Desenvolvimento pt1:

**CIÊNCIA DE DADOS E O FAKE NEWS**

A relação entre estes dois tópicos tem uma história recente, em se tratando da Ciência de Dados computacional, mas o fato da manipulação de dados ser usada em notícias é algo muito antigo [ McLachlan, 2017]. Porém, desde 2010, esse vem sendo um tópico de crescimento na indústria da computação, ao mesmo tempo que há uma crise de Fake News em vários países, virando tópico de leis nacionais na Europa, por exemplo. Os cientistas então viram oportunidades de como usar técnicas tanto para divulgação quanto para detecção de Fake News, surgindo um novo campo de pesquisa na área, o qual vem sendo extremamente importante e requisitado. O que a Ciência de Dados faz é usar de dados, hoje abundantes no mundo tecnológico, para melhorar relação com clientes, fazer propagandas direcionadas ou detectar fraudes, tudo isso analisando padrões [Provost, Fawcett, 2013]. A razão disso impactar na detecção de Fake News é que, assim como um email *spam*, há padrões para notícias querem chamar atenção, padrões esses que vêm da percepção humana[ Van der Linden, 2017]. A Ciência de Dados porém vai além disso, podendo ser usada de maneira a mudar a percepção das pessoas, e fazer com que as notícias falsas não sejam apenas detectadas, mas parem de ser amplamente divulgadas.

Neste capítulo oque abordaremos é um histórico da relação entre esses dois temas e daremos uma definição mais formal para Fake News, a qual será usada no capítulo 3, e uma secção sobre os cientistas de dados em si e suas responsabilidades ao trabalhar com esse assunto.

2.1. HISTÓRIA

Não podemos falar da relação entre os dois tópicos sem citar o ano de 1948, quando o jornal Chicago Daily Tribune imprime a icônica frase “Dewey defeats Truman”, quando na verdade o contrário havia acontecido. Isso foi um grande impacto na época, porém o fato foi usado como uma sátira pelo próprio Truman, que ridicularizou o adversário, gerando uma famosa foto dele segurando o jornal.



A repercussão disso levou os jornais a notarem como notícias como essa davam vendas. Na tirinha “The Evil Spirits of the Modern Day Press”[1888], já vemos sátiras de como a mídia controlava a população a partir de manchetes enganosas. Nessa época, porém, pouco a população podia fazer. A verdade é que Fake News, representando histórias enganosas amplamente divulgadas, data de muito antes, até em épocas medievais ou antes de Cristo [Weir William ,2009]. Histórias como o Cavalo de Troia, que não sabemos até a hoje da veracidade, são exemplos disso.

Por outro lado, os estudos de Ciência de Dados começaram muito tempo depois, por volta de 1960, mais tarde, numa conferência, o nome foi usado com seu significado atual pela primeira vez[Press Gil, 1996]. Antes disso, era mais um sinônimo de Ciência da Computação em si. Por enquanto era visto como um ramo da estatística computacional que trabalhava com dados.

Foram nos anos 2000 que estudos mais focados começaram a se desenvolver, com um estudo em particular no qual o autor visava marcar e expandir as áreas da chamada Ciência de Dados[ Cleveland, W. S. 2001], isso desencadeou uma série de pesquisas que aprofundaram várias temáticas e trouxe elas para o campo da mesma, dentre essas, o principal método de combate à Fake News, o reconhecimento de padrões.

A área de reconhecimento de padrões já era estudada antes, em áreas como Psicologia. Mas foi modificada levemente e entrou, junto com o aprendizado de máquina, no ramo da Ciência de Dados por volta dos anos 2000. Essas duas áreas têm muito em comum e é difícil delimitar começos e fins pelo menos com o histórico que temos hoje [Bishop Christopher, 2006]. Ambos os campos se baseiam fortemente na teoria matemática, porém, nesse estudo mostraremos apenas uma usagem prática do reconhecimento de padrões, logo, não irá se aprofundar neste tópico.

Estudos na área de reconhecimento de padrões, até então um ramo de estatística, e ainda não levava esse nome, foram se desenvolvendo desde 1960, quando eram puramente matemáticos, e não havia prospecção de uma utilidade prática nela, menos ainda como uma solução para Fake News. Os estudos rodavam em torno de resolução de problemas matemáticos, apesar de já existir princípios que levariam esse conceito para a computação, como abstração de grandes quantidades de dados e análises de informações complexas [V.N. Vapnik, A.Ya. Chervonenkis, 1974].

