métricas, conforme a Figura 86, e cinco grupos de imagens das curvas AUC, que consideram as médias dos resultados obtidos no *k-fold cross validation* (k=5).

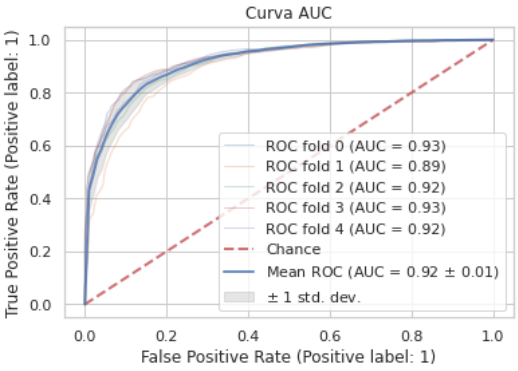
Figura 86 - Precisão, *Recall*, *F1-Score* e Acurácia para os Cenários CRFke.BrCorpus50W sem índices e com índices deslocados em 0, 30, 60 e 90 dias



	CRFFake.BrCorpus50W	CRFFake.BrCorpus50W0d	CRFFake.BrCorpus50W30d	CRFFake.BrCorpus50W60d	CRFFake.BrCorpus50W90d
Precisão	0,8497324	0,8434580	0,8494616	0,8464682	0,8459252
Recall	0,8344444	0,8274998	0,8297222	0,8341668	0,8375002
F1-Score	0,8416900	0,8350498	0,8393024	0,8398942	0,8413648
Acurácia	0,8429168	0,8365276	0,8409722	0,8408332	0,8419446

