Uma abordagem na classificação de tweets de domínio questionável com base em técnicas de Text Mining

Tópicos Avançados em Ciência de Dados [CC4061] - 2021/2022 – 2S

Grupo 7

Amanda Tavares
202103516

Arina Sanches
202100371

Lirielly Nascimento
202100370

Marta Barbosa
199201533

I. Resumo

Neste trabalho apresenta-se uma abordagem à construção de um sistema de classificação que permita detetar mensagens com informação considerada questionável na rede social Twitter, partindo, essencialmente da análise do conteúdo textual no feed dos utilizadores, recorrendo a técnicas de processamento de linguagem e análise de conteúdo para a extração de atributos que otimizem o modelo de classificação.

II. Introdução

Avaliar a credibilidade de um conteúdo publicado e partilhado numa rede social como o Twitter continua a ser um desafio de crescente importância da Internet. O possível anonimato dos utilizadores, combinado com interfaces digitais de serviços online, possibilita que programas de computador gerem automaticamente mensagens similares às realizadas por utilizadores humanos [1]. A disseminação de informação enganosa, na forma de notícias falsas, com intuito de manipular a opinião pública, muitas vezes com fins políticos é, em parte, um dos objetivos deste tipo de programas, também denominados por *bots*. A maior parte das técnicas de deteção de bots numa rede social como o Twitter, recorrendo a Machine Learning, apoia-se, essencialmente na análise do tipo de interação social que estas contas têm, observando-se o número de seguidores, número de contas que seguem, frequência das mensagens e algumas particularidades das mesmas como a ocorrência de menções, *hashtags* e *retweets* [2], nem sempre se considerando aspetos no próprio texto que poderão ser relevantes para classificar o tipo de informação que está a ser disseminada como questionável ou não. É nesta linha que o presente trabalho foi desenvolvido, aplicando técnicas de mineração de texto e análise de sentimento, exploradas ao longo das aulas da UC de Tópicos Avançados em Ciência de Dados, na qual foi proposto o desenvolvimento do presente projeto para a construção e otimização de um modelo de classificação, com base em técnicas de processamento da linguagem natural (NLP) e técnicas de análise de conteúdo [1], que consiga reconhecer *tweets* oriundos de uma fonte questionável.

Para o efeito, foi fornecido um *dataset* com 17.950 *tweets* relativos às últimas eleições presidenciais norte americanas, realizadas em 2020. Os *tweets* apresentam-se já classificados através da variável "questionable domain", separando-se em "TRUE" ou "FALSE", sendo esta desproporcional, verificando-se o valor "TRUE" em apenas 16.71% dos casos. O *dataset* era composto por onze variáveis: "id", "user friends count", "user followers count", "user favourites count", "description", "title", "favorite count", "retweet count", "user verified" e "contains profanity", sendo as duas últimas variáveis booleanas. O ponto de partida deste trabalho foi a análise exploratória de dados, onde se procurou compreender a distribuição das variáveis e utilizar técnicas de análise textual de forma a encontrar novas *features* de forma a construir um modelo que conseguisse classificar se um *tweet* não visto anteriormente está a disseminar informação questionável ou não. O modelo foi sendo sucessivamente aperfeiçoado, através da adição ou remoção de *features* e com a utilização de técnicas de balanceamento do conjunto de dados.

III. Análise Exploratória de Dados

Num primeiro momento, foi feito um pré processamento dos dados onde se verificou que a variável "id" não agregava valor ao modelo de classificação, sendo, por isso, descartada; a variável "title" continha a mesma informação dada pela variável "description" sendo igualmente descartada; não foram observados *missing values* no *dataset*.

Foram criadas as variáveis booleanas "mentions", "hashtag" e "URL" para identificar a sua existência nos *tweets*. Foi também feita uma contagem por *tweet* de cada um destes elementos. Mais tarde foi descartada a variável "URL" por não acrescentar nenhum elemento diferenciador, dado que todos os *tweets* apresentam exatamente uma hiperligação.

Para melhor entendimento dos dados, foram analisadas todas as distribuições das diferentes variáveis de modo a

identificar diferenças significativas que separassem os *tweets* em relação à variável "questionable domain".

A matriz de correlação indicou que o único par de variáveis com alta correlação (0.82) era *retweet count* e *favorite count*. Optou-se, assim, por manter a variável *retweet count*, uma vez que seria redundante manter as duas, em termos de contribuição para o modelo.

