Resumo: “Towards automatically filtering fake news in Portuguese”

20/07/2020

Renato M. Silva, Roney L.S. Santos,Tiago A. Almeida, Thiago A.S. Pardo

Feito por Sergio Magalhães Contente

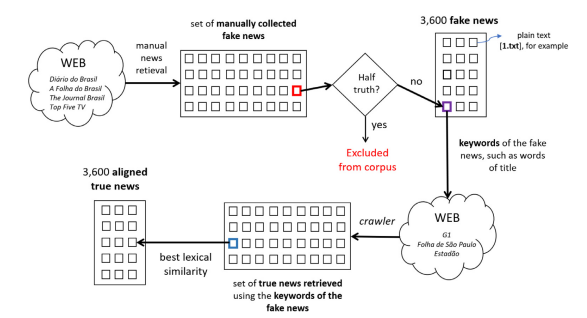
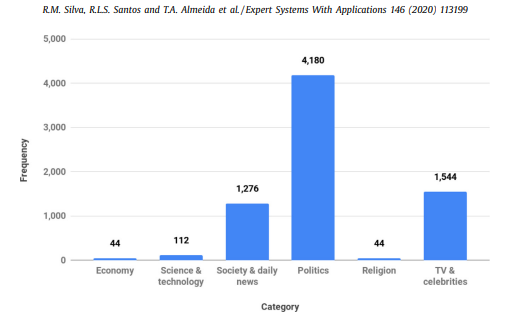
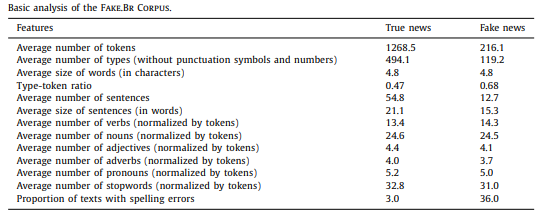
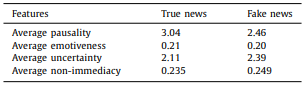
# Introdução

* 1. Começa alertando sobre o fato de atualmente notícias falsas estarem imitando o conteúdo de jornais oficiais. Diferente das notícias onde a fonte é incerta ou não há um empenho na pesquisa sobre o assunto do conteúdo, as chamadas notícias falsas abordadas no trabalho estão atreladas à ideia de enganar propositalmente.
  2. Afirma que as notícias falsas influenciam não só eleições políticas no mundo, mas também questões de saúde pública (e.g. conspirações contra a campanha de vacinação) e tragédias humanas - linchamentos públicos e violência.
  3. Afirma que existem estudos que mostram que cerca de 50% a 63% das pessoas conseguem separar notícias falsas enganadoras (deceptions).
  4. Declara duas afirmações sobre detecção automática de fake news:
     1. Esses modelos são mais objetivos que julgamentos humanos, pois esses podem ser mais enviesados que aqueles.
     2. Muita informação online pode sobrecarregar quem julga e levar a erros/delays. Assim, NLP pode detectar padrões linguísticos a fim de facilitar o julgamento. O problema é que essa abordagem é dependente de cada língua, assim surge a necessidade de criar corpora para cada cultura.
  5. Por fim, eles se propõe a responder às seguintes perguntas:
     1. Quais são os melhores métodos atuais para detectar notícias falsas automática e atualmente?
     2. Qual é a melhor feature para classificação de notícias falsas?
     3. Qual o impacto de diferentes estratégias de classificação?
     4. O tamanho do texto influencia os resultados da classificação?

# Trabalho relacionado

* 1. Nesta parte, busca-se referenciar o que já foi feito até agora para detecção de fake news.
  2. Começando com a separação de 3 tipos de deception:
     1. Notícias enganosas para propósitos humorísticos, utilizando-se de sarcasmo e ironia a fim de se produzir sátiras e paródias;
     2. Conteúdo falso com a intenção de enganar e espalhar desinformação - fake news, no geral, se encaixam nessa categoria;
     3. Informações não-validadas que são aceitas publicamente - rumores;
  3. Separação taxonômica de grupos que disseminam fake news: indicadores de linguagem verbal e não-verbal. Os verbais são divididos em baseados em linguística ou baseados em conteúdo. O primeiro está relacionado aos atributos da língua - gramática, semântica, erros ortográficos, conteúdo - e são os mais dedicados à NLP. Já o segundo é focado no que *é* de fato transmitido (e.g. fact checking). Comportamento não-verbal tem a ver com features paralinguísticas (e.g. tom de voz, pausas, traços, participação do indivíduo no discurso) ou de comunicação proxêmica (e.g. movimentos do corpo da pessoa, linguagem corporal)
  4. Para a linguística, o artigo diz que outros já encontraram features importantes no que tange à detecção de fake news:
     1. Quantidade de verbos e modificadores - adjetivos e advérbios
     2. Complexidade do texto - sua extensão
     3. Pausas - taxa de ocorrência de pontuação
     4. Incerteza - número de verbos modais e voz passiva
     5. Não-imediatismo - número de pronomes pessoais
     6. Expressividade - número de advérbios e adjetivos em relação a nomes e verbos)
     7. Diversidade
     8. Informalidade
  5. Alguns classificadores de machine learning utilizados:
     1. Bag of Words
     2. Part of Speech Tags
     3. Syntactic Information
     4. Métricas de legibilidade
     5. Classes semânticas de palavras
  6. Outros esforços para identificar fake news:
     1. Análise de contradições, mentiras, distorções, modificações na frase, superlativos, falta de informação, meias verdades, mudança de assunto, informações erradas e mal-entendidos.
     2. Identificação de notícias hiper partidárias (extremamente parciais): verbos factivos, assertivos e implicativos; pistas de subjetividade (palavras polarizadas) e pistas psicolinguísticas.
     3. Para fazer um fact-checking, normalmente, utiliza-se em estruturar os dados para depois fazer as análises
        1. Grafos de conhecimento (<https://www.ibm.com/cloud/learn/knowledge-graph>) com infoboxes da Wikipedia

