Detecção Automática de Notícias Falsas para o Português

Relatório Final de Projeto

Bolsista: Rafael Augusto Monteiro
Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo
Período: Agosto de 2017 a Julho de 2018

Resumo

Este relatório refere-se às atividades desenvolvidas durante o período de vigência da bolsa do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) - no período de Agosto de 2017 a Julho de 2018 – do projeto “Detecção Automática de Notícias Falsas para o Português”. Neste projeto, fazendo uso de técnicas de Processamento de Linguagem Natural e Aprendizado de Máquina, foram estudados alguns métodos para detecção automática de notícias falsas disseminadas na web. Os resultados obtidos são promissores e, pelo que se conhece da área, são os primeiros do tipo produzidos para a língua portuguesa.

1. Introdução

As transformações nos meios de comunicação, provocadas pela popularização de redes sociais, fóruns, blogs e outros ambientes na web, permitiram que os usuários se tornassem agentes ativos na divulgação de informações. Torna-se cada vez mais comum o compartilhamento e disseminação de conteúdo informativo, como notícias, análises e opiniões.

Contudo, tal conteúdo disseminado nem sempre é verídico. Notícias enganosas, opiniões forjadas sobre produtos e serviços, manchetes sensacionalistas e até mesmo textos satíricos são confundidos com suas contrapartidas verdadeiras, deixando em cheque a confiança na enorme quantidade de informações compartilhadas diariamente.

A detecção de conteúdo enganoso na web torna-se ainda mais preocupante quando consideramos a capacidade humana de diferenciar o que é enganoso ou não. Estudos realizados por Bond e DePaulo (2006) [1], por exemplo, avaliam que humanos acertam, em média, 54% dos julgamentos sobre verdades e mentiras, o que é próximo a uma escolha aleatória. Portanto, é de grande importância criar ferramentas que auxiliem na detecção da veracidade das informações presentes na web.

Uma das formas de detectar conteúdo falso é realizar a checagem dos fatos apresentados, através de jornalismo investigativo. Infelizmente, a complexidade em averiguar a veracidade dos fatos torna tal tarefa inviável para o grande volume de informações. Outra forma é por meio da análise de características textuais. Ao criar conteúdo enganoso, os autores costumam (involuntariamente) deixar marcas linguísticas e dicas textuais que permitem inferir, com alguma confiança, se um texto é falso ou não. Felizmente, tal análise pode ser feita automaticamente, por meio de

técnicas de Processamento de Linguagem Natural (PLN), subárea da Inteligência Artificial que estuda o processamento automático de linguagens faladas e escritas.

Inspirado nos trabalho de Pérez-Rosa e Mihalcea (2015) [2] e Pérez-Rosa et al. (2017) [3], nossa abordagem busca utilizar técnicas de PLN para extrair atributos de textos verdadeiros e falsos coletados da web e, com base em Aprendizado de Máquina (AM) (Mitchell, 1997) [4], criar classificadores que possam inferir se uma notícia é verdadeira ou falsa.

Na Seção 2, são descritos de forma detalhada os objetivos do projeto e é apresentado o cronograma original de atividades previstas. Na Seção 3, são descritos os métodos, técnicas e ferramentas utilizados em cada etapa do projeto. Na Seção 4, são apresentados e discutidos os resultados obtidos. Por fim, na Seção 5, são feitas algumas considerações finais sobre o projeto.

É importante ressaltar que este projeto foi desenvolvido em parceria com o aluno de doutorado do ICMC Roney L. S. Santos.

2. Objetivos

O objetivo deste projeto foi (i) investigar métodos de detecção automática de informações falsas na web, escritas em português, e, em paralelo, (ii) introduzir o aluno à área de pesquisa em PLN, com seus desafios e métodos, que subsidiaram a pesquisa do item (i).

Como ponto de partida para a investigação de métodos para detecção automática de informações falsas, foi utilizado o trabalho de Pérez-Rosa e Mihalcea (2015), que avaliou a eficácia de atributos textuais para a tarefa, utilizando técnicas de AM para produzir um classificador de informações falsas. Portanto, inicialmente, os objetivos do projeto foram:

a) Compilar um córpus de notícias falsas e verdadeiras para subsidiar o treinamento e teste dos métodos de AM;
b) Adaptar e criar os recursos e ferramentas linguístico-computacionais necessários para os métodos estudados;
c) Implementar os métodos utilizados em Pérez-Rosa e Mihalcea (2015), além de outros métodos da literatura (a depender do tempo disponível no decorrer do projeto), com o intuito de demonstrar e comparar a eficácia de cada método;
d) Avaliar os métodos implementados sobre o córpus compilado, destacando suas vantagens e limitações.

