Resumo: “Automatic Detection of Fake News”

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea

23/08/17

Feito por Sergio Magalhães Contente

# Introdução

* 1. Começa falando que a Jumpshot Tech Blog contou que o Facebook era responsável por 50% do tráfego de sites de fake news e 20% do tráfego de sites com boa reputação. A maioria dos adultos americanos (62%) recebem notícias por redes sociais.
  2. Até a época desse estudo, aproximações de fake news eram apenas baseadas em notícias satíricas e fact-checking. Entretanto, usar notícias satíricas para análise de fake news pode gerar fatores conflitantes como humor e absurdidade.
  3. No estudo foram criados 2 datasets: um deles feito por crowdsourcing misturada com anotação manual enquanto a outra foi feita coletando dados diretamente da web.
  4. Depois o modelo foi comparado com pessoas em si fazendo fact-checking.

# Trabalho relacionado

* 1. Determina que existem 3 tipos de fake news:
     1. Humorísticos com propósitos de serem paródias e sátiras;
     2. Conteúdo falso, com a intenção de enganar de fato;
     3. Rumores, ou seja, notícias não-confirmadas, mas aceitas publicamente.
  2. Utiliza de um estudo de sátiras para afirmar que os melhores classificadores foram feitos utilizando features de absurdidade, pontuação e gramática.
  3. Diz sobre uma aproximação estilométrica (estilo de escrita) para identificar notícias falsas e verdadeiras
  4. Também refere que outros trabalhos utilizou-se stop words, n-gramas, indícios de legibilidade, número médio de palavras por parágrafo.

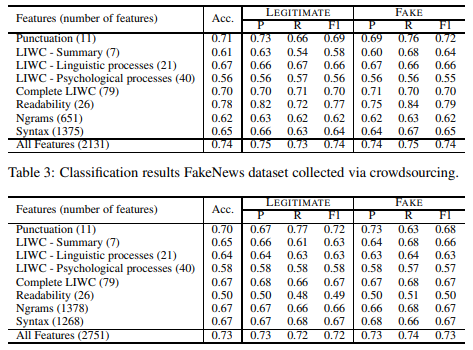
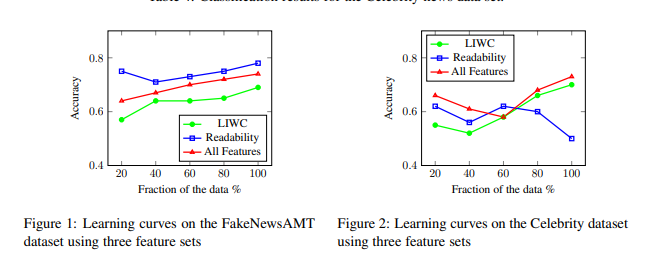
# Fake News Datasets

* 1. Até o momento desse estudo, existia apenas dados vindo de fontes satíricas como (“The Onion”) que geralmente está atrelado a ironia e humor.
  2. Necessidade de um dataset para fake news:
     1. O corpus deve conter tanto notícias verdadeiras quanto falsas
     2. Deve conter apenas notícias no formato textual
     3. Deve haver um chão de veracidade verificável
     4. Deve ser homogêneo em tamanho
     5. Estilo de escrita
     6. Contém notícias de um tempo pré-definido
     7. Deve ser entregue da mesma maneira e pelo mesmo propósito para casos reais ou falsos
     8. Deve ser publicamente disponível
     9. Deve levar em consideração diferenças culturais e de línguas
  3. Coletando notícias verdadeiras:
     1. Inicialmente coletaram dataset de domínios confiáveis (e.g. ABCNews, CNN, USAToday, NewYorkTimes, FoxNews, Bloomberg, CNET)
     2. Para garantir a veracidade, manualmente conduziram um fact-checking no conteúdo - verificando a fonte e fazendo referência-cruzada para diversas fontes.
  4. Coletando notícias falsas:
     1. Fizeram um crowdsourcing via Amazon Mechanical Turk (AMT) o qual foi utilizado com sucesso no passado para coletar dados enganosos, incluindo reviews de opiniões e tópicos controversos (e.g. aborto, sentença de morte)
     2. O problema de utilizar o AMT:
        1. Língua das notícias usadas pelos jornalistas podem ser diferentes da língua dos trabalhadores do AMT.
        2. Notícias são geralmente mais longas que reviews de consumidores, o que aumenta o trabalho dos trabalhadores do AMT.
  5. Eles então restringiram para apenas notícias estadunidenses e instruíram os trabalhadores da AMT para fazer o dataset normalmente.
  6. Eles conseguiram coletar 240 fake news - cada um consistindo de 132 palavras aproximadamente e 5 sentenças.
  7. O nome ficou de FakeNewsAMT
  8. Para o WebDataset, o segundo, o alvo era mais o conteúdo falso que aparecia naturalmente na web. Coletaram principalmente sobre figuras públicas - celebridades, já que possuem a maior tendência de serem fake news. Esse dataset é chamado de Celebrity