Foi realizada uma análise de co-ocorrências entre as *hashtags* de maior frequência para *tweets* de domínio questionável e não questionável, sendo possível verificar uma tendência dos assuntos de cada grupo, entretanto não foi obtida uma quantidade significativa de observações entre os pares. O mesmo foi feito para as menções e foram obtidos resultados semelhantes.

Para a análise do texto, procedeu-se à construção do *corpus* a partir da variável "description", que agrega o conteúdo de texto de cada *tweet*. Dado que, nos *tweets*, há muito ruído e como modelos pré-treinados podem se beneficiar da preparação de dados [3], procedeu-se a uma preparação cuidada para respetiva mineração do texto para a possível obtenção de *features* adicionais às já existentes no *dataset* que permitissem a otimização de resultados por parte de um modelo de classificação. Procedeu-se, assim, à remoção de pontuação, símbolos, números e urls, assim como *stopwords*, desnecessárias para análise em questão. Refira-se, no entanto que, antes deste procedimento, fez-se uma contagem do número de pontos de exclamação por *tweet*, considerando-se posteriormente como uma *feature* a considerar pelo modelo. Foi também realizado o mesmo procedimento relativamente ao número de letras maiúsculas por *tweet*.

Numa primeira análise do texto, foram construídas as respetivas *wordclouds*, de acordo com a variável questionable domain, onde foi possível detetar diferenças significativas, embora ambas dando ênfase à palavra "trump", mostrando que, independentemente de os *tweets* serem ou não de domínio questionável, o candidato presidencial Donald Trump é o principal elemento presente nos *tweets*, em geral. No entanto, na *wordcloud* para domínio questionável "False" observam-se termos mais relacionados com os candidatos e a eleição em si, nomeadamente as menções @realdonaldtrump e @gop, uma menção ao partido republicano. Já na *wordcloud* relativa a *tweets* de domínio questionável, o principal destaque vai para o primeiro nome dos dois candidatos às eleições presidenciais, "joe" e "trump", e para a menção "@gatewaypundit", que se refere ao site norte americano *Gateway Pundit*, conhecido por divulgar notícias falsas, de extrema direita, publicar boatos e teorias da conspiração, o que não deixa de ser uma primeira abordagem promissora (Fig.1).

Procedeu-se a uma análise mais detalhada em termos da frequência de palavras e associações entre as mesmas que se evidenciassem. Concluiu-se que, embora a palavra *trump* não seja diferenciadora, em termos de domínio questionável, o mesmo já não acontece com "via" e a menção @gatewaypundit. As associações foram estudadas através da construção de redes de acordo com cada domínio e a análise de n-gramas, em particular, bi-gramas, que estudam a frequência da associação

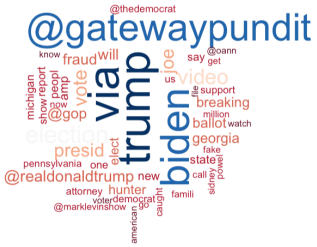

Fig. 1: Wordcloud para *tweets* de "Questionable Domain"

de duas palavras observadas no *corpus* (Fig. 2). Concluiu-se com mais expressividade a associação "via gatewaypundit" em *tweets* de domínio questionável, como uma possível *feature* diferenciadora dos dois grupos de *tweets*, sendo, assim, acrescentada ao *dataset* original como uma variável booleana. Ao longo das várias experiências com diferentes modelos, foram acrescentados mais bi-gramas e palavras que se traduziram em melhorias assinaláveis na performance do modelo em classificar *tweets* de fonte questionável.

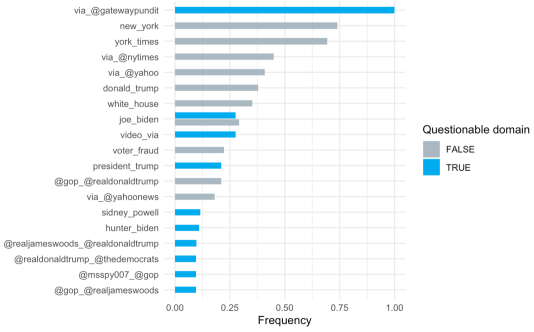

Fig. 2: Bi-gramas

Foram ainda exploradas justaposições de palavras relativamente à menção @gatewaypundit no *corpus* relativo ao *tweets* domínio questionável (Fig.3). Desta análise construiu-se um pequeno dicionário de *collocations*, denominado por *bag of words* [4], considerando-se as seis mais fortes ("via", "video", "georgia", "breaking", "@gop", "@realdonaldtrump"), e, em cada *tweet*, atribuiu-se a frequência total dos termos encontrados nesse mesmo dicionário, sendo adicionada como uma *feature* a considerar pelo modelo.