Em relação ao item (a), não foi encontrado nenhum córpus público e etiquetado de notícias falsas escritas em português. Portanto, a coleta de um córpus foi uma importante tarefa dentre as realizadas no projeto.

Com relação ao item (b), as ferramentas e recursos necessários para implementar os métodos estudados já existiam, como o etiquetador morfossintático NLPNet (Fonseca e Rosa, 2013) [5] e o léxico LIWC (Balage et al., 2013) [6], o qual lista as palavras do português e suas possíveis associações com classes semânticas

e com sentimentos, além de outras ferramentas básicas de pré-processamento de texto, como o stemmer do projeto Snowball[1] e listas de *stopwords,* entre outras.

No item (c), foi prevista a implementação de todos os métodos utilizados por Pérez-Rosa e Mihalcea, além da adaptação de eventuais outros métodos encontrados na literatura, como utilização de atributos linguísticos propostos por Zhou et al. (2004) [7] para indicação de conteúdo enganoso.

Finalmente, no item (d), uma análise comparativa entre todos os métodos implementados é proposta. A análise deve destacar as medidas de precisão, cobertura/revocação e medida-f de cada método para avaliar os erros e acertos na classificação de informações falsas. Ainda, métricas e representações de AM, como medida geral de acurácia e matriz de confusão, também serão utilizadas na avaliação.

Por fim, os resultados obtidos devem ser divulgados em resumos e artigos de divulgação em eventos científicos, como os realizados pela própria instituição (por exemplo, SIICUSP) e outros de alcance geral, como o PROPOR (*International Conference on the Computational Processing of Portuguese*), conforme se comenta no fim deste relatório.

O cronograma originalmente previsto para a realização do projeto é mostrado abaixo, no Quadro 1, onde as principais atividades são listadas. O cronograma é dividido em bimestres. Todas as etapas foram cumpridas, conforme se relata a seguir.

Quadro 1. Cronograma originalmente previsto para a realização do projeto

Atividades	Bim. 1	Bim. 2	Bim. 3	Bim. 4	Bim. 5	Bim. 6
Familiarização com a área de PLN e de AM						
Revisão literária sobre métodos de detecção de notícias e informações falsas na web						
Compilação de córpus de referência de notícias/informações falsas						
Criação/extensão/estudo de recursos e ferramentas linguístico-computacionais necessárias ao projeto						
Implementação do método de Pérez-Rosa e Mihalcea (2015)						
Implementação complementar de outros métodos de literatura e de métodos *baseline*						
Avaliação comparativa dos métodos						
Disponibilização de protótipo computacional com o melhor método estudado						
Produção de resumos e artigos para eventos da área						
Escrita de relatórios semestrais						

[1] http://snowball.tartarus.org/

Como resultados esperados do projeto, destacam-se: (i) aprofundamento de conhecimentos do aluno nas áreas de PLN e AM, cumprindo um dos principais objetivos desse programa de bolsas; e (ii) contribuições científicas na área, como a criação do primeiro córpus disponível publicamente e etiquetado de notícias falsas escritas em português e o estudo de métodos e a disponibilização de um software de detecção de notícias falsas para o português.

A seguir, as etapas principais do trabalho realizado são brevemente relatadas.

3. Metodologia

3.1. Coleta e Criação do Córpus

De acordo com Rubin et al. (2015) [8], existem três tipos principais de conteúdo enganoso em textos: (i) os humorísticos, utilizados para diversão, usando sarcasmo para fazer sátiras e paródias; (ii) os de conteúdo falso, que tem o propósito claro de enganar e causar confusões, e (iii) os boatos, que não possuem confirmação e geralmente são aceitos publicamente. Notícias falsas são um tipo de texto com conteúdo falso, assim como as avaliações falsas, por exemplo, que são criadas especificamente para prejudicar ou promover algo.

Os autores também definem características importantes para um córpus de notícias falsas: o córpus deve conter notícias falsas e verdadeiras, para que métodos preditivos possam encontrar padrões e regularidades; o córpus deve estar em um formato acessível para ferramentas de PLN, como texto simples; as notícias do córpus devem ser homogêneas em tamanho, na medida do possível (caso necessário, a normalização do tamanho - via truncagem, por exemplo - pode ser feita); as notícias devem ser homogêneas em tópicos abordados (negócios, ciência, política, etc); as notícias devem ser coletadas com base em um período de tempo pré-definido, já que há variação de estilo de escrita no tempo; informações extras, como nome do autor, *link* da notícia, etc., também devem ser coletadas, dado que tais informações podem auxiliar na checagem de fatos.

