# 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

[D04] The effect of seasonal epidemics and potentially pandemics repre- sents a significant issue for public health.

[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.

[D06] In the spring of 2009 a new strain of influenza, pandemic influenza A (H1N1) [pH1N1], emerged, beginning in Mexico and quickly spreading to the United States and around the world [8].

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. […] The rapid adoption of social media and the internet in general has opened the door for novel developments in epi- demiology

[D04] propose a classification model that classifies tweets related to Zika and thus enables us to extract helpful insights into the community.

# 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.

[D04] The effect of seasonal epidemics and potentially pandemics repre- sents a significant issue for public health. In this context, early warnings and real time tracking of the spread of disease is highly desirable.

**Aqui um outro problema de TI:**

[D04] key problem is how to filter the noise coming from tweets with similar terms but irrelevant content

# 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.

[D03] The large amount of social media data combined with the small amount of ground truth data and the general dynamics of infectious diseases present unique challenges

[D03] - Methods for validation as- sume that the training and testing data are independent of each other. … due to the strong spatial and temporal nature of infectious disease dynamics… a lack of multiyear social media datasets – this assumption may result in an inaccurate model.

[D04] detecting disease outbreaks through an automated, scalable Cloud-based system for collecting, tracking and analyzing social media data.

**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.

**Trecho livre q Podemos colocar:**

**Few research has been done in building intelligent models for community support on social networking sites, and thus, our approach demonstrates one such novel method blab la bla.**

**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)*

--- acho q isso eu estou coletando na introdução (ideia 1)

# Fundamentação

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

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

-- tenho um arquivo separado com a fundamentação do q é o twitter...

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

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

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

**[ilinet]**

**[U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) ]**

**[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.

**[Google Flu Trends]**

# 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... por maturidade talvez**

**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.

**Pensar nisso......... Analise de sentimento entra na maturidade 2 talvez.... (em geral fica filtragem por sentimento e palavras). Maturidade 1 fica com a analise textual e pura dos tweets**

**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. … // [D03] 238 million tweets .. from the continental United States between October 2, 2011 and May 26, 2012 – a 34 week span [D03] trained our models on 6 subsets: entire dataset grouped by each week; tweets that contained at least one of the following ILI related keywords; [outros 3 idem mas com info especificas de uma reuniao] … [alem disso gerou uma curva senoidal doida] generate one-thousand sine curves with random wavelengths and add noise generated by a normal distribution with a standard deviation of 0.1 to each point [q ele vai usar pra comparer depois acho]. [D03] We found that even irrelevant tweets and randomly generated datasets were able to assess disease lev- els comparatively well. This could serve as a ground level for evaluating other models: if a model can do only slightly better with seemingly relevant data than with seemingly ir- relevant or random data, then it is probably not learning much from the tweets and its ability to fit the data can be attributed to other factors. […] found that even irrelevant tweets and randomly generated datasets were able to assess disease lev- els comparatively well.

**beleza... mas só observou q isso acontece ... não deu nenhuma saída ou pista do q fazer !!!!... de qquer forma a tx de correlação eh bem baixa. .85 para o SVM**,

[D03] comparing the results of a traditional influenza related tweet dataset to a dataset of tweets that has not been filtered for a specific topic, a dataset of tweets related to a topic that is irrelevant to influenza, and a set of frequencies generated from random sine waves … [the result is ]… (i) seemingly irrele- vant tweets are moderately successful in assessing influenza prevalence, (ii) generated frequencies are often as good as measured frequencies from social media, and (iii) the choice of the validation method greatly affects the model’s reported performance. …

[D06] evaluated the accuracy of each U.S. GFT model by comparing weekly estimates of ILI (influenza-like illness) activity with the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet). For each GFT model we calculated the correlation and RMSE (root mean square error) between model estimates and ILINet […] Estimates of ILI produced by the GFT model developed in 2008 correlated highly with historical CDC ILI data […] compared the performance of the two models by calculating the Pearson correlation between model estimates and ILINet data. […] The updated model included approximately 160 search query terms related to influenza activity, […] The original model estimates were highly correlated with ILINet data during the pH1N1 period overall

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Aramaki et al. [12] used a different approach to obtain a better accuracy in their results. They recognised the fact that a tweet containing “influenza” or “flu”, may not mean that the twitter user has got flu; they may be just talking about the topic. To tackle this, they classify tweets into negative (suspicious, questions or news) and positive (the most accurate tweets). They compared several classifier methods such as Logistic Regression, Naive Bayes, Nearest Neighbour and Support Vector Machine (SVM). They based their studies comparing data from the Infection Disease Surveillance Center in Japan. They found that 42% of their analysed tweets were negative. This high value directly impacted the final results. For their analysis, they used an SVM-based clas- sifier. They noted that their method outperformed the Google Trend-based methods applied to the same area of Japan.