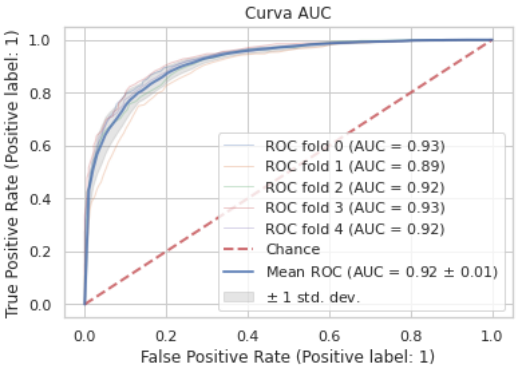
Fonte: autor

Figura 87 - Curva AUC Cenário CSVMFake.BrCorpus50W



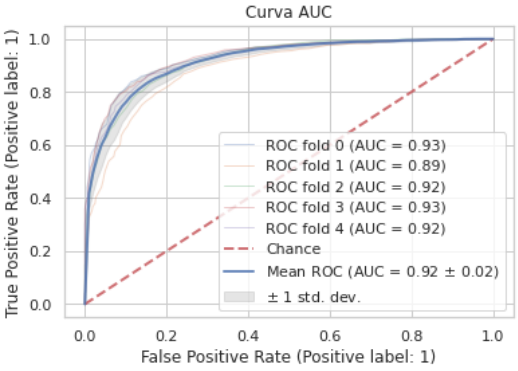
Fonte: autor

Figura 88 - Curva AUC Cenário CSVMFake.BrCorpus50W0d



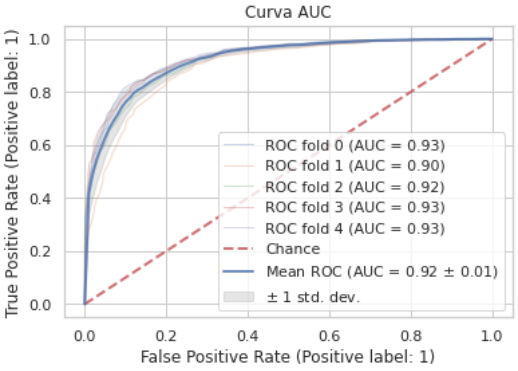
Fonte: autor

Figura 89 - Curva AUC Cenário CSVMFake.BrCorpus50W30d



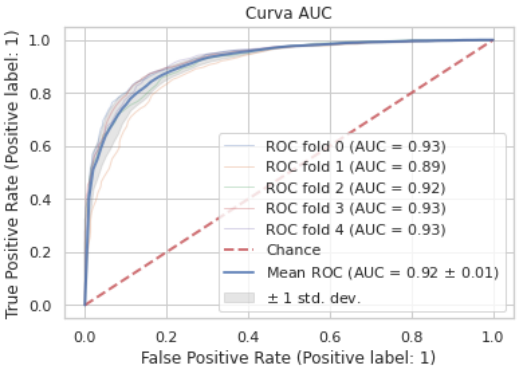
Fonte: autor

Figura 90 - Curva AUC Cenário CSVMFake.BrCorpus50W60d



Fonte: autor

Figura 91 - Curva AUC Cenário CSVMFake.BrCorpus50W90d

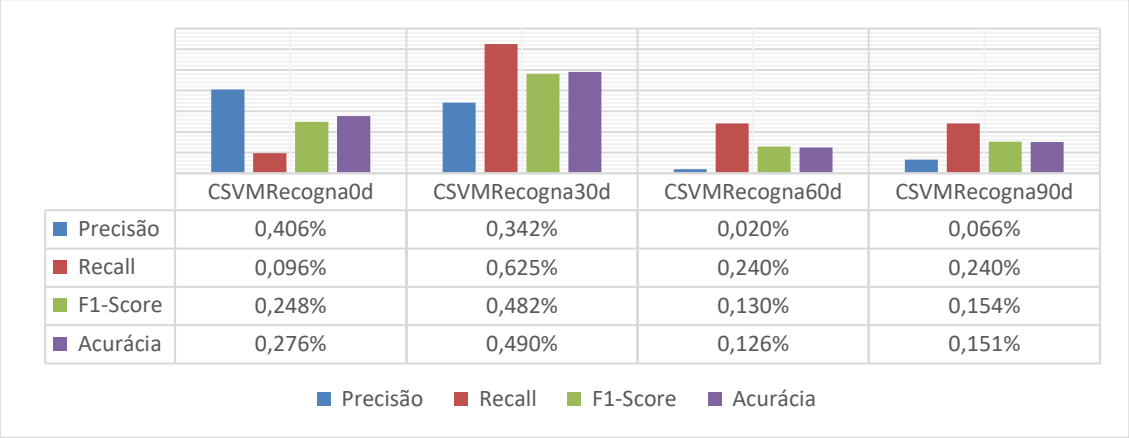


Fonte: autor

5.13 ANÁLISE GERAL E COMPARATIVA

Abaixo seguem imagens que indicam o aumento percentual incremental ao serem inseridos os índices sociais e macroeconômicos.

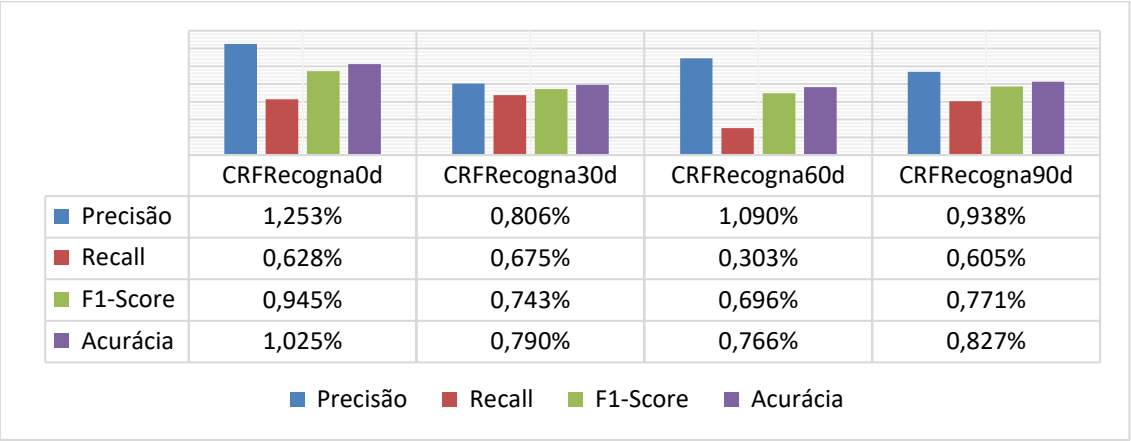
Figura 92 – Aumento incremental com o SVM e base Recogna



Fonte: autor

Na Figura 92 tem-se que, para o SVM com a base Recogna, o melhor desempenho (entre 0.3 e 0.6 pontos percentuais para todas as métricas) encontra-se entre 0 e 30 dias, embora já exista uma melhoria com outros períodos de deslocamento. Os melhores resultados com os deslocamentos em 30 dias se devem ao fato de que o efeito da alteração do humor enxerga-se após um tempo entre o espalhamento das notícias, daí a alteração nos índices se dá com um determinado atraso. Nota-se também que houve um aumento mais significante no poder da medida de *recall*, que indica uma boa efetividade para indicar-se o que foi classificado como *fake news* corretamente dentro de tudo que de fato é uma *fake news*. Essa conclusão demonstra que o os índices contribuem positivamente para a classificação de uma notícia de fato falsa, já a acurácia indica que houve uma melhora na classificação geral. Vale ressaltar que, dada a sensibilidade do tema, pode se ter que é mais positivo termos incremento em taxas de que indicam uma notícia que de falto é falsa, reduzindo-se assim as taxas de indicação de notícias verdadeiras como notícias falsas.

Figura 93 - Aumento incremental com o RF e base Recogna

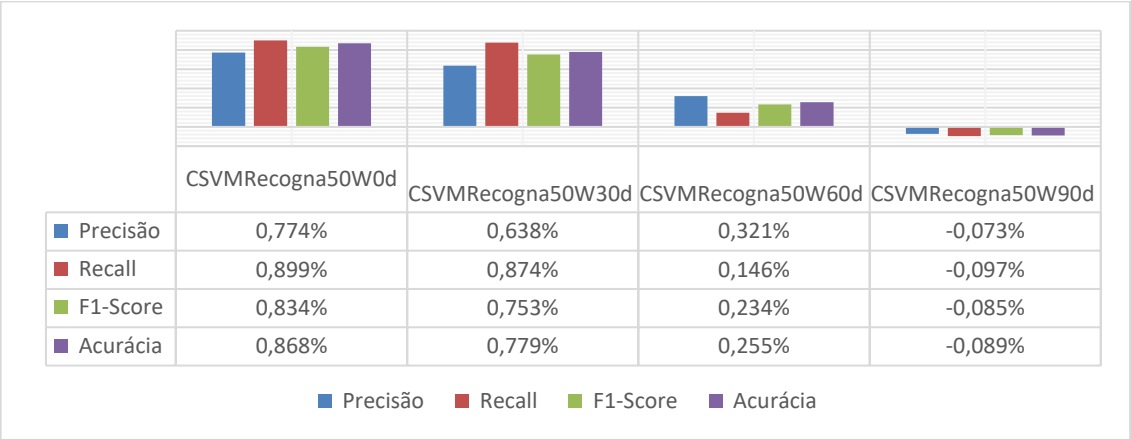


	CRFRecogna0d	CRFRecogna30d	CRFRecogna60d	CRFRecogna90d
Precisão	1,253%	0,806%	1,090%	0,938%
Recall	0,628%	0,675%	0,303%	0,605%
F1-Score	0,945%	0,743%	0,696%	0,771%
Acurácia	1,025%	0,790%	0,766%	0,827%

