NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

SAÚDE DIGITAL

Previsão da incidência da gripe com base no Twitter

Cristiana Filipa Cruz Rosa

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

LOMBADA MGI

2018

Saúde Digital

Previsão da incidência da gripe com base no Twitter

Cristiana Filipa Cruz Rosa

MGI

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SAÚDE DIGITAL:

PREVISÃO DA INCIDÊNCIA DA GRIPE COM BASE NO TWITTER

por

Cristiana Filipa Cruz Rosa

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence.

Orientador: Professor Doutor Roberto Henriques

Coorientador: Mestre Ivo Bernardo

Novembro 2018

AGRADECIMENTOS

Primeiramente quero agradecer à NOVA IMS por cinco anos de estudos que culminam neste trabalho. Foi um privilégio estudar com a qualidade que proporcionam aos seus alunos, e por tornarem toda a faculdade num grande grupo de amigos.

Ao meu co-orientador, Ivo Bernardo, por acreditar em mim, me orientar e ajudar em todo este longo e atribulado caminho. Foi um privilégio ter sido orientada por alguém com tanta sabedoria.

Ao Dário Jesus, por todo o apoio emocional, preocupação e motivação que me deu. Obrigada pelas palavras certas nos momentos certos, por todo o carinho e respeito para com o meu trabalho.

À Mariana Nunes, Soraia Lopes, Francisco Espírito Santo, Luís Constantino e Rodrigo Carreiras, pelo tempo que dispenderam das suas vidas para me ajudarem a desenvolver este trabalho mais depressa.

Aos meus pais, Maria Isabel e Mário João, por toda a educação que me deram, por me permitirem seguir os meus sonhos e por me apoiarem em cada passo. Sou o que sou hoje graças a vocês!

Aos meus irmão, Francisco Rosa e Guilherme Rosa, por todo o apoio, principalmente na fase final deste trabalho, por toda a preocupação com a finalização a tempo, pela pergunta de todos os dias “E a tese, já está feita?”. Agora sim!

RESUMO

As redes sociais fazem parte do quotidiano da grande parte da população mundial, onde existe uma forte partilha de conteúdos sobre diversos assuntos. Por esta razão, as redes sociais tornaram-se num repositório de dados, de onde é possível retirar informação valiosa e explorar os interesses da população em tempo-real (Recuero, 2005).

Pensemos no seguinte: quantas vezes vemos notícias no telejornal das quais já tínhamos tomado conhecimento através do Facebook ou Twitter? É neste ponto que percebemos que talvez alguns acontecimentos que impactam a população podiam ser detetados previamente.

Posto isto, o objetivo deste trabalho passa por utilizar as publicações do Twitter (*tweets*) relacionadas com a gripe, e perceber se estas mantêm uma relação com a incidência desta doença, uma das que mais preocupa a saúde pública portuguesa. Este tema torna-se particularmente relevante quando olhamos para a pandemia de gripe que ocorreu em 2009 e que se alastrou mundialmente (Centers for Disease Control and Prevention, 2010). Se esta gripe fosse prevista atempadamente, os países poderiam ter tido mais tempo para se prepararem, recolherem os recursos necessários para combater o surto e avisar a população dos procedimentos a tomar, reduzindo o número de afetados e consequente propagação.

A metodologia do estudo assenta em técnicas de *Data* e *Text Mining*. Começamos pela definição dos termos relacionados com a gripe para filtragem de *tweets*, seguida da recolha dos dados através de uma API (*Application Programming Interface)* e seu pré-processamento. Para obter todos os registos do *dataset* classificados, de modo a ser possível posterior aplicação estatística para as comparações desejadas, foram testados vários algoritmos de classificação - *Random Forest*, *Naïve Bayes* e Regressão Logística - tendo-se obtido melhores resultados com *Random Forest*. Este algoritmo foi então utilizado para classificar todo o *dataset* utilizando um subconjunto dos dados classificados manualmente como treino.

Na análise de resultados foram feitas diferentes comparações entre dados oficiais e dados do Twitter tendo em conta duas taxonomias diferentes para classificação e o desfasamento temporal, ou seja, considerando que a incidência no Twitter é detetada antes da incidência oficial. A relação foi testada aplicando regressão linear e concluímos que existe uma capacidade de previsão da taxa de incidência gripal através dos dados do Twitter, sendo esta dependente tanto do desfasamento temporal com da taxonomia aplicada.