# The Fake.Br Corpus

* 1. Escrito no formato de plain text a fim de facilitar a análise NLP
  2. Deve possuir versões verdadeiras e falsas sobre um mesmo tópico para encontrar padrões em instâncias positivas e negativas.
  3. Dado que a língua está em constante transformação, o córpus deve pertencer a um período de tempo específico para seu propósito
  4. Deve ter metadados (e.g. URL, autores, datas, comentários) para ser útil para os algoritmos de checagem de fatos.
  5. Para o Fake.Br, a maioria das notícias foram coletadas entre 2016 até 2018 incluindo seus metadados. A figura acima representa o esquema de construção desse córpus. Algumas características de sites que publicam notícias falsas (de acordo com a Monitor Tool of the Political Debate in Digital Media do grupo de pesquisas Research Group on Public Policies for Access to Information):
     1. Autor não é citado.
     2. Títulos são sensacionalistas - clickbait.
     3. Erros gramaticais e de concordância juntamente de adjetivos como “ladrão” e outros de forte sensibilidade.
     4. Notícias escritar com letras em caixa alta e múltiplos pontos de exclamação.
     5. Não indica quando o fato ocorreu nem contém outras fontes e referências.
     6. Não há uma página que identifique os administradores ou jornalistas responsáveis pela notícia.
     7. Site possui um layout confuso e poluído, camuflando-se como um site de notícias confiável.
  6. Após isso, fizeram uma busca manual para encontrar tais sites e encontraram 4: “Diário do Brasil”, “Folha do Brasil”, “Jornal do Brasil” e “Top Five TV”. E checaram manualmente a veracidade das notícias para evitar a coleta de meias verdades. Para tal, utilizaram os portais online de notícias: “Agência Lupa”, “Fato ou Fake”, “Aos Fatos” e “Boatos.org”.
  7. Após isso, fizeram um webcrawler para coletar notícias verdadeiras de agências confiáveis como “G1”, “Folha de São Paulo” e “Estadão”. Para isso, utilizaram as keywords extraídas das notícias falsas, como nomes e verbos dos títulos, palavras mais frequentes com exceção de stopwords.
  8. Para fazer a correspondência da notícia verdadeira com a falsa, fizeram uso da medida de similaridade léxica cossenóide.
     1. 
     2. O vetor f é a notícia falsa e o v é a verdadeira. Utilizou-se uma representação vetorial das palavras (e.g. Bag of Words) . Quanto mais próximo de 0, mais diferente os vetores são entre si.
  9. Por fim, separaram em categorias para cada tipo de notícia: Ciência, esportes, tecnologia, lazer, etc.
     1. 
  10. Então fizeram a análise básica do Corpus obtido utilizando features de NLP tradicionais:
      1. 
  11. Observa-se que como as notícias verdadeiras possuem mais tokens que as falsas, isso pode dificultar a análise por machine learning por conta de enviesamento da classificação.
  12. Por fim, fizeram outra análise de features:
      1. 
  13. Para tal, a pausa foi dada pelo número de marcas de pontuação dividido pelo número de sentenças. A emotividade foi dada pela soma de adjetivos e advérbios dividido pela soma de nomes e verbos. A incerteza é baseada na ocorrência de verbos modais e voz passiva. Por fim, a não imediaticidade é baseada na frequência das 1as e 2as pessoas.

# Experimentos

* 1. Utilizando as features das tabelas acima, foram feitos alguns experimentos.
  2. Cada amostra foi representada por 3 jeitos diferentes: Bag of Words (BoW), Word2Vec e FastText. Para a BoW foi usada a TF-IDF codificação para ajustar os pesos de tokens de cada documento. Para o Word2Vec e FastText utilizou-se vetores de 300 dimensões treinados com aproximação Skip-Gram. Para cada documento do corpus, obtiveram-se vetores pré-treinados para cada palavra e então computaram as médias.

