# Introdução e Motivação

*(surtos e pandemias de doencas)*

**Ideia 1: A idéia aqui é falar do controle falho (ou complexo) de monitoramento online das pandemias. Na verdade talvez não seja falho, mas cauteloso, formal... O que não impede que tenhamos um modelo de alerta q compõe e contribui com info a esses órgãos.**

**Ideia 2: falar da morosidade/celeridade como o problema a ser resolvido**

**Ideia 3: uso de modelos e redes sociais para auxiliar o controle e predicaos de surtos**

Ideia 1

[D01] The CDC (Centers for Disease Control and Prevention) currently diagnoses millions of cases of infectious diseases annually, generating population disease distributions that, while accurate, are far too delayed for real-time monitoring.

[D02] In March 2011 the most powerful earthquake and tsunami in Japan’s history caused horrifying devastation on the country’s northeastern coast. Along with a massive loss of life, the entire infrastructure of the region was destroyed: buildings were crushed and telephone lines were down. However, the mobile internet was still available, and resourceful doctors decided to use Twitter to inform chronically ill patients where they could obtain essential medicines.

[D02] After the Haitian earthquake, researchers used HealthMap’s automated surveillance system to chart the cholera outbreak. HealthMap looks at trends in the volume of reporting in informal sources, such as Twitter and news media, as well as collecting some data from official reports. Gathering data by this route made information on the distribution of cholera cases available two weeks before official sources released it.

Ideia 2

[D01] The ability to instantly compile and monitor such distributions is critical in identifying outbreaks and facilitating real-time communication between health authorities and health-care providers.

[D02] Public health agencies rely on traditional methods of surveillance to monitor outbreaks of disease. These include collection of diagnostic information from doctors and laboratory reporting of test results. Although this way of gathering data is very accurate, it can take a long time to identify new outbreaks and orchestrate a response. And time is critical when trying to prevent rapid spread of a disease.

Ideia 3

[D01] This task, however, is made challenging due to the lack of instantly available public health information, creating a need for the analysis of disease spread on frequently updated social media websites.

[D02] What has caught the attention of infectious disease experts is the growing number of informal sources of information that can provide a much faster picture of outbreaks.

[D02] monitor news and social media sites, including blogs, to pick up clues about emerging public health threats, but the information is less accurate and needs to be verified.

[D02] “The speed is useful,” “an extra week or two can be massively important in preparing a response.”

[D03] Data mining social media has become a valuable resource for infectious disease surveillance.

# Problema e Sub-Problemas

*(celeridade e realtime da evolução do surto)*

**Problema**

[D01]This task, however, is made challenging due to the lack of instantly available public health information

[D01] An approximate three week delay is incurred in the generation of the disease distribution due to the time-consuming process of aggregating national patient re- ports.

**Mas não pode ser apenas ESSE problema... pois ele já tem uma solução... temos que buscar um outro problema dentro desse problema. Algo mais específico que vai nos linkar com a contribuição do trabalho (QUE EU AINDA NÃO SEI QUAL É ... mas que vai aparecer)**

[D02] Public health agencies rely on traditional methods of surveillance to monitor outbreaks of disease. These include collection of diagnostic information from doctors and laboratory reporting of test results. Although this way of gathering data is very accurate, it can take a long time to identify new outbreaks and orchestrate a response. And time is critical when trying to prevent rapid spread of a disease.

[D02] “The speed is useful,” “an extra week or two can be massively important in preparing a response.”

[D02] Obtaining information directly from the public through informal sources is particularly valuable when local outbreaks are not covered by traditional surveillance systems. In many countries surveillance systems are not as robust as in the UK because of social, economic, or political constraints, and natural disasters can also disrupt collection of data.

# Contribuição e Hipótese

**Ideias:**

[D01] need for the analysis of disease spread on frequently updated social media websites.

[D01] a novel pipeline based model to generate a real-time, accurate depiction of infectious disease propagation using Twitter data.

[D01] holistically characterize disease spread using Twitter

[D01] generating a real-time ILI distribution exclusively from Twitter data.