PALAVRAS-CHAVE

Saúde; Gripe; Twitter; Text mining; Data Mining.

ABSTRACT

Social networks are part of the daily lives of a big part of the world population, where there is a strong sharing of content about several subjects. For this reason, social networks have become a repository of data, from which it is possible to extract valuable information and exploit the interests of the population in real time (Recuero, 2005).

Consider the following: how often do we see news on the newscast which we already knew through Facebook or Twitter? At this point, we realize that perhaps some events that impact the population could be detected earlier from different news channels.

Therefore, the objective of this work is to use Twitter publications (tweets) related to the flu to understand if these ones have a relationship with the incidence of this disease, one of the most worrying of Portuguese public health. This issue becomes particularly relevant when we look at the global flu pandemic of 2009 (Centers for Disease Control and Prevention, 2010). If flu dissemination was predicted in a timely way, perhaps countries would have more time to prepare, collect the resources needed to fight the outbreak and talk with the population about the procedures to take, reducing the number of affected and consequent spread of the disease.

The methodology of this study is based on Data and Text Mining techniques. We started by defining the terms related to flu, and then apply them for filtering, followed by data collection through an API (Application Programming Interface) and its preprocessing. To have all dataset records classified, in order to be possible later statistical application to perform the desired comparisons, Random Forest, Naïve Bayes and Logistic Regression were tested, obtaining better results with Random Forest. This algorithm was then used to classify the entire dataset using a subset of the data manually classified as training.

In the analysis of results, different comparisons were made between official data and Twitter data considering two different taxonomies for classification and time lag, that is considering that the incidence on Twitter is detected before the official incidence. The relationship was tested with linear regression and we concluded that there is a capacity of prevision of flu incidence through Twitter data, being this prevision dependent both on time lag and applied taxonomy.

KEYWORDS

Health; Flu; Twitter; Text mining; Data Mining.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

API	*Application Programming Interface*
ATAM	*Ailment Topic Aspect Model*
CDC	*Center for Disease Control and Prevention*
SNS	Serviço Nacional de Saúde
SVM	*Support Vector Machine*
RF	*Random Forest*
OOB	*Out-of-bag*
NB	*Naïve Bayes*

1. INTRODUÇÃO

1.1. ENQUADRAMENTO

A gripe é uma das maiores preocupações da saúde pública portuguesa, afetando o país social e economicamente por estar diretamente associada a uma maior procura de cuidados de saúde (George, 2006). É uma doença de início súbito provocada por um vírus que afeta principalmente o aparelho respiratório (George, 2006) e que pode afetar qualquer pessoa, tendo maior incidência nos meses de inverno, quando se registam temperatura mais baixas.

Atualmente a afluência às urgências hospitalares e aos centros de saúde é tão elevada que os tempos de espera ficam absurdamente altos, tendo a gripe um papel pejorativo nos meses da sua incidência (Tavares, 2017; George, 2006). Entre 2014 e 2015, o povo português foi abalado por uma grande epidemia de gripe, registando-se mortes e muitas horas extraordinárias por parte de médicos e enfermeiros, e segundo alguns autores, o caos instalou-se nos serviços de urgência hospitalar (Leal, Ferreira & Carvalho, 2015).

São casos como estes que denotam a falta de qualidade e eficiência da gestão dos hospitais, e a falta de preparação para surtos de doenças (Leal, Ferreira & Carvalho, 2015). Apesar dos esforços feitos pelo Sistema Nacional de Saúde para combater falhas nos serviços de prestação de cuidados de saúde, muito ainda tem de ser feito para que a qualidade aumente. Para isso, e tendo em conta que a tecnologia está cada vez mais a alastrar-se para todas as áreas, com grandes desenvolvimentos e significativas melhorias nos seus campos de atuação, surgiu um conceito que relaciona as tecnologias de informação com a saúde – saúde digital. Segundo Paul Sonnier, fundador de *Health Digital Group*, saúde digital engloba sistemas e aplicações de tecnologia de informação e comunicação aplicadas à área da saúde para monitorizar, gerir e melhorar a saúde da sociedade, ajudando assim a aumentar a qualidade e a reduzir a ineficiência da prestação de serviços de saúde (Sonnier, 2013).

