Paula Amaral
Universidade Nova de Lisboa
pt-maths-in@spm.pt

O ADMIRÁVEL MUNDO NOVO DO BIG DATA

"Big Data research is bound to become more widespread, and this will require more awareness on the part of data scientists, policymakers and a wider public about its contexts and often unintended consequences."
Ralph Schroeder

O parágrafo em destaque neste artigo lança um repto interessante que vai exigir uma reflexão profunda por parte dos cientistas na atual revolução digital, no admirável mundo novo do Big Data. É em torno de um problema muito mediático, o da análise de dados das redes sociais, que iremos focar-nos neste artigo, procurando resumir algumas das ideias apresentadas por Stefano Iacus, professor de Estatística e diretor do Laboratório de "Data Science", da Universidade de Milão, na conferência de apresentação da PT-MATHS-IN, que decorreu no ISEP – Porto, no dia 2 de junho. Aliando à sua atividade científica uma experiência empresarial, como cofundador da Spin-off, "Voices of the Blog", este orador apresentou parte do trabalho desenvolvido nos últimos anos sobre o tema da análise de textos das redes sociais para a previsão de resultados eleitorais. O parágrafo introdutório da página do "Voices of the Blog" (*http://www.voices-int.com*) é muito sintomático daquilo que o Big Data representa nas ciências sociais: "Quando milhões de pessoas se tornam utilizadores de plataformas da web e os seus debates e afirmações são transformados em dados, integrá-los e interpretá-los é o desafio – Big Data – que o mundo enfrenta hoje na política, na economia e na sociedade." Este é também o objeto de estudo de uma área que é designada por "Sentiment Analysis", também referida por alguns autores como "Opinion Mining" e "Emotion AI". No que se segue, adotaremos a designação de Análise de Sentimento e Opinião (ASO).

A ANÁLISE DE SENTIMENTO E OPINIÃO

A ASO consiste na análise automática de textos para determinar a polaridades das opiniões de uma determinada população de utilizadores, relativamente a um determinado assunto. São exemplo as intenções de voto, a satisfação relativa a produtos comerciais ou o sentimento associado a uma questão moral ou ética. Os textos podem ser obtidos, por exemplo, nas redes sociais como o Twitter, o Instagram e o Facebook. A massificação do uso da Internet e, por consequência, das redes sociais, disponibilizou um volume de informação impensável há umas décadas. No sítio "Internet live stats" (*http://www.internetlivestats.com*) podemos obter uma ideia concreta do volume de utilização da Internet atualizada em tempo real, como a figura 1 sugere.