Perto da década de 80, surgem alguns estudos sobre a área de computação especificamente, no começo ainda com outro nome, mas agora visando o uso do processamento de computadores para resolver os problemas antes matemáticos. O reconhecimento de padrões não era ainda uma área por si só dentro da estatística, mas os tópicos falados são, hoje, do ramo do aprendizado de máquina. Na época, a *IEEE Transactions on Computers* teve diversos artigos publicados sobre esse tema. Era a febre da computação atingindo a estatística e era uma questão de tempo até as pesquisas ficarem mais concretas.

Nos anos 2000, junto com a já citada publicação de Cleveland, surgiram diversos estudos sobre aplicações mais reais do reconhecimento de padrões, agora com algoritmos que já abriam o campo para ainda mais pesquisas[Mineichi Kudo, Jack Sklansky 2000].

Nos próximos anos houve um aumento em tudo relacionado a Ciência de Dados, a descoberta constante de novos e melhores algoritmos levavam a técnicas da estatística a serem usadas em novos campos, empresas lucraram muito com a mineração de dados. Ao longo dos anos, isso só foi aumentando. Em 2011 houve um aumento de 15,000% em trabalhos procurando por Cientistas de Dados, seminários e conferências sobre o tema tomaram conta de convenções de computação[Keith D. Foote, 2016].

Foi durante esse tempo que a aplicação de detecção de padrões começou a ser usada para problemas reais, a detecção de Fake News viria a ser uma dessas aplicações.

Nessa época, o Facebook começava a crescer, e a Ciência de Dados, especificamente o reconhecimento de padrões, começou a ser usada com uma aplicação de manipulação, para mostrar à pessoas o que elas querem ver. Esse tipo de uso se transformaria na grande crise de Fake News, a qual ocorreu entre os anos de 2015 e 2017 na eleição presidencial Norte Americana, que fez com que notícias falsas fossem amplamente divulgadas, e mostrou o quanto isso funcionava como massa de manobra. A partir desse evento, portais de notícia seguiram a tendência. Se preocupar se uma notícia é verdadeira antes de compartilhá-la virou uma preocupação global.

Surgem então, a partir de 2016, pesquisas ,dentro da área do reconhecimento de padrões, relacionadas à Fake News. Por ser recente, ainda não há estudos muito aprofundados, porém já se tornou o maior desafio dessa área nos último anos, sendo extremamente destacada em conferências de Ciência de Dados. Os estudos ainda estão sendo desenvolvidos e vemos nessas conferências as versões ainda em construção sendo apresentadas por cientistas mas não há ainda referências de fato de trabalhos englobando esse meio específico em 2018, sendo que há artigos sobre temas paralelos usados como base para essas pesquisas que serão apresentados ao longo desta monografia.

2.2. FATOS VS. VERDADES

A definição da palavra Fake News é extremamente importante para a Ciência de Dados resolver tal problema, e consequentemente, para essa pesquisa. A preocupação dos pesquisadores é que ao tentar censurar-las, eles acabem também censurando outros tipos de informação. Isso levou-os a um extenso debate moral e ético sobre onde os cientistas devem atuar.

A maior questão é diferenciar as notícias falsas e manipuladoras de duas coisas: erros e sátiras. O já citado exemplo de “Dewey Defeats Truman”, por exemplo, apesar de ser uma notícia falsa, não era Fake News da maneira que queremos detectar, já que foi um erro de impressão. Uma sátira em alguma tirinha ou site de notícias engraçadas poderia também divulgar algo como “Trump aponta *Alien* como possível ministro da economia”, que seria uma mentira, mas também não seria Fake News.

O desafio é então definir o que exatamente é que deve ser procurado por algoritmos de reconhecimento de padrões para detectar Fake News. A ideia é que apenas as notícias maliciosas, divulgadas com a intenção de enganar pessoas, sejam alvos da censura. Enganar é uma palavra essencial aqui, pois se alguém acredita em algo que a ciência prova o contrário, por exemplo, e divulga, não significa que essa pessoa está tentando enganar ninguém. Do ponto de vista dela, ela está apenas divulgando a verdade. Isso nos leva a ainda outro debate.

É uma grande questão então para os cientistas, definir qual escopo do que estão trabalhando, essa também é uma questão ética que será abordada na próxima subseção. Vamos então definir e discutir os termos que serão utilizados neste trabalho.

2.2.1 - Fatos

Serão dito fatos, tudo que for provado cientificamente. São coisas como evolução das espécies, teorias físicas, provas matemáticas e principalmente, eventos bem documentados. A principal preocupação é a divulgação de dados que não são fatos, como fatos. Não é um problema a divulgação de teorias, desde que o intuito seja de informação.