[D01] may provide a tool to epidemiologists for faster response to unknown infectious diseases.

[D01] infectious disease model premised on real-time Twitter data that incorporates a multi-step approach to identify “disease-linked” relevant tweets.

[D01] A correlation with the CDC ILI distribution (r = 0.983) representing an improvement over current state-of-the-art Twitter-based methodologies across one year.

[D01] Proof of robustness of our approach to external noise as signified by its correlation coefficient of 0.947 with mathematical disease simulations.

[D01] achieving a high level of noise invariance as a result.

[D02] the medical potential of this untapped source of data is beginning to be recognised. Infectious disease experts and computer scientists are working together to use this open data to improve disease surveillance.

[D02] monitor news and social media sites, including blogs, to pick up clues about emerging public health threats, but the information is less accurate and needs to be verified.

[D02] Social media represents a new frontier in disease surveillance.

[D02] Quicker detection means more time to prepare resources. […] “The speed is useful,” “an extra week or two can be massively important in preparing a response.”

[D02] the main benefits are shortening the length of time it takes to detect outbreaks to improve responses and allow healthcare agencies faster communication with the public.

**Contribuição original:**

*… ainda temos q encontrar*

* **Resolver problema de confiabilidade nos dados**
* **Maturidade das ferramentas**
* **Entregar Add-value**
* **Reputação?**
* **Mentiras e lixos nas info**

[D02] As more and more information becomes available the background noise of these sites increases exponentially and with it, rumours and half truths. More models are needed to filter and validate the data from these informal sites.

[D02] The challenges posed by the veracity of social media information remain central whether it is used for gathering disease intelligence or urgent doctor-patient communication.

[D03] there are consider- able risks associated with incorrectly predicting an epidemic.

**O nosso tem q ser algo diferente... nem q seja um pouco, mas a proposta tem q ser outra. Isso já está bem manjado.**

xx

# Exemplos do Problema

*(evidencias dos efeitos de atraso desde Ebola ao Corona)*

# Fundamentação

*(Twitter, rede social, sentimento, PLN, Classificadores, Machine Learning, Modelos* matemáticos)

**[Twitter – definição do q é o twitter ... 1 parágrafo]**

**[ILI – CDC – SIR – SEIR ...tem mais]**

**[ECDC European Centre for Disease Control and Prevention (ECDC) }**

**[US Centers for Disease Control and Prevention (CDC) ]**

# Trabalhos Relacionados

*Separar em 3 grupos: geração #1, geração #2, geração #3Covid*

* *Tentar identificar a separação das gerações*
* *Ver a diferença das arquiteturas*
* *Diferenças dos resultados*
* *Diferenças das técnicas*
* *Diferencas das fontes de dados (twiiter etc)*
* *Diferenças das avaliações estatísticas*
* *Diferenças das fontes de comparação (SIR, CDC, Trends)*
* *Diferentes da Maturidade dos modelos*

[D01]

Prior studies have utilized Twitter data to analyze textual sentiment, public anxiety regarding stock market prices, and opinions of restaurants and movies (Citar alguns desses trabalhos ou um survey)

**Mas eu acho q temos vários trabalhos já feitos de twitter para controle de doenças e é isso q iremos mostrar aqui.... em fases talvez... culminando com a discussão final do NOSSO problema e possivelmente a distinção do nosso trabalho**

**Tentar separar em grupos ou gerações...**

**Geração 1 - maturidade**

**ideia**

[D01]presented a keyword-based Tweet distribution to ap- proximate CDC curves or formulated a regression problem, employing supervised machine learning techniques to model disease spread over time.

**Os trabalhos Aparentemente eram limitados e esparsos**

**State of the art da época para a 2012-13 flu season a correlação de 0.877 (menor que o conseguido por [D01] na geração 2.**

**A meta era: aim of ascertaining the efficacy of the social media platform in modeling infectious illness frequency. Mas nao sei se conseguiu realmente gerar uma ILI em tempo real**

**crítica**

[D01][…] fail to adequately eliminate irrelevant tweets, posing significant issues to learning-based predictors that subsequently train using irrelevant data. […] presenting severe problems to distributions that aim to characterize influenza-like illnesses (ILI). Finally, many prior methods are unable to plot real-time ILI distributions, rendering them unable to provide early-warning benefits for health care providers.