Na análise mais pormenorizada dos *tweets*, foram detetadas repetições significativas nos *tweets* de domínio questionável, com uma distribuição bastante diferenciada, quando analisados os dois grupos. No entanto, embora seja uma característica específica dos *bots*, a sua consideração não revelou ser significativa na implementação do modelo, tendo sido abandonada esta exploração. Foram realizadas análises de sentimento recorrendo a vários léxicos, mas que nem sempre se revelaram suficientemente expressivas em termos de resultados a utilizar de forma eficaz pelo modelo. Da análise de sentimento, recorrendo ao léxico proposto por Loughran

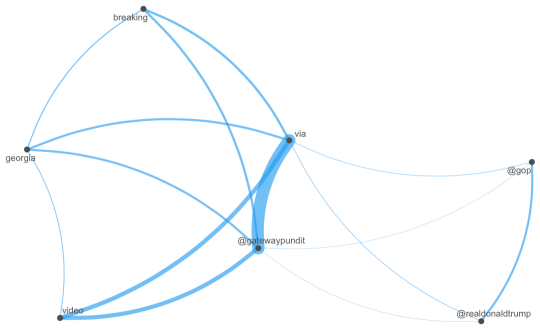

Fig. 3: Grafo de associações com @gatewaypundit

foi possível identificar a presença das palavras *"exclusive"* e *"breaking"* como as que figuravam em primeiro lugar no top 10 dos sentimentos positivos e negativos, respetivamente, no caso dos *tweets* de domínio questionável. Estas features foram também consideradas ao longo deste processo, sendo que breaking já tinha sido detetado anteriormente.

Entendeu-se ser ainda pertinente realizar uma análise de sentimento por *tweet*. Para conseguir resultados mais específicos, realizou-se uma análise frase a frase, procurando-se as emoções associadas, que foram posteriormente adicionadas como uma *feature* ao *dataset*. De seguida foi aplicada uma técnica de *Association Rules* com o objetivo de inferir quais conjunto de emoções estariam associados aos respetivos *questionable domain*.

A otimização do modelo desenrolou-se ao longo de várias etapas de processamento da linguagem natural presente nos *tweets*, algumas revelando-se infrutíferas e outras mais promissoras. Foram observadas diferenças significativas no *corpus* na análise do número de frases e número de palavras, *tokens*, verificando-se que, no caso dos *tweets* de domínio questionável, quase 50% é composto apenas por uma frase, sendo que nos de domínio confiável, mais de 50% apresentam entre duas e três frases. No número de palavras dos *tweets* de domínio questionável, há também diferenças assinaláveis, apresentado a distribuição um enviesamento à direita, indicando que a mediana é inferior à média de palavras. No caso do domínio confiável, há uma maior diversidade no número de palavras por *tweet*, resultando numa distribuição não tão enviesada. Foram, assim, incluídas como *features* a considerar pelo modelo de classificação o número de frases e o número de palavras por *tweet*.

IV. Modelos de Classificação

Os modelos escolhidos foram *Logistic Regression*, *kNN (k = 11)*, *Naive Bayes* e *Boost Trees* (*XGBoost*). As métricas utilizadas para avaliar os modelos foram *Precision, Recall e F1*. Incialmente, o *dataset* foi dividido em treino, com 80% dos dados e teste, com os restantes 20%. A seguir, foram utilizadas outras técnicas para tentar minimizar o impacto de ter dados não balanceados.

Foi gerado um *Baseline Model* para ser utilizado como ponto de partida e referência para análise de resultados futuros. Este modelo foi gerado utilizando as variáveis explicativas que já estavam presentes, inicialmente, no conjunto de dados e com as variáveis booleanas que indicavam a presença de menções, URLS e *hashtags*. A variável "questionable domain", que divide os *tweets* serem oriundos de fonte questionáveis ou não, é a variável dependente do modelo. Dentre os modelos de *machine learning* utilizados, o que obteve melhor performance foi o *kNN* com *F1* igual a 0.35.