Inicialmente, com o intuito de capturar instâncias representativas do problema abordado, foi feita uma coleta de mensagens com conteúdo enganoso. As mensagens, que eram difundidas via aplicativos de comunicação, foram coletadas por colaboradores do projeto, extraídas de sites de checagem de fatos e páginas em redes sociais. No total, foram coletadas cerca de 70 mensagens com boatos ou notícias falsas, com sua falsidade confirmada por pesquisa manual. Porém, dada a dificuldade na coleta de mensagens compartilhadas, e que boa parte das mensagens coletadas possuíam recursos como áudios, vídeos e imagens, que fugiam ao escopo deste projeto, a utilização de tais mensagens no projeto foi descartada.

Em seguida, foi iniciada uma coleta semi-automática de notícias falsas difundidas nos sites Diário do Brasil, A Folha do Brasil, The Jornal Brasil e Top Five TV. As notícias foram baixadas utilizando um *crawler*, e passaram por uma análise manual, garantindo que todas contivessem conteúdo falso. As notícias que utilizavam fatos verdadeiros para apoiar conclusões enganosas, também chamadas de meias verdades (Clem, 2017) [9], não foram incluídas no córpus. No total, foram coletadas 3600 notícias falsas.

Para garantir o balanceamento do corpus, foram coletadas notícias verdadeiras que fossem similares às notícias falsas. De início, a coleta das notícias verdadeiras era feita manualmente, buscando por palavras-chave de cada notícia falsa nos portais G1, Folha de São Paulo e Estadão, que são reconhecidos publicamente como veículos de notícias confiáveis. Eram escolhidas notícias que citassem o maior número possível de temas, entidades, instituições ou pessoas apresentadas na notícia falsa. Porém, a coleta manual se mostrou muito desafiadora, devido à alta subjetividade na escolha da notícia mais similar, e na quantidade de notícias apresentadas em cada busca. Portanto, uma abordagem semi-automática foi novamente utilizada. Por meio de um *crawler*, cerca de 40.000 notícias verdadeiras dos portais foram coletadas. Em seguida, para cada notícia falsa do córpus, foi selecionada a notícia verdadeira mais similar (com base na métrica da similaridade do cosseno, como apresentada por (Salton e McGill, 1986) [10]). Por fim, uma verificação manual foi feita, para garantir que a notícia verdadeira escolhida era similar em tópico à notícia falsa. O Quadro 2 apresenta trechos de notícias verdadeiras e falsas coletadas. É importante notar que a notícia verdadeira não necessariamente desmente a notícia falsa, como apresentado nos exemplos do Quadro 2.

Quadro 2. Exemplos de notícias verdadeiras e falsas coletadas

Notícia Real	Notícia 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.	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 cancelar o carnaval de 2018.
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 Paranoá. ‘Prática permite que mulher possa optar por participar de tais eventos sociais’, diz texto.	Acabou a mordomia ! Ingresso mais barato pra mulher é ilegal. Baladas que davam meia entrada para mulher, ou até mesmo gratuidade, estão na ilegalidade agora. Acabou o preconceito com os homens nas casas de show de todo o Brasil.

Analisando-se a distribuição de temas das notícias coletadas no córpus, tem-se que: 58% das notícias coletadas falam sobre política; 21,4% são sobre TV e celebridades; 17,7% são sobre sociedade e eventos do dia a dia; 1,5% são sobre ciência e tecnologia; 0,7% são sobre economia; e 0,7% são sobre religião.

3.2. Extração de Atributos

Para criar um classificador automático utilizando aprendizado de máquina, é necessário que atributos textuais das notícias sejam extraídos e representados em um formato numérico. Os atributos extraídos, segundo as propostas de Pérez-Rosa e Mihalcea (2015) e Zhou et al. (2004), foram:

- **Unigramas / *bag of words*:** representação em *bag of words* da ocorrência ou não das palavras presentes nos textos, utilizando valores booleanos. A extração foi feita após conversão das palavras do texto para caixa baixa, remoção de numerais, pontuação e *stopwords* (palavras de classe fechada, como preposições, pronomes e artigos), e radicalização das palavras.
- **POS Tags (etiquetas morfossintáticas, ou classes gramaticais):** número normalizado de ocorrências de cada etiqueta morfossintática no texto, indicado pelo etiquetador/tagger NLPNet (Fonseca e Rosa, 2013).
- **Classes semânticas do LIWC:** número normalizado de ocorrências de cada uma das 64 classes semânticas indicadas pelo *Brazillian Portuguese LIWC Dictionary* (Balage et al., 2013).
- **Pausalidade, Emotividade, Incerteza e Não Imediatismo:** número normalizado dos atributos definidos como possíveis indicadores de conteúdo enganoso em textos eletrônicos, a saber:
    - Pausalidade: indicativo da frequência de pausas em um texto, calculado pela razão entre o número de sinais de pontuação e o número de sentenças.
    - Emotividade: indicativo da expressividade linguística em um texto, calculada pela soma de adjetivos e advérbios dividida pela soma de substantivos e verbos.
    - Incerteza: medida pelo número de verbos modais e ocorrências de voz passiva no texto.
    - Não Imediatismo: calculada pelo número de pronomes de 1a e 2a pessoa no texto.

As normalizações citadas acima foram feitas dividindo-se os números calculados pela contagem total de palavras presentes no texto. Sendo assim, todos os valores calculados eram números no intervalo entre 0 e 1, garantindo o funcionamento adequado dos algoritmos de AM.

A extração de atributos foi implementada por meio de *scripts* da linguagem Python. A radialização foi feita utilizando a implementação do *stemmer* do projeto Snowball presente no pacote NLTK. A contagem dos atributos Pausalidade e Emotividade foi feita utilizando o parser NLPNet, enquanto a contagem dos atributos Incerteza e Não Imediatismo foi feita utilizando o parser PALAVRAS (Bick, 2000) [11].

3.3. Aprendizado de Máquina

Com o intuito de classificar automaticamente as notícias, foram explorados algoritmos de classificação baseados em Aprendizado de Máquina de diferentes paradigmas. Inicialmente, alguns experimentos foram conduzidos durante a construção do córpus, utilizando a ferramenta Weka[2], amplamente utilizada na área. Finalizada a construção do córpus, novos experimentos foram realizados, utilizando a biblioteca Scikit-Learn, da linguagem Python.

Todos os testes realizados foram conduzidos utilizando o método de validação cruzada em 5 *folds*, conforme feito por Pérez-Rosa e Mihalcea (2015). O método

[2] https://www.cs.waikato.ac.nz/ml/weka/

divide o conjunto de dados em 5 partes de tamanho igual, utilizando uma parte para teste e as demais partes para treinamento dos classificadores. Em seguida, o processo é realizado novamente, selecionando-se outra parte do conjunto de dados para testes, e as demais para treinamento. O processo é repetido até que todas as partes já tenham sido utilizadas como conjunto de testes. Ao final do processo, são classificadas todas as instâncias do conjunto de dados. Ao comparar a classificação de cada instância com sua classe real, é possível obter métricas de desempenho, como acurácia, precisão e revocação, por exemplo.

Os primeiros testes foram realizados durante a construção do córpus, com 674 notícias, sendo 337 notícias falsas e 337 notícias verdadeiras, utilizando os atributos unigramas, POS tags e classes semânticas do LIWC, e os algoritmos de classificação SVM, Naive-Bayes e Random Forest, que são amplamente utilizados em tarefas de classificação de textos, e árvores de decisão J48, que permitem visualizar de forma compreensível a estrutura de decisão gerada. A acurácia geral obtida por cada classificador, utilizando-se os atributos separadamente, é apresentada na Tabela 1, sendo que o melhor resultado para cada atributo é destacado em negrito.

Tabela 1. Resultados dos experimentos preliminares

Classificador	*Bag of words*	LIWC	POS
Naive-Bayes	74,04%	68,25%	64,10%
J48	67,66%	64,10%	62,61%
RandomForest	76,26%	**72,11%**	**73,44%**
SVM	**81,16%**	61,13%	64,84%

Por meio desses testes, foi possível observar a eficácia dos classificadores SVM e RandomForest, que foram utilizados novamente nos testes seguintes. Ainda, o algoritmo J48 apresentou baixo desempenho, além da visualização da árvore de decisão gerada ser confusa devido ao alto número de nós, e não foi utilizado nos testes seguintes.

Após a finalização do córpus, foram realizados novos experimentos, utilizando todas as 7.200 notícias coletadas. Os algoritmos utilizados foram os mesmos, excluindo-se o J48 e adicionando-se a rede neural *Multilayer Perceptron* (MLP), algoritmo conexionista que apresenta bons resultados em bases de dados com grande número de exemplos.