Fonte: autor

Na Figura 93 tem-se que, para o RF com a base Recogna, o melhor desempenho (entre 0.6 e 1.2 pontos percentuais para todas as métricas) encontra-se entre 0 e 30 dias, embora já exista uma melhoria com outros períodos de deslocamento. O desempenho com o RF segue superior ao encontrado com o SVM. Os melhores resultados com os deslocamentos em 0 e 30 dias se devem ao fato de que o efeito da alteração do humor enxerga-se após um tempo entre o espalhamento das notícias, daí a alteração nos índices se dá com um determinado atraso. Encontra-se também uma melhora entre outros períodos de deslocamento, indicando que o efeito positivo pode se dar, não necessariamente pelo valor do índice em si, mas pelo efeito de sua alteração. Nota-se também que houve um aumento mais significante no poder da medida de *precisão*, diferentemente do que ocorre no teste anterior, com o SVM, porém o RF indica melhoras mais significativas. O RF trabalha melhor com um conjunto de características quantitativas, já o SVM lida melhor com uma mistura no conjunto de caraterísticas, podendo lidar com características categóricas e quantitativas. Daí, entende-se o fato do RF ter desempenho superior no conjunto de dados considerado.

Figura 94 - Aumento incremental com o SVM e base Recogna modificada

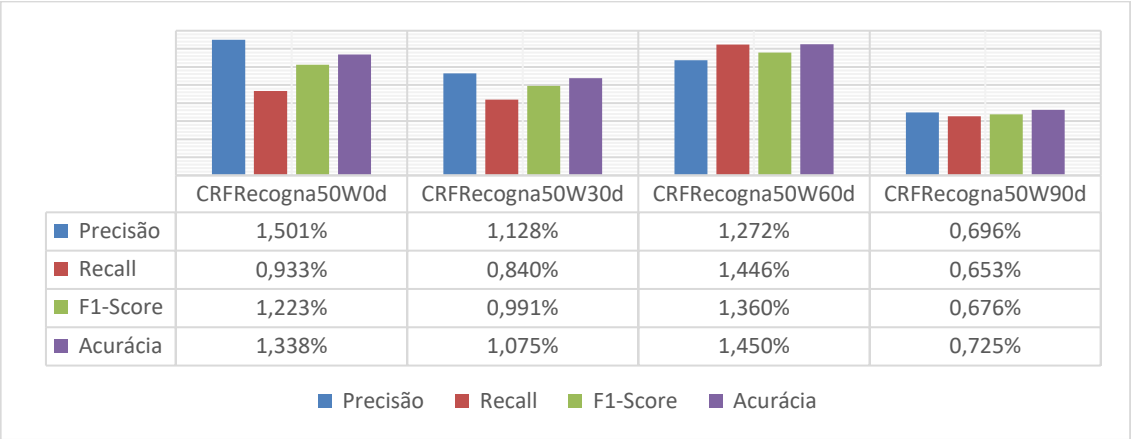


	CSVMRecogna50W0d	CSVMRecogna50W30d	CSVMRecogna50W60d	CSVMRecogna50W90d
Precisão	0,774%	0,638%	0,321%	-0,073%
Recall	0,899%	0,874%	0,146%	-0,097%
F1-Score	0,834%	0,753%	0,234%	-0,085%
Acurácia	0,868%	0,779%	0,255%	-0,089%

Fonte: autor

Na Figura 94 tem-se que, para o SVM com a base Recogna modificada (apenas 50 primeiras palavras para cada notícia), o melhor desempenho (entre 0.6 e 0.9 pontos percentuais para todas as métricas) encontra-se entre 0 e 30 dias, embora já exista uma melhoria com outros períodos de deslocamento. Nota-se que com a base modificada o desempenho do SVM é superior. Os melhores resultados com os deslocamentos em 0 e 30 dias se devem ao fato de que o efeito da alteração do humor enxerga-se após um tempo entre o espalhamento das notícias, daí a alteração nos índices se dá com um determinado atraso. Nota-se também que houve um aumento mais significante no poder da medida de *recall*, diferentemente do que ocorre no teste anterior, com o RF. Como já mencionado, o RF trabalha melhor com um conjunto de características quantitativas, já o SVM lida melhor com uma mistura no conjunto de caraterísticas, podendo lidar com características categóricas e quantitativas. Daí, entende-se o fato do SVM ter desempenho inferior no conjunto de dados considerado. Embora isso seja verdade, nota-se que com a alteração da base de dados, com a redução da quantidade de palavras consideradas, o incremento no desempenho é superior. Isso se deve, pois, a limitação do tamanho do vetor de caraterísticas linguísticas reduz a importância do valor do conteúdo textual na tarefa de classificação, e aumenta a importância das características dos índices.

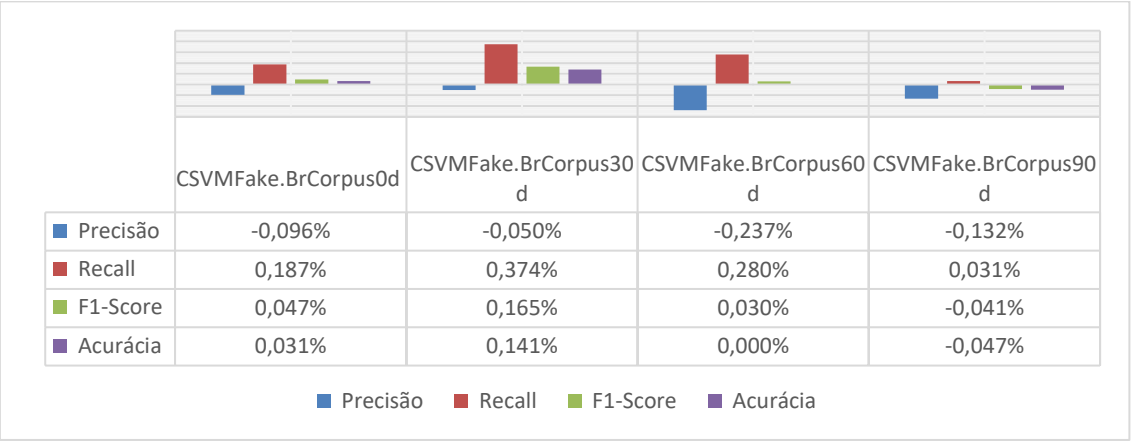
Figura 95 - Aumento incremental com o RF e base Recogna modificada