Este volume de informação representa um manan-

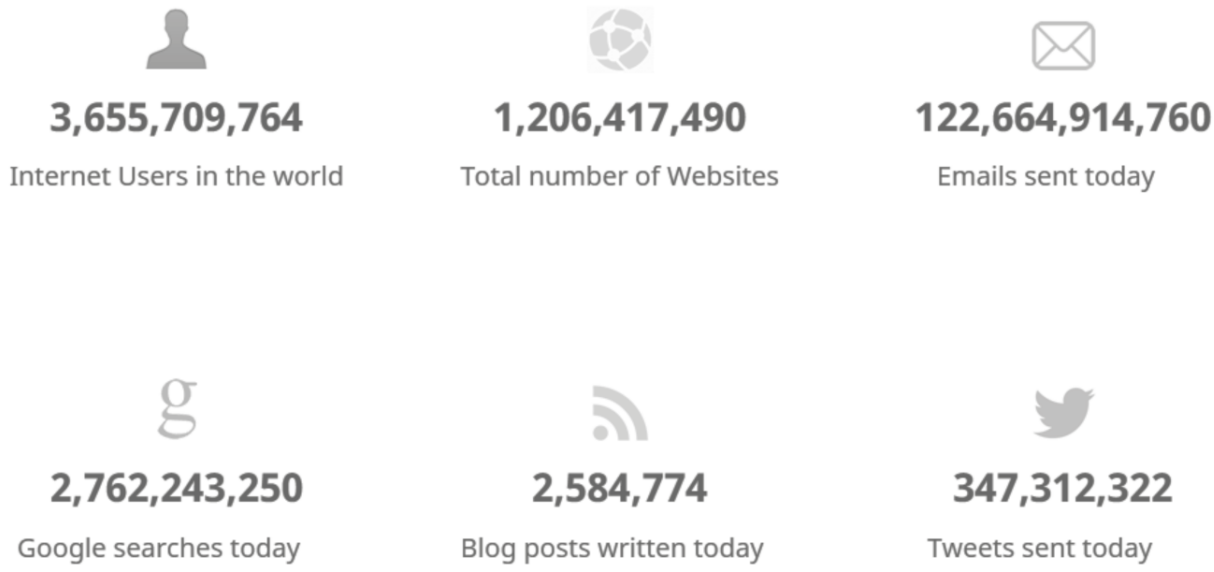

Figura 1. Cenário relativo à utilização da Internet no dia 12 de junho às 12h.

cial de oportunidades a serem exploradas em variados domínios da ciência, da economia, da sociedade, e da política, entre outros. Para empresas e governos, o conhecimento atual (*Nowcasting*) das opiniões dos clientes/cidadãos permite a tomada de decisões eficazes de acordo com aquilo que são os seus objetivos. A ASO constitui um recurso alternativo às sondagens ou aos dispendiosos e morosos sensos. O papel da matemática, mais concretamente da estatística, a par da inteligência artificial, assume destaque neste domínio da gestão da informação.

A ASO tem por base, tal como foi referido, a análise de um conjunto de N textos, $T = T_1, T_2, \ldots, T_N$ e nesses textos são conferidas as ocorrências de L expressões (*stems*), de um corpo de palavras ou sequência de palavras, $S = s_1, s_2, \ldots, s_L$, consideradas relevantes na identificação das possíveis polaridades das opiniões. Essa escolha é por si só complexa e não prescinde de um envolvimento humano direto. Esses termos devem permitir identificar a polaridade do texto (a favor ou contra) ou a sua classificação num conjunto mais alargado de M categorias ou classes, $\mathcal{D} = D_0, D_1, \ldots, D_M$, por exemplo: a favor, indeciso, neutro ou contra. Uma categoria adicional deve ser sempre considerada para cobrir as situações em que não é possível classificar a polaridade do texto. Essa categoria (D_0) é considerada ruído e surge, por exemplo, quando se trata de um texto incompleto, ou quando a opinião expressa é irrelevante ou se encontra fora do contexto. Acontece precisamente que no rastreio de textos de redes sociais esta característica é a predominante.

Depois de definido este corpo de expressões, e com base no mesmo, pretende-se classificar o maior número possível de textos de acordo com a polaridade/sentimento que expressam. Esta tarefa só poderá, como é óbvio, ser realizada por um classificador automático. Um exemplo muito simples e genérico de um classificador, no caso de duas classes D_1 e D_2 linearmente separáveis, consiste na determinação de um hiperplano $g(x) = w^T x + b$, tal que

$$\begin{cases} g(x) > 0, \, x \in D_1 \\ g(x) < 0, \, x \in D_2 \end{cases},$$

onde x é um vetor que caracteriza um determinado objeto. No caso em que os objetos são textos, a cada texto j, $j = 1, \ldots, N$, encontra-se associado um vetor de características (*features*) $x^j \in 0, 1^L$, tal que:

$$x_i^j = \begin{cases} 1 & \text{se o texto } j \text{ contiver o termo } s_i \\ 0 & \text{caso contrário} \end{cases} \quad i = 1, \ldots, L.$$

Deste modo, o corpo dos textos é mapeado numa matriz $0/1$ de dimensão NxL.