Até aos dias de hoje muitos desenvolvimentos se fizeram na saúde em termos tecnológicos, no entanto, no que diz respeito a previsão de epidemias, muito ainda tem de ser feito. Fora do âmbito tecnológico, existem várias medidas nacionais para combater a incidência da gripe. Destas medidas são destacadas duas: uma vacina sazonal administrada antes da chegada do inverno e a vigilância epidemiológica (George, 2006). O problema aqui exposto é que, não obstante da eficácia destas medidas para diminuir a incidência da doença, ambas não têm capacidade de prever uma epidemia, tendo sido este comportamento epidémico[1] alcançado com muita frequência nos últimos anos (George, 2006).

A produção de vacinas é limitada e os vírus da gripe estão em constante mutação, ficando a imunidade não assegurada. A vigilância restringe-se ao quadro clínico de doentes que recorrem aos serviços de urgência com sintomas de gripe; esta pode estimar uma crescente incidência, no entanto apenas confirma um surto de gripe depois dos casos serem confirmados laboratorialmente, ou seja, muitos casos ficam por diagnosticar e a real incidência da gripe na população é apenas extrapolada

[1] O comportamento epidémico é declarado em função da incidência gripal dos últimos dez anos, tendo sido na época 2017/2018 declarada epidemia de gripe entre a semana 52/2017 e 7/2018 (Pechirra, Cristovão, Costa, Conde, Guiomar, Rodrigues, Silva, Torres & Machado, 2018).

(Corley et al., 2009). Assim levanta-se a necessidade de previsão para que o governo esteja preparado para uma incidência maior e possa direcionar os seus esforços geográfica e temporalmente.

No presente trabalho pretende-se então criar meios de previsão da incidência da gripe, recorrendo às tão utilizadas redes sociais, que nos podem dar exabytes de informação acerca de pessoas, e recorrendo a técnicas de *Data Mining* e *Text Mining* para tratamento de dados e análise de resultados.

Em Portugal, no espaço de 7 anos registou-se um aumento de utilização de redes sociais, sendo o Twitter uma das mais utilizadas (Agência Lusa, 2016). Esta rede social tem uma grande dinâmica em termos de acontecimentos no momento, e é muitas vezes utilizada quase como diário da vida dos utilizadores.

1.2. RELEVÂNCIA E OBJETIVOS

O tema da saúde emerge no trabalho de projeto pois é do meu interesse pessoal o bem-estar da minha família e amigos, bem como o da sociedade em geral. Para além disso, quero dar o meu contributo na obtenção de mais conhecimento e suporte teórico e prático para o possível desenvolvimento de uma ferramenta que poderá melhorar a saúde pública, tanto em termos de diminuição da força da incidência de gripe como diminuição dos problemas no serviço de urgência hospitalar aqui expostos. Juntei a esta vertente a vertente tecnológica e social pois as redes sociais suscitam-me interesse por assumirem um lugar de destaque no que concerne a recolha de informação acerca da sociedade, onde vejo grandes oportunidades de descoberta de padrões nas mais diversas áreas.

Segundo o relatório da época 2015/2016 do Programa Nacional de Vigilância da Gripe, foram registados em Portugal 1.273 casos de Síndrome Gripal, dos quais 46 morreram devido a infeção respiratória e 9 confirmados com vírus da gripe (Guiomar, Pechirra, Cristovão, Costa, Conde, Rodrigues, Silva, Machado & Nunes, 2016). Na época de 2014/2015 registaram-se 1.366 casos e destes 14 morreram, 3 confirmados com vírus da gripe (Guiomar, Costa, Cristovão, Pechirra, Rodrigues & Nunes, 2015). Estes números, apesar de não parecerem elevados, são preocupantes, pois como podemos constatar, uma simples gripe pode levar a complicações, sintomas severos, internamentos e até mortes.