Fonte: autor

Na Figura 95 tem-se que, para o RF com a base Recogna modificada (apenas 50 primeiras palavras para cada notícia), o melhor desempenho (entre 0.9 e 1.5 pp) encontra-se entre 0 e 60 dias, embora já exista uma melhoria com outros períodos de deslocamento. Assim como no SVM, nota-se que com a base modificada o desempenho é superior. Os melhores resultados com os deslocamentos em 30 e 60 dias se devem ao fato de que o efeito da alteração do humor enxerga-se após um tempo entre o espalhamento das notícias, daí a alteração nos índices se dá com um determinado atraso. Novamente, há indícios de que a alteração dos índices em si, e não os valores, podem indicar um melhor valor de discriminação, já que é melhor também em outros tempos de deslocamentos. Como já mencionado, o RF trabalha melhor com um conjunto de características quantitativas, já o SVM lida melhor com uma mistura no conjunto de caraterísticas, podendo lidar com características categóricas e quantitativas. Daí, entende-se o fato do RF ter desempenho superior no conjunto de dados considerado. Adicionalmente, com a redução da quantidade de palavras, nota-se de maneira mais evidente o melhor desempenho do modelo RF. Como já mencionado, isso se deve, pois, a limitação do tamanho do vetor de caraterísticas linguísticas reduz a importância do valor do conteúdo textual na tarefa de classificação, e aumenta a importância das características dos índices.

Figura 96 - Aumento incremental com o SVM e base Fake.Br

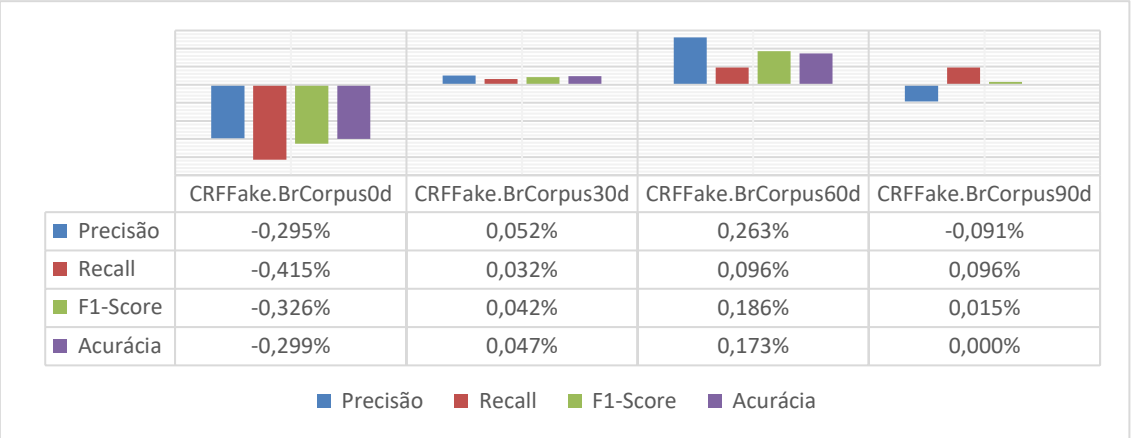


	CSVMFake.BrCorpus0d	CSVMFake.BrCorpus30d	CSVMFake.BrCorpus60d	CSVMFake.BrCorpus90d
Precisão	-0,096%	-0,050%	-0,237%	-0,132%
Recall	0,187%	0,374%	0,280%	0,031%
F1-Score	0,047%	0,165%	0,030%	-0,041%
Acurácia	0,031%	0,141%	0,000%	-0,047%

Fonte: autor

Na Figura 96 tem-se que, para o SVM com a base Fake.Br, os desempenhos não se mantêm positivos como para a base Fake Recogna, isso se deve pela diferença existente entre as bases, Fake.Br e Fake Recogna, embora tenhamos um desempenho positivo para 60 dias. Daqui, entende-se que um estudo das características linguísticas das bases consideradas pode beneficiar o estudo, entendendo-se melhor quando há uma possível aplicação para índices adicionais no modelo classificatório.