Com base nos resultados da análise exploratória dos dados, foram sendo sucessivamente adicionadas novas *features* ao modelo, de forma a melhorar a sua capacidade de classificação. Inicialmente foram adicionadas duas novas *features* ao modelo, referentes as contagens do número de menções, *hashtags* e prosseguiu-se com o treino e teste. De seguida, com base na análise de frequência, foram incluídas as *features*: "@gatewaypundit", "@gop", "@realdonaldtrump", "trump" e "amp", cada um delas indicando a presença ou ausência destas palavras na descrição dos *tweets*. Com estas iterações não foram observadas melhorias significativas nas métricas.

Posteriormente foram incluídos os bi-gramas mais relevantes, já mencionados anteriormente na análise exploratória, assim como "exclusive" e "break", que resultaram da análise de sentimento. A adição destas *features*, resultou num aumento no valor das métricas para os modelos testados. Dentre estes, o *XGBoost* obteve melhor resultado, com o *F1* igual a 0.58.

Sabendo que o *dataset* tem 16.71% dos *tweets* classificados como "TRUE" e 83.29% classificados como "FALSE", em termos de questionable domain, foram utilizadas técnicas para balancear a amostra, a fim de melhorar os modelos. Uma delas foi o algoritmo Rose (Random OverSampling Examples), que consiste em criar uma amostra de dados sintéticos, ampliando o espaço de recursos de exemplos de classes minoritárias e maioritárias. Operacionalmente, os novos exemplos são extraídos de uma estimativa condicional de densidade do *kernel* das duas classes, como descrito em Menardi e Torelli (2013) [5]. O modelo *Naive Bayes* foi o que conseguiu o melhor resultado, tendo *F1* igual a 0.56. No entanto, o modelo anterior conseguiu resultados melhores, pelo que foi decidido não continuar com esta técnica.

Foi também aplicada a técnica de *random under-sampling*, sendo que, nesta abordagem, todas as observações da classe minoritária são mantidas no conjunto de dados, enquanto que observações da classe maioritária são removidas aleatoriamente, de forma a reduzir o tamanho da mesma. De seguida, foi aplicada a técnica de *random over and under sampling*, onde são aleatoriamente duplicadas as amostras da classe minoritária e removidas amostras da classe maioritária. As abordagens acima referidas, não melhoraram a performance dos modelos, verificando-se que os valores de F1 diminuíram para todos os modelos testados em relação aos modelos obtidos anteriormente. Por fim, foi ainda aplicada a técnica de *Under and Over sampling with Smote* que resultou numa melhoria significativa da performance dos modelos, destacando-se o *kNN* com *F1* igual a 0.70. Obteve-se assim um aumento de 0.12 no *F1*, em relação ao modelo que tinha a melhor performance até agora.

Seguidamente, foi avaliado o impacto da inclusão das *features* referentes às palavras associadas com as palavras "trump", "gop", "realdonaldtrump" e "amp" aos modelos, utilizando os dados não balanceados. O *XGBoost* continuou sendo o melhor modelo e obteve-se um ligeiro aumento no valor do *F1*, que passou a ser 0.62. De seguida, foi adicionado ao modelo, a informação relativa a proporção de letras maiúsculas em cada um dos *tweets*. Mais uma vez foi observado um aumento no valor do *F1*, tornando-se igual a 0.67. Com a inclusão das *features* que indicam a presença ou ausências de combinações de emoções, descobertas na análise exploratória, assim como a contagem de sentimentos por *tweet*, não houve mudanças significativas na performance dos modelos. Numa tentativa de melhorar o desempenho do modelo, o *Under and Over sampling with Smote* foi aplicado aos dados. Esta técnica foi a escolhida por ser a que produziu melhores resultados, comparativamente com as demais já testadas. O melhor modelo foi o *kNN*, com *F1* igual a 0.78. Neste ponto da análise, estavam sendo consideradas 40 *features*. Foi então considerada a relevância de cada uma delas para o modelo. Dentre as 40 *features*, apenas 17 foram escolhidas, mantendo-se das features adicionadas ao dataset a proporção de palavras contendo apenas letras maiúsculas, menções, sentimentos identificados, presença de *hashtag*, menções de palavras palavras relevantes e bi-gramas. Utilizando o *kNN*, o valor do *F1* ficou igual a 0.79 (modelo KNN_V14), pelo que concluímos que a remoção de um número significativo de *features* não prejudicou a performance do modelo.