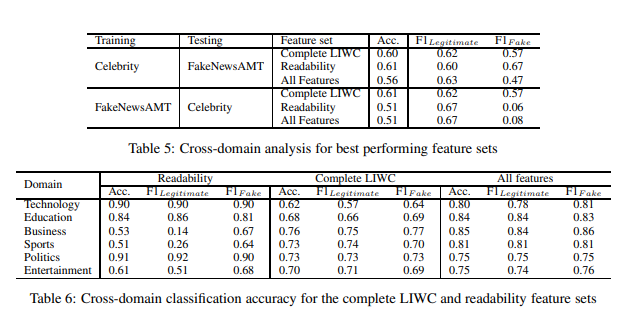
# Features linguísticas

* 1. Ngramas: extraíram unigramas e bigramas derivados da bag of words. Codificados como TF-IDF para normalizar os pesos de cada palavra (tamanho).
  2. Pontuação: 11 tipos de pontuação foram levados em conta derivadas da LIWC. Isso inclui pontos finais, vírgulas, traços, pontos de exclamação e interrogação.
  3. Psicolinguísticas: Usando o lexicon LIWC para extrair as proporções das palavras que caem na categoria psicolinguística. Por exemplo, processos psicolinguísticos (e.g. emoções positivas, percepções), categorias sumarizadores (e.g. palavras por sentença), categorias part-of-speech (e.g. artigos, verbos). É útil para analisar contextos diversos que podem ser alvos de falsidades. No trabalho, eles usaram apenas 3 categorias para tal:
     1. Categorias summary: pensamento analítico, tom emocional
     2. Processos linguísticos: funções das palavras, pronomes
     3. Processos psicolinguísticos: afetividade, social
  4. Legibilidade: Número de caracteres, palavras complexas, tamanho das palavras, número de sílabas, tipo de palavras, número de parágrafos. Métricas: Flesch-Kincaid, Flesch Reading Ease, Gunning Fog, and the Automatic Readability Index (ARI)
  5. Sintaxe: Extraíram um set de features derivadas de regras de produção baseadas em árvores CFG (Context-Free-Grammar). Também codificadas em TF-IDF

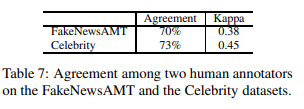
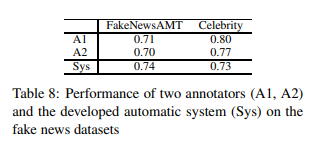
# Modelo computacional para detecção de Fake News

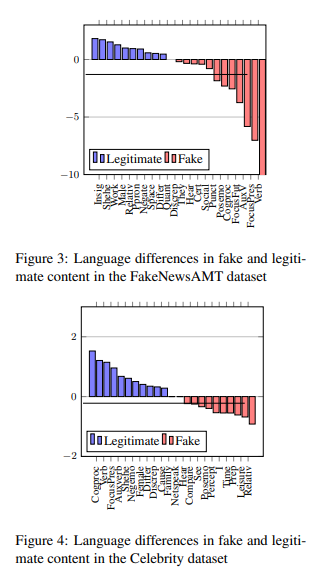
* 1. 
  2. Utilizou-se um classificador SVM e validação cruzada 5-fold com acurácia, precisão, recall e F1 sendo as médias de 5 iterações.
  3. Curvas de treinamento: investigou se largas quantidades de dados de treinamenton podem melhorar a identificação de conteúdos falsos.
  4. 
  5. Com exceção da legibilidade, as outras features apresentaram um aumento regular, ou seja, quanto maior os dados de treino, melhor.

# Análise cross-domain

* 1. Aplicabilidade do método entre os domínios utilizando os 2 melhores sets de features (Legibilidade - “Readability” - e LIWC) junto com o classificador que utiliza todas as features
  2. 
  3. Os resultados mostram que há uma perda significativa de acurácia quando não há uma separação entre domínios. Explicações possíveis:
     1. Propriedades linguísticas de deceptions em um domínio é estruturalmente diferente de outro
     2. Uma feature aplicada em um domínio não estava performando bem no respectivo domínio em primeiro lugar
  4. Então, testou-se a melhor feature de cada domínio para domínios opostos e constatou que a melhor explicação é a primeira.

# Performance humana:

* 1. Criaram uma interface de anotação onde as pessoas indicavam se uma notícia era verdadeira ou não. Eles também perguntaram se os anotadores haviam lido ou escutado sobre determinado tópico recentemente - menos de 5% leram anteriormente as notícias - isso diminui o viés da análise.
  2. Cada um analisou os 2 datasets (FakeNewsAMT e Celebrity) e não houve oferta monetária para evitar enviesamento.
  3. 
  4. Em adição, comparou-se com o classificador automático e obteve-se esse resultado:
  5. 
  6. Isso mostra que humanos detectam notícias falsas no domínio de celebridades melhor que o modelo, porém em contextos sérios e políticos o sistema é melhor.

1. Insights posteriores
   1. Observaram que classificadores que utilizavam a informação semântica codificada no lexicon do LIWC possuíam boa performance através dos domínios.
   2. Avaliaram quais classes do LIWC mostravam diferenças significativas entre notícias falsas e legítimas.
   3. Um valor positivo indica uma associação entre a classe do LIWC e um conteúdo legítimo e um negativo indica uma associação entre uma classe do LIWC e um conteúdo falso.
   4. 
   5. Resultando que palavras usadas em conteúdos verdadeiros são associadas com processos cognitivos e também possuem mais palavras de função como “he”, “she”, negações e também expressa relatividade. Já para as fakes, usa-se mais advérbios, verbos e pontuação.