Figura 97 - Aumento incremental com o RF e base Fake.Br



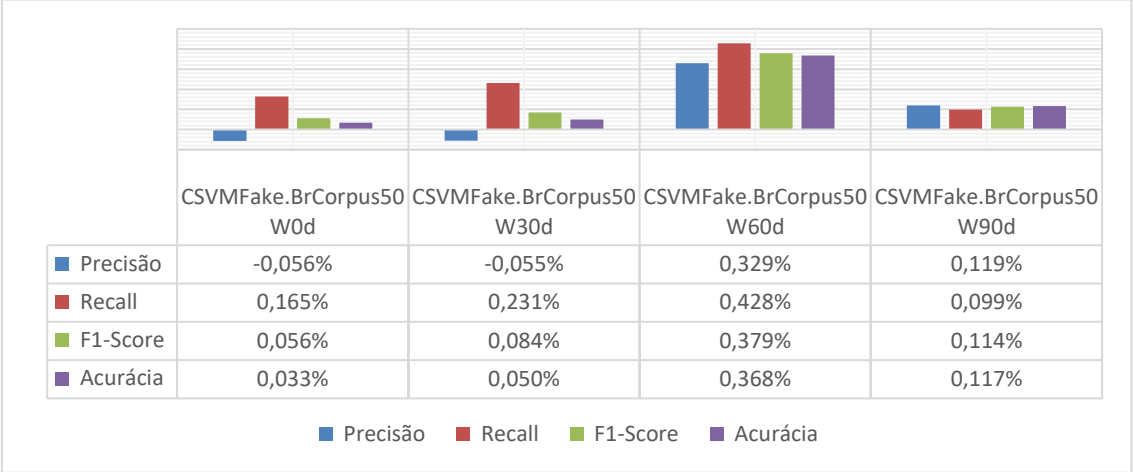
	CRFFake.BrCorpus0d	CRFFake.BrCorpus30d	CRFFake.BrCorpus60d	CRFFake.BrCorpus90d
Precisão	-0,295%	0,052%	0,263%	-0,091%
Recall	-0,415%	0,032%	0,096%	0,096%
F1-Score	-0,326%	0,042%	0,186%	0,015%
Acurácia	-0,299%	0,047%	0,173%	0,000%

Fonte: autor

Na Figura 97 tem-se que, para o RF com a base Fake.Br, os desempenhos não se mantêm positivos como para a base Fake Recogna, isso se deve pela diferença existente entre as bases, Fake.Br e Fake Recogna, embora tenhamos um desempenho positivo para 60 dias. Daqui, entende-se que um estudo das características linguísticas das bases consideradas pode

beneficiar o estudo, entendendo-se melhor quando há uma possível aplicação para índices adicionais no modelo classificatório.

Figura 98 - Aumento incremental com o SVM e base Fake.Br modificada

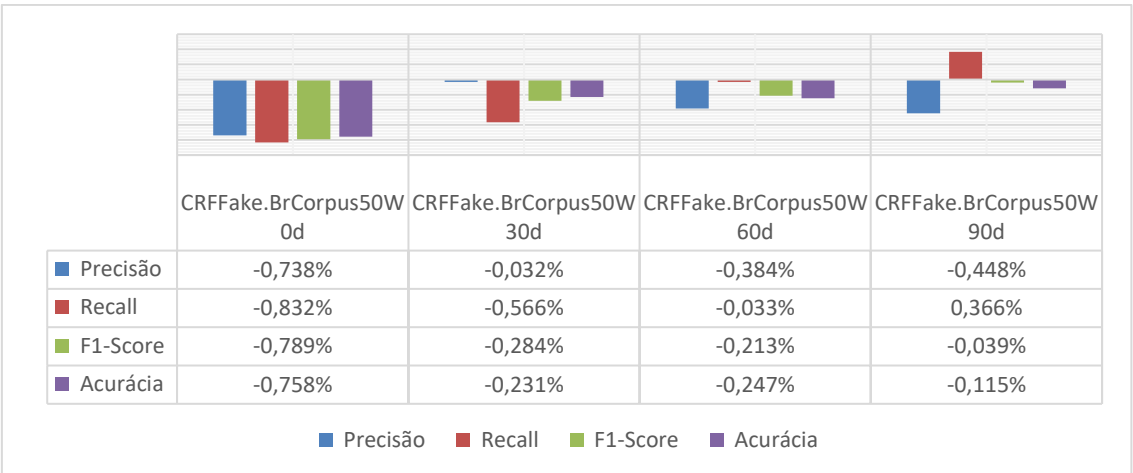


	CSVMFake.BrCorpus50W0d	CSVMFake.BrCorpus50W30d	CSVMFake.BrCorpus50W60d	CSVMFake.BrCorpus50W90d
Precisão	-0,056%	-0,055%	0,329%	0,119%
Recall	0,165%	0,231%	0,428%	0,099%
F1-Score	0,056%	0,084%	0,379%	0,114%
Acurácia	0,033%	0,050%	0,368%	0,117%

Fonte: autor

Na Figura 98 tem-se que, para o SVM com a base Fake.Br modificada (apenas 50 primeiras palavras para cada notícia), o melhor desempenho (entre 0.3 e 0.4 pp) encontra-se entre 60 e 90 dias. Assim como com a base Fake Recogna, nota-se que com a base modificada o desempenho é superior. Como já mencionado, isso deve pelo fato de que há uma maior importância dos índices quando há a redução das características linguísticas.