Posto isto e tendo em conda que a epidemia tem sido frequente em Portugal, é preciso também alertar que a qualquer momento pode ocorrer uma pandemia de gripe, que se trata de uma epidemia espalhada por diversas regiões do planeta, em proporções e gravidade muito maiores (George, 2006). As pandemias ocorrem duas a três vezes por século, separadas por entre dez a cinquenta anos (George, 2006). A última ocorreu em 2009 e provocou a morte a 203 mil pessoas no mundo (Gerschenfeld, 2013). Tendo em conta o histórico de pandemias, a probabilidade de existir uma pandemia nos próximos anos é elevada, sendo que o intervalo médio entre as últimas seis é de aproximadamente 26 anos (George, 2006).

Assim, o presente trabalho de projeto assenta em quatro objetivos principais:

- O objetivo primário é o de explorar os contributos existentes relacionados com os temas aqui abordados, aprofundando conhecimentos teóricos;

- Em segundo lugar, numa vertente mais de desenvolvimento, o objetivo é identificar palavras chave que estejam relacionadas com a gripe, aplicar algoritmos de *Text Mining* e assim aprofundar conhecimentos práticos nas ferramentas utilizadas;

- Em terceiro lugar, pretende-se compreender se o conteúdo publicado no Twitter permite prever a afluência da gripe em Portugal;

- Por último, e tendo em conta os conhecimentos adquiridos neste estudo e os conhecimentos adquiridos com o uso das redes sociais, identificar outras redes sociais nas quais seja possível aplicar a mesma metodologia.

2. REVISÃO DA LITERATURA

2.1. TWITTER E REDES SOCIAIS

Com o aparecimento da Web 2.0 em 2004, uma segunda geração de serviços na rede compartilhados por todos os utilizadores, estes passam também a gerar conteúdo de forma colaborativa e não só a consumi-lo (Barreto, 2011). Com esta partilha de informação foram-se criando relações e estabelecendo contactos, fazendo com que se acentuasse o crescimento das redes sociais. As redes sociais impuseram-se mundialmente e atualmente são das ferramentas que mais geram informação. Recuero (2005) afirma que as redes sociais suscitaram uma série de mudanças no comportamento das pessoas e no fluxo de informação da sociedade, facilitando a sua circulação.

O Twitter, lançado em julho de 2006 nos Estados Unidos da América, com o intuito de funcionar como SMS na internet, é uma rede social de microblogging usado por milhões de pessoas que se conectam entre elas, cada uma na sua própria rede. Os utilizadores podem partilhar diversas informações através de mensagens curtas até 140 caracteres (Huberman, Romero & Wu, 2008). Nestas mensagens podem ser usados cardinais antes de uma certa palavra ("#") de modo a demarcar o seu assunto.

Considerando o seguinte exemplo de uma mensagem no Twitter:

"Tempo de festa, amigos e sardinhas! #santospopulares #Lisboa"

Publicado às 21:00 do dia 12/06/2017

Com este tweet conseguimos perceber que o utilizador se encontra nos santos populares em Lisboa no dia 12 de junho de 2017 às 21h, sendo que esta mensagem foi indexada aos assuntos "santospopulares" e "Lisboa"; todos os utilizadores que pesquisarem por estes assuntos no Twitter poderão encontrar a mensagem deste utilizador.

De uma forma mais teórica, um tweet contem entidades e locais, para além do conteúdo textual. As entidades são os assuntos demarcados com cardinal, URL ou fotos e vídeos. Os lugares são os lugares geográficos mencionados no tweet (Russell, 2013).

Com o exemplo dado podemos concluir que as mensagens no Twitter são uma grande fonte de informação acerca da atividade dos utilizadores. Conseguiremos, então, através do Twitter saber se uma pessoa está doente? Conseguiremos prever a disseminação dessa doença? Conseguimos usar notícias para perceber a incidência da gripe em determinados pontos e prever a sua propagação a zonas vizinhas?