Como não é possível neste artigo elaborar em detalhe a construção do classificador, refira-se apenas que, no caso presente, um pequeno subconjunto dos N textos, selecionado aleatoriamente, denominado por conjunto de treino, é analisado e classificado (com supervisão humana) de acordo com a sua categoria $D_i \in \mathcal{D}$ (entre as M consideradas). Com esses textos previamente classificados, é desenvolvido um modelo de classificação automática que será depois testado com outro subconjunto de textos, denominado conjunto de teste. Esse classificador poderá ser desenvolvido utilizando redes neuronais, SVM (*support vetor machine*) ou inferência Bayseana e, uma vez adotado, servirá para classificar automaticamente milhões de outros textos.

Na prática, M é geralmente menor do que 10, correspondendo às categorias/classes distintas de opiniões, L (dimensão do corpo de palavras) é da ordem das centenas, enquanto N pode corresponder a milhões de textos. Como tal, a dimensão do espaço dos vetores das características é da ordem $K = 2^L$, pelo que a primeira simplificação consiste em considerar apenas um subconjunto $\underline{S}$ de S dos vetores efetivamente observados no conjunto de treino, permitindo assim uma redução considerável para $\underline{K} = \#\underline{S}$.

O objetivo final da análise consiste em estimar a distribuição de probabilidade das categorias de opiniões/sentimentos, $P(D)$, $D \in \mathcal{D}$, a partir da distribuição empírica das opiniões/sentimentos dos N textos. Matricialmente,

$$P(D) = P(D \mid \underline{S})\, P(\underline{S})\,, \tag{1}$$

onde $P(S)$ representa a distribuição dos vetores de características em S. No entanto esta metodologia tem, na opinião de Stefano Iacus, falhas que explicam alguns insucessos em previsões baseadas na análise de textos das redes sociais.

PORQUE FALHA A APREDIZAGEM AUTOMÁTICA NAS REDES SOCIAIS?

Relativamente à equação (1), o facto de o conjunto dos textos de treino ser muito menor do que o total dos N textos, de as categorias D_i para $i \neq 0$, serem expressas para um pequeno subconjunto de vetores S_j de S e de D_0 ser a categoria mais frequente, tem como consequência que as probabilidade condicionadas em (1) assumem valores muito próximos de zero e, pelo contrário, a categoria D_0 é sobrestimada quando a agregação do conjunto de treino e teste for efetuado, pois muitas categorias serão enviesadamente estimadas como sendo D_0. Assim, uma alternativa consiste em determinar $P(D)$ através de,

$$P(\underline{S} \mid D)\, P(D) = P(\underline{S}), \tag{2}$$

obtendo-se, pela inversa generalizada de $P(S \mid D)$,

$$P(D) = [P(\underline{S}|D)^T P(\underline{S}|D)]^{-1} P(\underline{S}|D)^T P(\underline{S}), \tag{3}$$

Esta ideia é muito simples do ponto de vista matemático, mas permite desenvolver uma estratégia muito mais eficiente em ASO, como parecem atestar os resultados que foram reportados pelos colaboradores do "Voices of the Blog", tendo obtido diversos sucessos em previsões, por exemplo, de resultados eleitorais (eleições 2016 EUA; ver figura 2) e referendos (BREXIT).