Figura 99 - Aumento incremental com o RF e base Fake.Br modificada

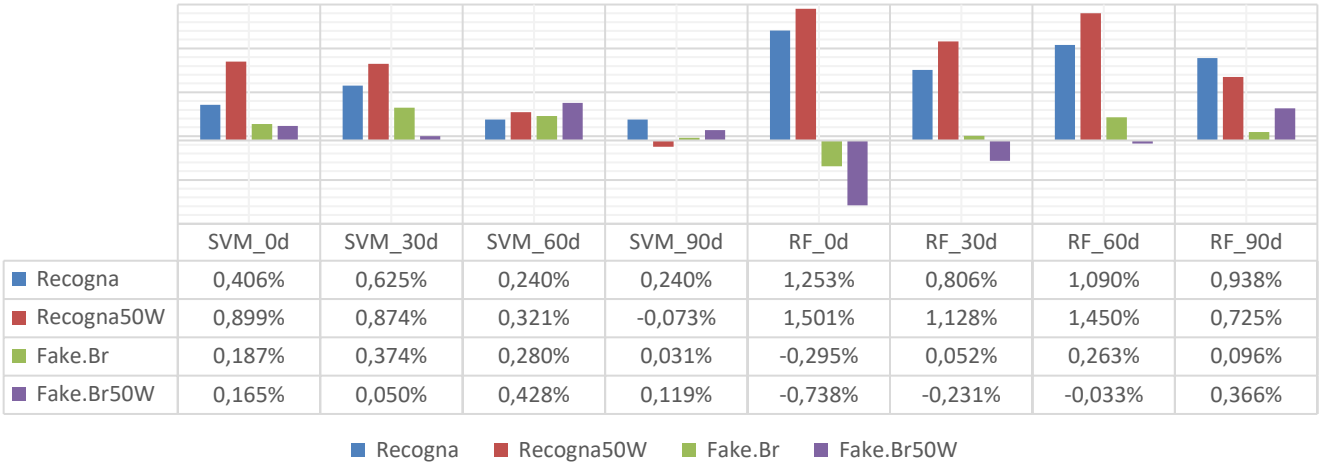


	CRFFake.BrCorpus50W0d	CRFFake.BrCorpus50W30d	CRFFake.BrCorpus50W60d	CRFFake.BrCorpus50W90d
Precisão	-0,738%	-0,032%	-0,384%	-0,448%
Recall	-0,832%	-0,566%	-0,033%	0,366%
F1-Score	-0,789%	-0,284%	-0,213%	-0,039%
Acurácia	-0,758%	-0,231%	-0,247%	-0,115%

Fonte: autor

Na Figura 99 tem-se que, para o RF com a base Fake.Br modificada (apenas 50 primeiras palavras para cada notícia), os desempenhos não se mantêm positivos como para a base Fake Recogna modificada (apenas 50 primeiras palavras para cada notícia), isso pode se dever a diferença existente entre as bases.

Figura 100 – Indicação da melhoria incremental ao se comparar os modelos sem e com os indicadores sociais e macroeconômicos



	SVM_0d	SVM_30d	SVM_60d	SVM_90d	RF_0d	RF_30d	RF_60d	RF_90d
Recogna	0,406%	0,625%	0,240%	0,240%	1,253%	0,806%	1,090%	0,938%
Recogna50W	0,899%	0,874%	0,321%	-0,073%	1,501%	1,128%	1,450%	0,725%
Fake.Br	0,187%	0,374%	0,280%	0,031%	-0,295%	0,052%	0,263%	0,096%
Fake.Br50W	0,165%	0,050%	0,428%	0,119%	-0,738%	-0,231%	-0,033%	0,366%

Fonte: autor

Analisando-se a Figura 100, na qual foi coletado o maior valor dentre as métricas consideradas (seja acurácia, *recall*, precisão ou *F1-score*), nota-se um incremento suave, porém constante quando se observam as métricas obtidas nos testes com a base **FakeRecogna** com os modelos *Random Forest* e SVM, demonstrando assim a possibilidade de melhoria na tarefa de classificação de *fake news* com esses modelos ao serem inseridos índices sociais e macroeconômicos como variáveis de entrada.

Para a base **FakeRecogna modificada**, na qual considera-se a utilização de apenas as cinquenta primeiras palavras de cada notícia, nota-se que há um incremento ainda superior em relação ao observado com a base de notícias completas ao serem inseridos os indicadores sociais e macroeconômicos. Daqui, pode-se concluir que a redução de características linguísticas que podem ser relevantes no modelo é compensada pela inclusão de outras variáveis, que no caso são os índices sociais e macroeconômicos. Portanto, a relevância das características linguísticas é reduzida.

Em ambas as bases (FakeRecogna e FakeRecogna modificada) nota-se que o modelo vencedor é o *Random Forest*, embora o SVM ainda demonstre melhoria. Como já mencionado, isso se deve pois o modelo *Random Forest* lidar melhor com características quantitativas, já o SVM lida bem com uma mistura de características quantitativas e categóricas. Vale lembrar também que os modelos SVM e RF lidam melhor com mais dimensões do que o modelo *Naive Bayes*, que tendeu a ignorar o acréscimo de variáveis, dado que já lidava com um número de variáveis consideráveis no modelo que utiliza apenas características textuais sem os indicadores sociais e macroeconômicos. Ressalta-se também que o modelo *Naive Bayes* é

paramétrico, ou seja, pressupõe-se um tipo de distribuição probabilística definida, que para o caso foi considerada a distribuição gaussiana. Já os modelos *Random Forest* e SVM são não paramétricos, portanto, os dados não necessitam ser provenientes de um tipo de distribuição de probabilidade.

Nota-se também que os indicadores deslocados em 0 e 30 dias indicam os melhores desempenhos, porém existem melhores em outros períodos. Daqui, nota-se que pode haver um benefício considerando-se a variação dos índices, e não o valor normalizado do índice em si.

Considerando-se a base **Fake.Br**, nota-se um comportamento diferente, com um aumento muito mais singelo, e em alguns casos uma redução no desempenho. Nota-se também que o desempenho com a **modificação da base Fake.Br** (considerando-se apenas as cinquenta primeiras palavras) no modelo não é compensada pela inclusão dos indicadores sociais e macroeconômicos. Isso pode se dever pois a base **Fake.Br** deve ter características diferentes das apresentadas na base **FakeRecogna.**

6 DISCUSSÕES SOBRE OS RESULTADOS

De maneira a consolidar o entendimento dos resultados apresentados, e evidenciar eventuais lacunas e ameaças à validade da pesquisa, tem-se a necessidade da discussão dos resultados apresentados na seção anterior.

Ressalta-se que os resultados indicam a possibilidade de que não sejam considerados os valores dos indicadores sociais e macroeconômicos, e sim a variação dos indicadores, já que a alteração do humor geral pode ser mais suscetível à variação do índice, e não ao valor nominal. Nota-se isso pois, embora existam melhorias focadas em determinados períodos de deslocamento, esse efeito não é exclusivo nesses períodos, podendo se estender por períodos mais longos ou mais curtos.