Nas iterações seguintes, foram utilizados os dados não balanceados e as 17 *features* selecionadas anteriormente. Sucessivamente foram adicionadas novas *features* e avaliado o impacto nos novos modelo gerados. Foram adicionadas *features* que indicam: a ausência de duplicados, o bag of words, a contagem do números de pontos de exclamação, o número de palavras que começam com letras maiúsculas, a quantidade de algarismos, o número de *tokens* e o número de frases por tweet. Com a introdução destas *features* o *F1* passou a ser 0.67, tanto para o *kNN*, como para o *XGBoost*.

Numa fase final do trabalho, foi feita uma nova revisão das *features* a atribuir ao modelo, concluindo-se que incluindo todas as que tinham sido encontradas ao longo do trabalho e considerando os dados balanceados, se obtinha o maior valor de *F1*, 0.84 (modelo KNN_V20). No entanto, foi ainda realizada uma experiência após uma nova seleção das *features*, concluindo-se a redução das mesmas em mais de 50% não alterou significativamente o valor de *F1* (modelo KNN_V21). No Quadro 1 apresenta-se os 3 melhores modelos obtidos no estudo com os resultados das métricas analisadas: *F1, Precision e Recall*.

Quadro 1 - Comparativo das características dos trabalhos

Modelo	**KNN_V14**	**KNN_V20**	**KNN_V21**
Precision	0.80	0.90	0.86
Recall	0.78	0.78	0.75
F1	0.79	0.84	0.80

Já na Fig. 4 exibe-se os resultados dos melhores modelos encontrados durante as experimentações, ordenados pelo maior valor da métrica F1.

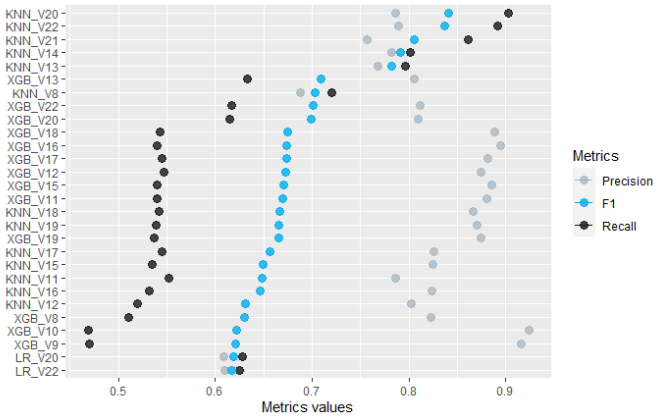

Fig. 4: Evolução das métricas.

V. Conclusão

Ao longo do trabalho desenvolvido foi possível concluir que a extração de *features* dos *tweets*, com base apenas na mineração de texto não é um processo fácil, dado a pouca extensão de texto, muitas vezes com difícil análise de sentimento. No melhoramento dos modelos foram, muitas vezes, abandonadas as estratégias correntes e delineadas novas que permitissem uma classificação mais otimizada. No entanto, todo este processo revelou-se satisfatório uma que resultou um modelo com um nível de Recall, Precision e F1 de, respetivamente, 0.8, 0.86 e 0.75. Embora não seja o modelo que apresenta melhores resultados, observados no (*Quadro 1 e Figura 4*), é aquele que considera menos *features* sem perder performance. Como uma futura abordagem, seria interessante testar a metodologia desenvolvida neste projeto num dataset maior, onde não fosse necessário realizar técnicas de balanceamento de dados e assim testar se há uma melhoria nas métricas de performance analisadas. Fica, no entanto, claro que, este tipo de classificação vai muito além do estudo das interações sociais do utilizador desta rede social, de forma a catalogá-lo como sendo de um domínio questionável, mas passa também por um problema de processamento da própria linguagem, procurando características diferenciadoras.

References

[1] Przybyła, P., "Detecting Bot Accounts on Twitter by Measuring Message Predictability" (2019)

[2] Romo, J. M., Araujo, L. , "Detecting malicious tweets in trending topics using a statistical analysis of language" (2012)

[3] SreeJagadeesh Malla, P. J. A. Alphonse , "Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news" (2022)

[4] Bernardes, V., "Linguistic and Emotion-based Identification of Tweets with Fake News: A Case Study" (2021)

[5] Lunardon, N., G. Menardi, and Nicola Torelli. "R package'ROSE': Random Over-Sampling Examples." (2013).