Com o intuito de evitar viés na classificação devido à variação de tamanho entre as notícias verdadeiras e falsas, os experimentos finais foram realizados após o truncamento dos textos. Para isso, era analisado cada par de notícias verdadeira-falsa, e a maior notícia era truncada no mesmo número de palavras da menor notícia. Um exemplo do truncamento de texto é apresentado no Quadro 3.

Também foram exploradas técnicas para selecionar um número menor de atributos, com a intenção de melhorar a performance dos classificadores e gerar modelos de seleção mais compactos. Foi feita a seleção de atributos, baseada no ganho de informação mútua. Além disso, também se testou a análise de componentes principais (PCA). Ambas as técnicas foram utilizadas com a implementação presente

no pacote Scikit-Learn, com parâmetros padrões, e foram selecionados os 20 atributos mais relevantes.

Por fim, para o modelo *Bag of Words*, foi avaliada a seleção de palavras com mais ocorrências no córpus. Para isso, foram selecionadas apenas as palavras que ocorressem com frequência maior que um limite pré-estabelecido. Foram testados diversos valores para o limite.

A Seção 4 apresenta e discute todos os resultados obtidos nos experimentos finais.

Quadro 3. Exemplo do truncamento realizado nos textos

Notícias originais, sem truncamento:	
Notícia falsa	Notícia verdadeira
Janaína, a mulher que representa os brasileiros honestos, denuncia: "Querem acabar com o Sérgio Moro". O PT é uma vergonha para a política. Não há motivos para duvidar da advogada e professora de direito Janaína Paschoal. É só lembrar as declarações divulgadas pelos próprios petistas. Eles disseram: "O partido é o inimigo número 1 da Operação Lava Jato" Abaixo um vídeo onde Janaína Paschoal, na reunião na Comissão do Impeachment, lembra claramente das várias ações impetradas por parlamentares do PT e do PCdoB contra o juiz Sérgio Moro. O que Janaína demonstra claramente é a famosa "tática de assédio jurídico"	Alvo da Lava Jato, Renan sofre resistência no PMDB para liderar a bancada. Senador, que vai deixar presidência do Senado em fevereiro, tem o nome cotado para assumir a liderança do partido ou a Comissão de Constituição e Justiça. . O futuro do presidente do Senado, Renan Calheiros (PMDB-AL), após o fim de seu mandato à frente da Casa, em fevereiro, ainda está indefinido. A probabilidade maior é a de que Renan ocupe a liderança do PMDB na Casa – função que já desempenhou outras duas vezes – mas dentro da bancada há resistência ao nome dele, principalmente por conta dos processos do senador na Justiça e os inquéritos de que é alvo na Lava Jato. Peemedebistas ouvidos pelo G1 dizem que o fato de ele ser investigado pode desestabilizar e constranger o partido caso assuma a liderança da bancada. Em dezembro, o STF tornou Renan réu pelo crime de peculato [...]
Após o truncamento:	
Notícia falsa	Notícia verdadeira
Janaína, a mulher que representa os brasileiros honestos, denuncia: "Querem acabar com o Sérgio Moro". O PT é uma vergonha para a política. Não há motivos para duvidar da advogada e professora de direito Janaína Paschoal. É só lembrar as declarações divulgadas pelos próprios petistas. Eles disseram: "O partido é o inimigo número 1 da Operação Lava Jato" Abaixo um vídeo onde Janaína Paschoal, na reunião na Comissão do Impeachment, lembra claramente das várias ações	Alvo da Lava Jato, Renan sofre resistência no PMDB para liderar a bancada. Senador, que vai deixar presidência do Senado em fevereiro, tem o nome cotado para assumir a liderança do partido ou a Comissão de Constituição e Justiça. O futuro do presidente do Senado, Renan Calheiros (PMDB-AL), após o fim de seu mandato à frente da Casa, em fevereiro, ainda está indefinido. A probabilidade maior é a de que Renan ocupe a liderança do PMDB na Casa – função que já desempenhou outras duas

impetradas por parlamentares do PT e do PCdoB contra o juiz Sérgio Moro. O que Janaína demonstra claramente é a famosa "tática de assédio jurídico"	vezes – mas dentro da bancada há resistência ao nome dele, principalmente por conta dos processos do senador na Justiça

4. Resultados e Análises

4.1. Experimentos com Córpus Completo

A Tabela 2 apresenta uma comparação entre os atributos utilizados no treinamento do classificador SVM (utilizada a implementação LinearSVC do pacote Scikit-Learn, com os parâmetros padrões). São apresentados, para cada conjunto de atributos, a precisão, a revocação e a medida-f por classe, além da acurácia geral. Os experimentos foram conduzidos utilizando a validação cruzada em 5 *folds*.