Vale indicar também que foi considerado o *Bag Of Words* como modelo de Processamento Natural de Linguagem, quando existem técnicas mais avançadas que capturam o valor semântico da relação entre diversas palavras, como o *Word2Vec* e os *N-gramas*. Isso ocorreu, pois, as referências consideradas indicam bons resultados com o modelo *Bag Of Words*, porém mantém-se como ponto de atenção a possibilidade de utilização de outras metodologias.

Dos modelos classificatórios considerados; *Random Forest*, *Naive Bayes* e SVM, vale evidenciar-se a possibilidade de inclusão de técnicas mais recentes presente em estudos de inteligência artificial e aprendizado de máquina, como *deep learning* e Redes Neurais. Dado que o foco do estudo se prendia a identificação do possível efeito dos índices, não houve a exploração da hipótese com outras técnicas e modelos de classificação, porém mantém-se como ponto de atenção tal possibilidade. Ressalta-se também que o modelo *Naive Bayes* é paramétrico, e se vale de uma distribuição probabilística definida (para o estudo foi considerada a gaussiana). Já os modelos *Random Forest* e SVM são *não* paramétricos, portanto, os dados não necessitam ser provenientes de um tipo de distribuição de probabilidade. Também tem-se que o modelo SVM lida bem com a mistura entre variáveis categóricas e quantitativas, já o *Random Forest* tende a ter desempenho superior na utilização de variáveis apenas quantitativas, que é o caso do estudo apresentado.

7 CONCLUSÃO E ESTUDOS FUTUROS

Dado o que foi apresentado na seção anterior, mesmo embora existam pontos de eventual melhoria no estudo, que podem implicar no aperfeiçoamento do estudo e ser objeto de trabalhos futuros, demonstrou-se a possibilidade de serem incluídos índices sociais e macroeconômicos como variáveis quantitativas em um modelo classificatório que exerce a tarefa de detecção de *fake news*, ainda indicando o impacto positivo superior com possibilidade de redução da carga semântica da notícia, ou seja, com a redução da quantidade de palavras em uma notícia.

Como trabalhos futuros, dado que o processamento textual é algo custoso computacionalmente, uma possível linha de estudo seria a avaliação do ganho de desempenho com a redução *controlada* da quantidade de palavras em uma base de texto e a inserção de outros indicadores de contexto de dispersão da notícia, como os indicadores sociais e macroeconômicos. Isso é algo que pode ter aplicação prática, como na classificação de postagens em redes sociais, mensageiros instantâneos, portais de notícias e blogs, quando se imagina que a tarefa de classificação de *fake news* em tempo real se beneficiária de um processamento mais rápido.

Além disso, dado que houve uma diferença de desempenho dos resultados entre as bases FakeRecogna e Fake.Br, que ocorreu por diferenças de características linguísticas entre as bases, um possível estudo futuro poderia incluir avaliação e indicação de quais tipos de textos poderiam se beneficiar da inclusão de variáveis adicionais de contexto social e macroeconômico na tarefa de detecção de *fake news*, para assim ser avaliado quando seria cabível a inclusão dos indicadores no modelo classificatório dado o tipo de notícia a ser considerada. Considerando-se ainda a questão linguística, há também a possibilidade de serem considerados técnicas mais avançadas de processamento de linguagem natural, como o *Word2Vec* e N-gramas, dado que há uma dificuldade de classificadores lidarem com vetores de frequência muito extensos.

Além disso, podem ser considerados modelos de classificação mais robustos, com redes neurais e aprendizado profundo (*deep learning).*

Adicionalmente, é válido considerar uma linha de pesquise que avalie a possibilidade de inclusão de outros índices contextuais relevantes, como dados de pesquisa eleitorais, índices de crescimento econômico, taxa de câmbio, entre outros indicadores que se têm como hipótese que podem afetar o padrão de dispersão de notícias falsas na sociedade. Considerando-se os índices adicionais, vale também considerar a hipótese de ter-se a variação do índice, e não o

seu valor absoluto, tendo-se como hipótese que a variação em si que pode implicar em alguma alteração no humor da sociedade quanto à dispersão de uma quantidade superior de *fake news.*

REFERÊNCIAS

ABONIZIO, H. Q. et al. Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features. **Future Internet**, v. 12, n. 5, p. 87, 11 maio 2020.

AGARWAL, V. et al. Analysis of Classifiers for Fake News Detection. **Procedia Computer Science**, v. 165, p. 377–383, 2019.

BIN GUO et al. The Future of False Information Detection on Social Media: New Perspectives and Trends. **ACM Computing Surveys**, v. 53, n. 4, p. 68-68:36, 1 ago. 2020. Disponível em: <http://eds.b.ebscohost.com/eds/detail/detail?vid=3&sid=1e954d27-581e-48b6-bab6-667b34668f8f%40sessionmgr101&bdata=JkF1dGhUeXBlPWlwLHNoaWImc2l0ZT1lZHMtbGl2ZSZzY29wZT1zaXRl#AN=145361161&db=bth>.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://link.springer.com/article/10.1023/A:1010933404324>.

COELHO, A. **Stemmming para a língua portuguesa: estudo, análise e melhoria de algoritmo RSLP.** 2007. UFRGS, 2007. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/23576/000597277.pdf?sequence=1>. Acesso em: 21 abr. 2022.

DEMŠAR, J. Statistical Comparisons of Classifiers over Multiple Data Sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006. Disponível em: <https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf>.

FAUSTINI, P. H. A.; COVÕES, T. F. Fake news detection in multiple platforms and languages. **Expert Systems with Applications**, v. 158, p. 113503, nov. 2020. Acesso em: 22 abr. 2021.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, jun. 2006. Acesso em: 18 set. 2019.