2.2.2 - Verdades

Verdades é o que não pode ser considerado um fato, ou até vai contra um fato, mas não tem a intenção de denegrir, como crenças, ou teorias da conspiração. O ponto é que algumas pessoas acreditam em coisas e quando compartilham-na, não é com a intenção de enganar ninguém, pelo contrário, ela está tentando convencer outros a verem seu ponto de vista.

2.2.3 - Sátiras

Sátiras são notícias que não tem intenção de divulgar informações, e assim, não tem intenção de enganar ninguém, apenas dar algum teor cômico a algum evento ou fato, que pode ser enganador mas as pessoas sabem que se trata de uma piada.

2.2.4 - Fake News

Iremos chamar nessa pesquisa, e é como é visto por grande parte dos pesquisadores na área, de Fake News, notícias que divulgarem algo que for de encontro com um fatos históricos, ou não for nem um fato nem uma verdade. Assim, eliminaremos a censura de opiniões e pontos de vista. As sátiras não serão consideradas notícias, e assim, não serão também Fake News.

Tais definições servem para definir melhor o escopo do problema, e é o primeiro passo quando cientistas vão embarcar nesse meio de usar o reconhecimento de padrões para detecção de Fake News. Porém elas não bastam, apesar de termos essas definições, a existência de todas essas nuances dificulta muito algoritmos de detecção. Falaremos mais disso no capítulo 3, e como resolver esses problemas. O quesito de erros será melhor abordado na secção seguinte, onde será descrita a responsabilidade dos cientistas de dados na detecção de Fake News.

2.3. Responsabilidade dos cientistas

Cientistas são humanos, e assim como tal, podem cometer erros e ter seus próprios pontos de vista. Na ciência em si, isso é relevado pelo fato que é algo muito difuso e incerto. Se uma teoria for provada errada por qualquer outro pesquisador, haverá um debate acadêmico e a ela provavelmente não será aplicada em nenhuma ação real. A verdade é que muito já foi debatido sobre a falsificação de teorias ou métodos na ciência no passado [Lydia Patton, 2014], e apesar de ainda haver certo medo por parte da academia científica, o problema não é mais alarmante.

A Ciência de Dados passa por um problema similar, por ser muito nova, e ser uma aplicação prática, algumas de suas áreas sofre certo risco de cair em mãos erradas. O reconhecimento de padrões, e especificamente a detecção de Fake News é uma dessas áreas.

Por ser tratar de um assunto delicado, quando falamos a respeito da censura, a detecção deve ser feita de maneira completamente imparcial, dado alguns parâmetros, como as definições que demos na secção anterior. Porém definições podem ser mal interpretadas e cabe aos cientistas que vão trabalhar com os dados que não usem os algoritmos de forma a prejudicar alguém ou censurar alguma notícia.

Não cabe aos cientistas então, julgar o conteúdo da notícia, e sim, se ela se enquadra na definição feita por eles, do que é Fake News. Eles têm, porém, certa liberdade ao tratar de erros, citados na seção anterior. Apesar de não se tratarem especificamente de Fake News, a maioria dos algoritmos visa pelo menos alertar o leitor de que aquilo é um erro.

Então nesse meio, em que o escopo é extremamente sinuoso, e fazer algoritmos que consigam diferenciar entre uma notícia falsa e uma sátira, por exemplo, é extremamente difícil, é ainda mais importante que os cientistas que fazem esses algoritmos tenham noção de sua responsabilidade e tarefa na detecção de Fake News.

2.4. Considerações do capítulo

Neste capítulo foi discutida a relação dos dois tópicos desta pesquisa, um histórico de como Fake News sempre esteve presente na humanidade e como estudos da área de reconhecimento de padrões foram se desenvolvendo ao longo dos anos até se tornar robusta o suficiente para detectar essas notícias falsas. Vimos diversas pesquisas no começo da área que foram importantes, como a publicação de Cleveland, e o grande aumento na área da Ciência de Dados que ocorreu nos anos 2000, citado em livros como “Data Science, the sexiest job in the 21st century”.

Além disso foi visto a definição mais formal de Fake News que é usada pelos cientistas, e por qual motivo é importante dar essa definição, como os algoritmos podem ser criteriosos e que é uma dificuldade para que os algoritmos de detecção não virem uma censura.

Por fim, foi abordado o lado dos cientistas e como eles devem ter responsabilidade na hora de pesquisar e aplicar algoritmos para que a definição de Fake News seja seguida à risca e a sua detecção não se tornar algo perigoso.