A tabela também apresenta o uso combinado de alguns atributos. As três primeiras linhas apresentam os atributos inspirados em [2], enquanto a quarta linha apresenta a combinação deles; as quatro linhas seguintes apresentam os atributos de [7], seguidas pela combinação entre eles; em seguida, é feita a combinação entre o melhor atributo de [2] e o melhor atributo de [7]; por fim, a última linha apresenta a combinação de todos os atributos.

Tabela 2. Comparação do desempenho dos diferentes conjuntos de atributos em classificações com o algoritmo SVM

Atributos	Precisão		Revocação		Medida-f		Acurácia
	Falsa	Real	Falsa	Real	Falsa	Real	
POS tags	0,76	0,74	0,73	0,77	0,74	0,76	0,75
Classes Semânticas	0,73	0,73	0,73	0,73	0,73	0,73	0,73
Bag of Words	0,88	0,89	0,89	0,88	0,88	0,88	0,88
POS + Classes semânticas + *bag of words*	0,88	0,89	0,89	0,88	0,89	0,89	0,88
Pausalidade	0,52	0,52	0,58	0,46	0,55	0,49	0,50
Emotividade	0,57	0,56	0,53	0,61	0,55	0,58	0,56
Incerteza	0,51	0,51	0,46	0,57	0,48	0,54	0,51
Não-imediatismo	0,53	0,51	0,16	0,86	0,24	0,64	0,51
Pausalidade + emotividade + incerteza + não-imediatismo	0,57	0,56	0,53	0,60	0,55	0,58	0,57
Bag of words + emotividade	0,88	0,89	0,89	0,88	0,89	0,89	0,89
Todos os Atributos	0,89	0,89	0,89	0,89	0,89	0,89	0,89

Observa-se que o *bag of words* atingiu resultados muito bons, com uma medida-f de 88%, tanto para notícias falsas quanto verdadeiras. Ainda, a combinação do *bag of words* com os outros atributos não trouxe melhorias significativas. Diferentemente de [2], POS tags não apresentaram o melhor resultado.

Dado o alto desempenho do uso de *bag of words*, foi explorada a seleção de palavras que ocorrem no córpus com frequência acima de um determinado valor. A Tabela 3 apresenta os resultados encontrados para diferentes frequências mínimas. É importante ressaltar que a frequência mínima é avaliada no córpus inteiro, e não em cada texto.

Tabela 3. Comparação do desempenho do uso de *bag of words*, com o classificador SVM, onde são utilizadas apenas palavras acima de uma frequência mínima pré-definida no córpus

Frequência Mínima	Precisão		Revocação		Medida-f		Acurácia	Número de Palavras no Conj. de Dados
	Falsa	Real	Falsa	Real	Falsa	Real		
1	0,88	0,89	0,89	0,88	0,88	0,88	0,89	24.065
3	**0,88**	**0,89**	**0,89**	**0,88**	**0,89**	**0,89**	**0,89**	**10.480**
5	0,88	0,88	0,88	0,88	0,88	0,88	0,88	7.847
10	0,87	0,88	0,88	0,87	0,87	0,87	0,87	5.326
15	0,87	0,87	0,87	0,88	0,87	0,87	0,87	4.238
20	0,87	0,87	0,87	0,87	0,87	0,87	0,87	3.630
30	0,87	0,87	0,87	0,87	0,87	0,87	0,87	2.835
40	0,86	0,86	0,86	0,86	0,86	0,86	0,86	2.335
50	0,86	0,86	0,86	0,86	0,86	0,86	0,86	2.019
100	0,85	0,85	0,85	0,85	0,85	0,85	0,85	1.186
150	0,85	0,86	0,86	0,85	0,86	0,85	0,85	840
200	0,86	0,86	0,86	0,85	0,86	0,86	0,85	644
300	0,85	0,85	0,85	0,85	0,85	0,85	0,85	408
400	0,84	0,84	0,85	0,83	0,84	0,84	0,84	272
500	0,82	0,84	0,84	0,82	0,83	0,83	0,83	205
1000	0,77	0,81	0,82	0,75	0,79	0,78	0,78	55

Na primeira linha, é possível observar que, sem nenhum corte, foram escolhidas 24.065 palavras. Na linha seguinte, podemos observar que, ao utilizar a frequência mínima 3, é possível reduzir o número de palavras no conjunto de dados pela metade, o que diminui o tempo de treinamento e o tamanho do modelo gerado, sem perdas de performance. Nas linhas seguintes, é possível observar a queda do desempenho

conforme mais palavras são removidas do conjunto de dados, o que torna o corte de frequência prejudicial.