GARCIA, G. L.; AFONSO, L. C. S.; PAPA, J. P. FakeRecogna: A New Brazilian Corpus for Fake News Detection. **Lecture Notes in Computer Science**, p. 57–67, 2022. Acesso em: 22 ago. 2022.

GRANSKOGEN, T. **Automatic Detection of Fake News in Social Media using Contextual Information**. 2018. NTNU, 2018.

GUDERLEI, M.; ASSENMACHER, M. Evaluating Unsupervised Representation Learning for Detecting Stances of Fake News. **Proceedings of the 28th International Conference on Computational Linguistics**, 2020. Acesso em: 22 ago. 2022.

HALL, M. A.; WITTEN, I. H. **Data Mining : Practical Machine Learning Tools and Techniques 3rd edition Ed. 3**. [s.l.] Morgan Kaufmann Publishers, 2011.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29–36, abr. 1982.

HOFMANN, T.; SCHÖLKOPF, B.; SMOLA, A. J. Kernel methods in machine learning. **The Annals of Statistics**, v. 36, n. 3, p. 1171–1220, jun. 2008.

IBOVESPA B3 | B3. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm>.

ÍNDICE NACIONAL DE PREÇOS AO CONSUMIDOR AMPLO | IBGE. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/precos-e-custos/9256-indice-nacional-de-precos-ao-consumidor-amplo.html?=&t=o-que-e>.

JOÃO PAULO II. , 4 mar. 1979. Disponível em: <https://www.vatican.va/content/john-paul-ii/pt/encyclicals/documents/hf_jp-ii_enc_14091998_fides-et-ratio.html#%242Y>. Acesso em: 24 out. 2022.

KALIYAR, R. K. et al. FNDNet – A deep convolutional neural network for fake news detection. **Cognitive Systems Research**, v. 61, p. 32–44, jun. 2020. Acesso em: 21 mar. 2020.

KUHN, M.; JOHNSON, K. **Feature engineering and selection a practical approach for predictive models**. [s.l.] Boca Raton London New York Chapman & Hall/Crc, 2020.

QUINTANILHA, T.; DA SILVA, M.; LAPA, T. Fake news and its impact on trust in the news. Using the Portuguese case to establish lines of differentiation. **Communication & Society**, v. 32, n. 3, p. 17–33, 2019.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 20 dez. 2007. Acesso em: 29 out. 2020.

MAHESH, B. Machine Learning Algorithms - A Review. **International Journal of Science and Research (IJSR)**, v. 5, n. 1, p. 641–645, 5 jan. 2016.

MARTINEZ, T. S.; CERQUEIRA, V. dos S. Estrutura da inflação brasileira: determinantes e desagregação do IPCA. **Economia e Sociedade**, v. 22, n. 2, p. 409–456, ago. 2013. Acesso em: 30 nov. 2021.

MITCHELL, A. Key findings on the traits and habits of the modern news consumer. **policycommons.net**, 7 jul. 2016. Disponível em: <https://policycommons.net/artifacts/618417/key-findings-on-the-traits-and-habits-of-the-modern-news-consumer/1599362/>. Acesso em: 22 ago. 2022.

MONTEIRO, R. A. et al. **Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results**. Disponível em: <https://www.semanticscholar.org/paper/Contributions-to-the-Study-of-Fake-News-in-New-and-Monteiro-Santos/d551f16971f5dee10bcf2ba82229cda359a3c890>. Acesso em: 23 fev. 2022.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544–551, set. 2011. Disponível em: <https://academic.oup.com/jamia/article/18/5/544/829676>.

ORENGO, V. M.; HUYCK, C. A stemming algorithm for the portuguese language. **Proceedings Eighth Symposium on String Processing and Information Retrieval**, 2002. Acesso em: 22 ago. 2022.

OZBAY, F. A.; ALATAS, B. Fake news detection within online social media using supervised artificial intelligence algorithms. **Physica A: Statistical Mechanics and its Applications**, v. 540, p. 123174, out. 2019.

PENTINA, I.; TARAFDAR, M. From “information” to “knowing”: Exploring the role of social media in contemporary news consumption. **Computers in Human Behavior**, v. 35, p. 211–223, jun. 2014.

PÉREZ-ROSAS, V.; MIHALCEA, R. Experiments in Open Domain Deception Detection. **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, 2015. Acesso em: 22 ago. 2022.

PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS CONTÍNUA - PNAD CONTÍNUA | IBGE. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/trabalho/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=o-que-e>.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, mar. 1998.

BOKKA, K. R. et al. **Deep learning for natural language processing solve your natural language processing problems with smart deep neural networks**. [s.l.] Birmingham Packt June, 2019.

REIS, J. C. S. et al. Supervised Learning for Fake News Detection. **IEEE Intelligent Systems**, v. 34, n. 2, p. 76–81, mar. 2019. Disponível em: <https://sentic.net/supervised-learning-for-fake-news-detection.pdf>.

SCHONLAU, M.; ZOU, R. Y. The random forest algorithm for statistical learning. **The Stata Journal: Promoting communications on statistics and Stata**, v. 20, n. 1, p. 3–29, mar. 2020.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys**, v. 34, n. 1, p. 1–47, 1 mar. 2002. Disponível em: <https://dl.acm.org/citation.cfm?id=505283>.

SHU, K. et al. Fake news detection on social media: A data mining perspective. **ACM SIGKDD Explorations Newsletter**, v. 19, n. 1, p. 22–36, 1 set. 2017. Disponível em: <https://dl.acm.org/citation.cfm?id=3137600>.

TANDOC, E. C.; LIM, Z. W.; LING, R. Defining “Fake News”. **Digital Journalism**, v. 6, n. 2, p. 137–153, 30 ago. 2017.

ZHANG, H. **The Optimality of Naive Bayes**. 2004. Disponível em: <http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>.