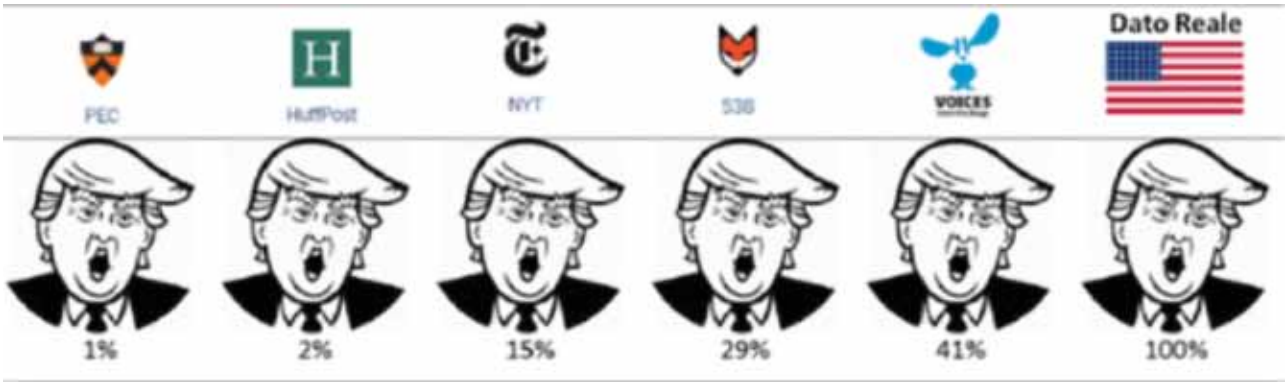

Figura 2. Comparação da previsão do *Voices of the Blog* com outras fontes.

Por fim, corroborando as palavras de G. King, "The best technology is human-empowered and computer-assisted", Stefano Iacus defende que "na maioria, os métodos não supervisionados em ASO falham por serem classificadas como neutrais afirmações que em 90% dos casos não o são, ou ainda como positivas quando são negativas. A dificuldade de uma análise automática lidar com a ironia, o sarcasmo, metáforas, ambiguidades e trocadilhos exige uma técnica que combine a supervisão humana com a automática. Por exemplo, a frase – Este filme tem *boas premissas*, um *bom enredo*, um *elenco excecional*, atores de *primeira classe*, o ator principal *dá o seu melhor*. Ainda assim é *péssimo* – contêm cinco expressões positivas contra uma negativa, levando a que, numa análise automática, pudesse ser classificada erradamente como uma opinião positiva. Assim sendo, enquanto que ontologias e NLP (*Natural Language Processing*) são adequadas em casos nos quais não existe grande subjetividade, como documentos oficiais, decisões de tribunais, e artigos científicos, é de desencorajar a sua aplicação exclusiva na análise de textos de redes sociais.

BIG DATA: MATHEMATICS IN INDUSTRY 4.0

Como foi já mencionado, este artigo pretende resumir algumas das ideias apresentadas por Stefano Iacus, professor de Estatística e diretor do Laboratório de "Data Science", da Universidade de Milão, na conferência "Big Data: Mathematics in Industry 4.0", promovida pela PT-MATHS-IN. O seu conteúdo lança um repto interessante que vai exigir uma reflexão profunda por parte dos cientistas nesta atual revolução digital, no admirável mundo novo do Big Data. Por essa razão a PT-MATHS-IN elegeu este tema para o seu workshop de apresentação, que decorreu no ISEP – Porto, no dia 2 de junho. Estiveram presentes mais de 130 participantes, de cerca de 35 organizações distintas, incluindo um número considerável de estudantes, estando a indústria também fortemente representada. Procurando abordar alguns dos principais aspetos do Big Data, este encontro contou com outros excelentes oradores, especialistas em áreas tão distintas como o atuariado (Gabriel Bernardino), a deteção de fraudes (João Resende) e a recolha de dados (Filipa Calvão). É de assinalar ainda a

Figura 3 – Sessão de abertura do encontro Big Data – Mathematics in Industry 4.0

visão estratégica do papel dos matemáticos na Europa do Big Data deixada pelo presidente da EU-MATHS-IN, Wil Schilders. O encontro conseguiu imprimir nos participantes uma visão muito abrangente não só do que comporta este tema tão vasto do Big Data, mas também da sua influência no modelo de sociedade do futuro.

REFERÊNCIAS

[1] A. Ceron, L. Curini, S. M. Iacus, *Politics and Big Data: Nowcasting and Forecasting. Elections with Social Media*, Routeledge. (2016).

[2] D. Hopkins, G. King, "A method of automated nonparametric content analysis for social science", *American J. Pol. Sci.* 54 (1) (2010) 229–247.

[3] S.M. Iacus, "Big data or big fail? the good, the bad and the ugly and the missing role of statistics", *Electronic J. Appl. Stat. Anal.* 5 (11) (2014) 4–11.

[4] S. M. Iacus, A. Ceron, L. Curini, "A fast, scalable and accurate algorithm for sentiment analysis of social media content", *Information Sciences*, 367-368, (2016) 105-124.