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Khan et al. [15] focused on early detection of outbreaks before they transform into full-blown pandemics. They identified three kinds of tweet generators: media and news that posts, tweets related to publicity or advertisement of medicines and tweets created by users that contain key words. The scope of their study was to analyse the last group of tweets to identify true reports of symptoms in the person tweeting. They applied NLP by training a subset of data and applying pre-processing techniques. They used unigram and bigram models to compare their proposed model to obtained a precision of 88.7% compared to other models.

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Lampos et al. [16] focused on reducing the impact of an epidemic disease such as the flu. They analysed tweets from the United Kingdom for a period of 24 weeks. They

A Social Media Platform for Infectious Disease Analytics 529

compared this data with official data from the Health Protection Agency. In this work, they used textual analysis and not just a count of occurrences of words, e.g. flu. They obtained a high correlation but noted that the results could be more accurate if they applied methods to remove media hype and discussions around health events.

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Achrekar et al. [17] tried to go further by attempting to predict flu trends. In their work, they regarded Twitter users as sensors. They removed retweets and tweets from the same user, and labelled tweets with data related to influenza-like illness (ILI). In order to collect data from Twitter, they use the search API using keywords such as flu, swine flu and H1N1. They compared their results with data from the Center for Disease Control and Prevention and obtained a strong correlation with a Pearson’s correlation coefficient of 0.9846.

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Signorini et al. [18] explored the analysis and tracking of H1N1 activity including symptoms and medications. Using the timestamp and geo-location information stored in tweets, they used Google Maps to show tweet distributions and their temporal dependencies.

[D04] [falando de outro trabalho.... q serve pra caracterizar a geração 1] Doan et al. [19] used tweets collected over 36 weeks. The number of tweets collected in that period of time was over 587 million from approximately 24.5 m users. They used this data to compare data from CDC’s U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet)3. This institution compiles more than 25 million patient registers per year. They used two methods to filter the information. The first method was based on four keywords: flu, cough, headache and sore throat. With this method, they got a Pearson’s correlation coefficient of 0.95. They compared various methods but one key conclusion they reached was that the keywords list was directly related to the official information, i.e. an increase or decrease in keywords correlates directly with the accuracy of official information.

[D04] falando de outro trabalho.... q serve pra caracterizar a geração 1] Hirose et al. [20] proposed a method to forecast influenza outbreaks through Twitter, again using ILINet3 as the official data source used for comparison. They filtered tweets into two groups: positives and negatives. Negatives are tweets contained symptom words but not really expressing influenza reports, e.g. “Football fever this weekend!!!”. On the other hand, they identified that positive tweets are more likely connected to real symptoms. They identified that performing multiple linear regressions could improve the accuracy of the prediction.

**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.

[D05] investigate the relationship between social media data and Google Trend data with specific focus on infectious diseases. We present an efficient and scalable system design leveraging a Cloud-based system architecture based on the Australian National eResearch Collaborative Tools and Resources (NeCTAR – www. nectar.org.au) Research Cloud.

**Critica: comparou com o Trends … pq? Isso não tem sentido .... comparar twitter com trends ??**

[D05] propose a classification model that classifies tweets related to Zika … also […] A special type of analytical methodology called the word clouds was then used, which when given an array of words, gives us insight into what words have the highest frequency and are important for the analysis […] in Australia […] utilized the Australia-wide NeCTAR Research Cloud. NeCTAR offers access to over 30,000 physical servers across multiple availability zones … total amount of data used in this work comprised 6,764,090 tweets

**Achava q eh geracao 1 …. Maturidade baixa... só classificou... usou muito pouco tweet (apenas 1471). Esse artigo só serve para caracterizar o nível baixo de maturidade... só classificador MASSSS ele usou uma analise de sentimento e vou considerar isso como geração 2: Keyword + Sentiment (**positive or negative sentiment )

[D05] Positive tweets were discarded since it can be assumed that tweets related to infectious disease should express negative sen- timent.

**Hummmm tão arbitrário… eu não gostei disso... pode haver tweet neutro ou ate mesmo positivo falando da doença.... Acho q o Sentimento tinha q ser considerado, mas não descartado. Fake e noticia devem ser descartados.... estranho esse procedimento deles**

**Geração 3 – talvez**

**Ideia**

xxxx

**A meta era: xxxxx**

**Crítica**

**Related works**

[xxx]

# 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.
* [D03] usou linear e multivariate regression (além de SVM)
* [D04] Naive Bayes
* three modules: a data collection module, a data processing module and a data analysis module.
* L1 penalty and L2 penalty with regards to the performance of LR and SVM.
* F1-score

# 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

[D04] A classification model to analyze the spread and emerging trends of the Zika virus in Twitter - 2016

[D05] A social media platform for infectious disease analytics – 2018

[D06] Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic - 2011