Além da remoção de palavras com poucas ocorrências, também foram exploradas duas outras técnicas de seleção de atributos: baseada em ganho de informação mútua e baseada em análise de componentes principais. Os métodos foram utilizados para selecionar os 20 melhores atributos do *bag of words*, e foram avaliados utilizando validação cruzada com o algoritmo SVM. Ao realizar a seleção com o ganho de informação mútua, o algoritmo SVM obteve 72% de acurácia geral, desempenho expressivamente inferior aos 88% obtidos sem a seleção. Ao substituir a seleção baseada em ganho de informação mútua pela análise de componentes principais, a acurácia geral foi de 84%, demonstrando a maior eficiência da técnica na redução de dimensionalidade do conjunto de dados.

Para avaliar a influência dos algoritmos de classificação, foram realizados testes com diferentes algoritmos. A Tabela 4 apresenta o desempenho de diferentes classificadores utilizando *bag of words* como atributos, com palavras que ocorrem no córpus com frequência mínima 3. As implementações utilizadas foram as disponibilizadas no pacote Scikit-Learn, com os parâmetros padrões (foram utilizadas as classes MultinomialNB para Naive-Bayes, LinearSVC para SVM, RandomForestClassifier para Random Forest, e MLPClassifier para MLP).

Tabela 4. Comparação do desempenho de diferentes algoritmos de classificação, utilizando *bag of words* como atributos, com palavras que ocorrem ao menos 3 vezes no corpus

Classificador	Precisão		Revocação		Medida-f		Acurácia
	Falsa	Real	Falsa	Real	Falsa	Real	
Naive-Bayes	0,88	0,87	0,86	0,88	0,87	0,87	0,87
Support Vector Machine (SVM)	0,89	0,89	0,89	0,89	0,89	0,89	0,89
Random Forest	0,75	0,81	0,83	0,72	0,79	0,76	0,77
Multilayer Perceptron (MLP)	0,91	0,90	0,90	0,91	0,91	0,91	0,90

É possível observar que o MLP atinge desempenho marginalmente superior ao SVM, mas com um tempo de treinamento maior: o treinamento e teste via validação cruzada do SVM demorou, em média, 11 segundos, enquanto o MLP levou cerca de 5 minutos. Ainda, o Naive-Bayes também apresentou uma boa performance, dada a simplicidade do algoritmo, enquanto o desempenho encontrado pelo algoritmo Random Forest nos testes preliminares não se manteve, apresentando o menor desempenho entre os algoritmos testados.

Na Tabela 5, apresentamos uma matriz de confusão para a classificação, utilizando *bag of words* como atributos e o algoritmo SVM. A matriz de confusão permite visualizar a distribuição dos acertos e erros por classe, podendo indicar um possível viés do algoritmo para uma determinada classe.

Tabela 5. Matriz de confusão da classificação utilizando *bag of words* e SVM

		Classes Reais	
		Real	Falsa
Classificado como	Real	3.192	432
	Falsa	408	3.168

Como indicado pela Tabela 5, os erros estão distribuídos de forma balanceada entre as classes, o que condiz com os resultados apresentados anteriormente. Acredita-se que, assim como para detecção de spam, é mais prejudicial classificar erroneamente notícias verdadeiras (que seriam filtradas, por exemplo) do que não detectar algumas notícias falsas. Portanto, ainda podem ser exploradas opções que permitam tais melhorias.

Com relação à distribuição dos erros em relação ao tema das notícias, pudemos observar as seguintes relações: 11,6% das notícias relacionados à política foram classificadas incorretamente; para notícias sobre TV e celebridades, o erro foi de 10,4%; 12,3% para notícias sobre o dia a dia e sociedade; 16,1% para notícias sobre ciência e tecnologia; 18,1% para notícias sobre economia; e 20,4% de erro na classificação de notícias sobre religião. Economia e religião foram as categorias com maior dificuldade na classificação, o que pode ocorrer devido ao pequeno número de notícias dessas categorias no córpus.