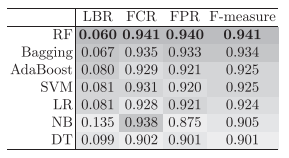
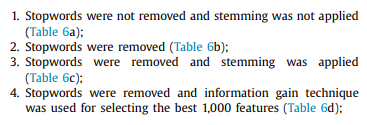
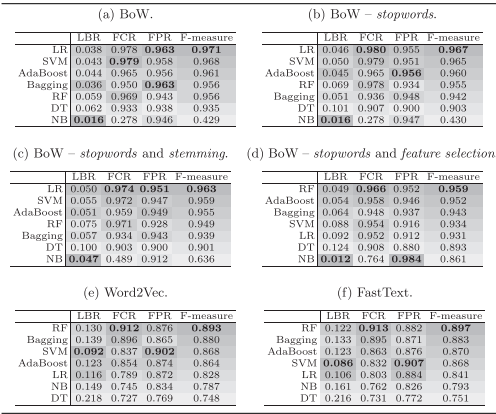
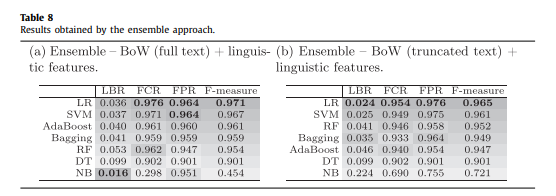
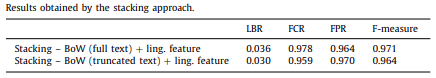
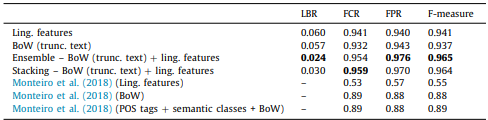
# Pré-processamento

* 1. Aplicou-se a normalização Z-score com a informação dos exemplos de treinamento. Além disso, todas as instâncias foram colocadas em caixa baixa e as dummy features como numerais, URLs e emails foram convertidas em ‘0’, ‘URL’ e ‘EMAIL’. Depois foram tokenizadas com base em espaços e marcas de pontuação.

# Métodos

* 1. Classificação:
     1. Regressão Logística (LR)
     2. Support Vector Machine (SVM)
     3. Decision Trees (DT)
     4. Random Forest (RF)
     5. Bootstrap aggregating (bagging)
     6. Adaptive Boosting (AdaBoost)
  2. Medidas de performance:
     1. Legitimate News Blocked Rate (LBR): equivalente ao falso positivo - notícias legítimas catalogadas como falsas.
     2. Fake News Caught Rate (FCR): equivalente ao recall - notícias falsas identificadas corretamente
     3. Fake News Precision Rate (FPR): Proporção de notícias classificadas como falsas e que realmente pertencem a classe de notícias falsas.
     4. F-measure: média harmônica do FCR e FPR.

# Resultados

* 1. 
  2. Esses resultados mostram que a RF foi a melhor.
  3. Para resultados obtidos com features geradas pelas técnicas de representação de texto (BoW, Word2Vec e FastText):
     1. 
     2. 
     3. Os resultados mostraram que remover stopwords e aplicar stemming não melhorou a performance dos métodos de classificação. Isso evidencia que o pré-processamento pode remover features importantes para a classificação de fake news. Feature selection também não foi efetiva como em (d).
     4. Os resultados com BoW foram obtidos com a Word2Vec e FastText. Isso porque esses modelos são usados para gerar vetores de palavras que foram treinados com documentos da Wikipedia, Google News, que são bem escritos. Assim, esses vetores não são adequados para analisar fake news. Provavelmente, se tivessem sido treinados com documentos mais poluídos (noisy), o resultado seria diferente.
     5. Além disso, os resultados da BoW foram melhores que os da tabela de features linguísticas (item a)
  4. Conjunto de previsões usando diferentes sets de features
     1. 
     2. A melhor aproximação foi aquela que combinou previsões obtidas com a LR. Na seção eles propõem uma aproximação stacking que usa um meta-classificador treinado com probabilidades dadas por 2 classificadores individuais. O primeiro é a LR treinada com features linguísticas e o segundo é a LR treinada com a BoW.
     3. 
     4. Por fim, eles compararam os resultados do estudo com o de um anterior (Monteiro et al. (2018)). Nessa comparação, eles mostram que resultados da literatura obtidos no Fake.Br corpus são inferiores aos do estudo. Porém, as features linguísticas extraídas desse corpus performaram muito pobremente no estudo do Monteiro et al. (2018), dando um F-measure de 0.55
     5. 
     6. Isso ocorre porque:
        1. Monteiro et al. (2018) possui menas features que a do estudo
        2. Monteiro et al. (2018) não possuia features normalizadas, mas o estudo utilizou a normalização Z-score.
        3. Monteiro et al. (2018) não performou grid-search para encontrar o melhor regularização do parâmetro de SVM.

# Conclusões

* 1. Q1: Os melhores métodos são LR, SVM e RF.
  2. Q2: As melhores features foram as geradas pelo BoW em detrimento das linguísticas, Word2Vec e FastText.
  3. Q3: A aproximação combinada e com stacking foi a que apresentou melhores resultados
  4. Q4: Os resultados com textos inteiros (e não truncados) foram melhores. Mas, acredita-se que exista algum viés no dataset, isso significa que provavelmente os resultados com os textos truncados possam ser melhores para aplicações no mundo real.