Vários estudos, nas mais diversas áreas, consideraram o Twitter como a sua fonte de dados. Em Lerman e Ghosh (2010) estudaram a rapidez e a disseminação de notícias no Twitter, Paul e Dredze (2012) estudaram a capacidade do Twitter para monitorizar várias doenças que afetam a saúde pública como gripe, infeções, alergias e obesidade, através de sintomas, tratamentos e palavras geralmente associadas a estas doenças. Em Sakaki et al. (2010) foram estudados eventos em tempo real como sismos e foi proposto um algoritmo para monitorizar tweets, de modo a encontrar o centro e a trajetória dos sismos. Em O'Connor et al. (2010) foi encontrada uma correlação entre a

opinião da população e a confiança na política e os sentimentos medidos através da análise de publicações no Twitter.

St Louis e Zorlu destacam no seu estudo duas plataformas, HealthMap e BioCaster, que usam programas para monitorizar notícias e sites sociais para recolher dados sobre ameaças à saúde pública, agindo mais rapidamente na sua detenção e preparação de recursos para combate (St Louis & Zorlu, 2012).

Um estudo feito em 2011 sobre a previsão da gripe suína mostrou que o surto desta gripe em 2009 podia ter sido identificado no Twitter uma semana antes do comunicado oficial (Szomszor, Kostkova, & De Quincey, 2011). Segundo os autores tal seria possível pois os dados oficiais demoraram tempo a serem recolhidos e processados, enquanto que a recolha de dados no Twitter é instantânea. Para além disso denotam que uma peça futura de informação seria obter a localização do utilizador aquando da partilha da mensagem.

A taxa de incidência gripal é apenas uma extrapolação de casos analisados e diagnosticados (Corley et al., 2009) e mesmo havendo precisão nos métodos tradicionais para monitorizar a disseminação e os surtos de doença, leva muito tempo obter confirmações e preparar resposta a esses surtos. Do outro lado, temos uma ferramenta eficiente com informação produzida pelos próprios afetados, o Twitter, capaz de fornecer informação em tempo real acerca da saúde pública (Paul & Dredze, 2012).

2.2. TEXT MINING

O *Text Mining* é um processo intensivo de descoberta de conhecimento e deteção de padrões nos dados através da extração destes de várias fontes de dados de texto (Hotho, Nürnberger, Paaß, 2005). O *Text Mining* tem uma grande importância em ambientes de dados ricos em texto, como é o nosso caso, pois a natureza dos seus dados é não estruturada. Dixon (1997) sugere que o processo englobe os seguintes passos:

1. Encontrar documentos que contenham informação relevante para o estudo e selecioná-los;
2. Extrair informação dos documentos selecionados;
3. Pré-processar os dados de maneira a convertê-los em dados estruturados e prepará-los para aplicação de técnicas de *data mining;*
4. Extração do conhecimento (padrões e tendências).

De maneira a estruturar os dados, Hotho, Nürnberger e Paaß (2005) destacam no seu estudo 3 métodos de *data mining* que podem ser aplicados a texto:

- Classificação, que se baseia em classificar novos dados em classes pré-definidas tendo em conta o que já sabemos de outros. Um exemplo dado em Hotho et al. (2005) é

classificar automaticamente notícias como sendo de desporto, política ou arte. Os algoritmos a utilizar podem ser árvores de decisão, *Naïve Bayes* (NB), *Nearest Neighbor* ou *Support Vector Machine* (SVM) (Hotho, Nürnberger, & Paaß, 2005);

- *Clustering*, uma técnica que agrega textos em classes de acordo com características ou conteúdos comuns. O objetivo é obter a maior homogeneidade possível *intra-cluster* e heterogeneidade *inter-cluster.* Algumas das abordagens para este método passam por algoritmos de *clustering* hierárquico, *k-means* ou *Self Organizing Map* (SOM) (Hotho, Nürnberger, & Paaß, 2005);

- Sumarização, que consiste em extrair do texto pequenas partes que o resumam. Assim cada texto pode ser mais rapidamente classificado através de palavras ou excertos chave.

Este trabalho de projeto incidirá sobre os métodos de classificação e os algoritmos inerentes. Ao longo dos anos foram aplicados inúmeros algoritmos de aprendizagem nos vários estudos feitos e aplicados a dados com origem em redes sociais. Nestes estudos vemos, não só diferentes algoritmos, mas também algumas diferenças nas classes aplicadas à classificação.