[…] fail to categorically eliminate tweets on premises other than hashtag analysis.

**Related works**

[D01] Bodnar and Salath ́e (2013) provide a comprehensive summary of these methods, using over 240 million tweets in their analysis. Their work concludes that the inclusion of “seemingly irrelevant” tweets in a sup- port vector machine multivariable regressor yields correlations as high as 0.783, suggesting that methods reporting lower r-values have failed to properly learn information from tweets, potentially fitting the data due to other associated factors. **Tenho que ler esse aqui**

**Geração 2 – talvez maturidade**

**Ideia**

Comparacao com outros modelos

[D01] evaluate the effectiveness of our model by comparing our Twitter-generated disease distribution with both the CDC ILI curve and SEIR (susceptible, exposed, infected, recovered) disease spread simulation distribution

**A meta era: aim of ascertaining the efficacy of the social media platform in modeling infectious illness frequency. (ILI)**

**Crítica**

**Related works**

[D01]

a novel pipeline based model to generate a real-time, accurate depiction of infectious disease propagation using Twitter data. […]

an amalgam of natural language processing and supervised machine learning, is invariant to mass media hype and significantly reduces the noise introduced by the use of tweets. […]

multi-step classification procedure, whereby tweets are categorized into distinct subsets from which only relevant tweets are considered. [..]

We further develop random forest and support vector machine classifiers to cull spam and identify tweets regarding infectious diseases, generating a real-time ILI distribution exclusively from Twitter data. [...]

generating a real-time ILI distribution exclusively from Twitter data. […]

The correlation coefficient between the Twitter disease distribution obtained via our approach and CDC data from mid-2013 to mid-2014 was 0.983, improving upon the best model published for the 2012-13 flu season.

**Geração 3 – talvez**

**Ideia**

xxxx

**A meta era: xxxxx**

**Crítica**

**Related works**

[xxx]

**Flu Survey**

**Flu Tracker**

[D02] The Flusurvey project, part of a European initiative to monitor influenza trends, collects data from over 2000 volunteers who log on every week to report any flu-like symptoms. It provides a useful addition to the traditional methods of surveillance because most people with flu do not see their general practitioner.

# Discussão e Crítica

*(Ideias e criticas ao q foi apresentado e lido)*

# Técnica e Arquitetura para o nosso futuro

*(Pra gente usar no futuro do nosso experimento - ou analisar pelo menos)*

* [D01]multi-step classification procedure, whereby tweets are categorized into distinct subsets (three unique categories of tweets: self-reported, non self-reported, and spam.)
* [D01] approach: Hashtag Specification P-Metric populatiry> Linguistic Term Association > Term Corpus Topic Modeling TF-IDF >Term Corpus Topic Modeling k-means> Term Corpus Topic Modeling> Term Corpus Topic Modeling:
* [D01] random forest and support vector machine classifiers to cull spam
* [D01] comparing our Twitter-generated disease distribution with both the CDC ILI curve and SEIR (susceptible, exposed, infected, recovered)
* Stanford Spinn3r dataset, a collection of over 100 million tweets from 2013—2014
* Pearson’s correlation coefficient
* Kullback-Leibler divergence.
* CDC ILI até os dias de hoje... tem CDC Brazil? Europa?
* Tem q fazer isso: [D01] eliminate vast selections of irrelevant data, especially from a noise-riddled network such as Twitter, and successfully model the disease distribution with the resulting salient infor- mation.

# Bibliografia

[D01] Disease propagation in social networks: a novel study of infection genesis and spread on twitter – 2016

[D02] Can Twitter predict disease outbreaks? -2012

[D03] Validating Models for Disease Detection Using Twitter - 2013