Por fim, foram realizados testes sem o truncamento dos textos. Com o algoritmo SVM e usando *bag of words* como atributos, foi possível atingir 96% de acurácia. Porém, é muito provável que o desempenho alto seja resultado de *overfitting*, dado que as notícias verdadeiras são, em geral, muito maiores do que as falsas.

4.2. Interface Web

Com o intuito de aplicar o conhecimento adquirido em uma aplicação real, foi desenvolvida uma interface web simples, que realiza a classificação automática de uma notícia em falsa ou verdadeira. Para isso, basta que o usuário insira o corpo da notícia em uma caixa de texto e selecione um dos modelos disponibilizados. As Imagens 1 e 2 mostram a versão inicial da interface.

A interface foi desenvolvida utilizando o framework Django, que permite o desenvolvimento de aplicações web utilizando a linguagem Python. O framework foi escolhido por permitir a fácil integração entre os classificadores da biblioteca Scikit-Learn em uma aplicação web. O sistema está disponível publicamente para acesso na URL https://fakenilcweb.herokuapp.com/.

Imagem 1. Exemplo da interface web desenvolvida

Imagem 2. Exemplo de resposta dada ao usuário após submeter uma notícia

5. Considerações Finais

Este projeto possibilitou a investigação de métodos para a detecção automática de informações falsas na web. Com base em 2 trabalhos prévios na área, foram estudados atributos variados e aplicados diversos métodos e técnicas de PLN e AM para a detecção de conteúdo enganoso em textos.

Como contribuição científica relevante deste projeto para a área de PLN, destacam-se os resultados obtidos, que podem servir como base para estudos futuros em detecção de conteúdo enganoso, e a criação do primeiro córpus público e anotado de notícias falsas em português.

Os resultados obtidos nos experimentos realizados durante o projeto foram positivos, com destaque para a acurácia de 89% obtida com o uso de *bag of words* e SVM. Ainda que sejam resultados impressionantes, o foco da abordagem foram as notícias com conteúdo falso, sendo a detecção de meias verdades, boatos e textos satíricos desafios a serem resolvidos em trabalhos futuros. Também é interessante que os modelos treinados neste projeto sejam avaliados em um córpus de validação, independente do córpus apresentado, com o intuito de estudar a generalização dos modelos desenvolvidos pela abordagem proposta.

Um artigo completo sobre as atividades realizadas durante este projeto foi submetido e aceito para publicação na décima terceira edição da *International Conference on the Computational Processing of Portuguese (PROPOR)*[3]. A conferência (de caráter internacional) é um dos principais eventos científicos na área

[3] http://www.inf.ufrgs.br/propor-2018/

de tecnologias da linguagem, sendo especificamente voltada para a língua portuguesa. Ainda, um resumo está em preparação para submissão ao *Simpósio Internacional de Iniciação Científica e Tecnológica da USP (SIICUSP)*, evento organizado pela instituição.

Mais informações sobre as ferramentas e recursos desenvolvidos neste projeto estão disponíveis no website do projeto OPINANDO[4], o qual este projeto de iniciação científica integra.

Agradecimentos

À USP, por apoiar a realização deste projeto.

Referências

[1] Bond, C. F. Jr. & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. Psychological Bulletin, Vol. 134, N. 4, pp. 477-492.
[2] Pérez-Rosas, V. & Mihalcea, R. (2015). Experiments in Open Domain Deception Detection. In the Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1120-1125.
[3] Pérez-Rosa, V. & Kleinberg, B. & Lefevre, A. & Mihalcea, R. (2017). Automatic Detection of Fake News. arXiv:1708.07104.
[4] Mitchell, T. (1997). Machine Learning. McGraw-Hill.
[5] Fonseca, E. R. and Rosa, J.L.G. (2013) Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, p. 98-107.
[6] Balage Filho, P.P.; Aluísio, S.M.; Pardo, T.A.S. (2013). An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. In the Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pp. 215-219.
[7] Zhou, L., Burgoon, J., Twitchell, D., Qin, T., Nunamaker Jr., J. (2004) A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems 20(4), 139–165
[8] Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: Three types of fakes. (2015). Proceedings of the Association for Information Science and Technology 52(1), 1–4
[9] Clem, S.: Post-truth and vices opposed to truth. (2017). Journal of the Society of Christian Ethics 37(2), 97–116
[10] Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. (1986). McGrawHill, Inc., New York, NY, USA
[11] Bick, E. The Parsing System PALAVRAS: Automatic Gramatical Analysis of Portuguese in a Constraint Grammar Framework. (2000). Aarhus University Press.

[4] https://sites.google.com/icmc.usp.br/opinando/