Em Corley et al. (2009), um estudo aplicado às tendências de gripe em vários *blogs* da *web*, foram recolhidos 44 milhões de *posts*, entre 1 de agosto e 1 de outubro de 2008, através de uma *Application Programming Interface* (API) e destes filtrados os de língua inglesa e relacionados com gripe, ou seja, os que no seu conteúdo continham as palavras "*influenza*" e "*flu*". Depois de todo o tratamento de dados, os *posts* foram agrupados por mês, semana e dia da semana e normalizados por dia tendo em conta a média de *posts* em toda a base de dados no dia da semana referente.

De seguida foi feita a correlação com o *Center for Disease Control and Prevention* (CDC) que é composto por 2400 prestadores de cuidados de saúde em 50 estados que reportam aproximadamente 16 milhões de visitas de pacientes por ano. Cada relatório reportado contém o número total de pacientes vistos e o número de casos de gripe por faixa etária. A correlação entre os dois *datasets* foi feita com o Coeficiente de *Pearson*, e foi obtido r = 0.767, com um nível de confiança de 95%.

Os autores deram também uma proposta para monitorizar a gripe tendo em conta apenas os *blogs* mais influentes no assunto. Para isso definiram três classes para classificar cada publicação: identificação própria de sintoma, a identificação de outra pessoa que tem sintoma (segunda mão) ou um artigo objetivo (ou opinião). Destas apenas as duas primeiras foram consideradas e assim foram identificados os cinco *blogs* que representam a maior frequência de posts relacionados com gripe.

Paul e Dredze (2012) apresentam no seu estudo um novo modelo, *Ailment Topic Aspect Model* (ATAM), que aprende com sintomas, tratamentos e palavras relacionadas com doenças e associa estes três tópicos às próprias doenças. Começaram com mais de 2 biliões de *tweets* e depois de toda a limpeza e pré-processamento obteram 11.7 milhões relacionados com saúde (tendo em conta apenas uma lista de 20.000 palavras chave aplicada para filtrar o *dataset*).

Para classificação dos *tweets* a serem usados em treino, foram usadas cinco classes: doente, onde a mensagem indicava que o utilizador estava realmente doente; saúde, onde a mensagem era

um comentário sobre algo relacionado com a saúde de outra pessoa; não relacionado; não inglês e; ambíguo, sendo estas últimas autoexplicativas. Para treino foi utilizado o algoritmo SVM com *unigram*, *bigram* e *trigram*. O *n-gram* representa uma sequência de *n* itens num dado texto. Por exemplo, considerando a sequência "Hoje está a chover muito", e considerenado *n-gram* ao nível de palavras, em *unigram* a sequência seria "Hoje", "está", "a", "chover", "muito"; em *bigram* seria "Hoje está", "está a", "a chover", "chover muito"; e em *trigram* seria "Hoje está a", "está a chover", "a chover muito". Com este classificador foi obtida uma precisão de 90.4% e aplicado aos 11.7 milhões tweets, resultou em 1.63 milhões relacionados com saúde.

Depois de toda a construção e aplicação do ATAM, os autores calcularam o número de *tweets* relacionados com gripe por semana e normalizaram pelo número total de *tweets* nessa semana em todo o *corpus*. Aplicaram a Correlação de *Pearson* entre a frequência de gripe nos seus dados e os dados de CDC e obtiveram um coeficiente de 0.934.

Em Sakaki et al. (2010) foi proposto um sistema de notificação de eventos baseado nos terramotos na zona do Japão. Os autores definiram que cada utilizador do Twitter era um sensor e cada publicação do mesmo era uma informação sensorial. A classificação desta informação sensorial foi feita da forma mais simples: cada *tweet* foi classificado como positivo caso se referisse ao evento em estudo, ou negativo caso contrário. Neste estudo foi também usado o algoritmo SVM mas com apenas estas duas classes. Foram utilizados três grupos de *features* para construir o classificador sendo elas *features* estatísticas, com o número de palavras no *tweet* e a posição da palavra-chave; *features* de palavras-chave, ou seja, as próprias palavas contidas no *tweet*; e *features* de contexto, as palavras antes e depois da palavra-chave. Sobre estes três grupos foram aplicadas duas diferentes pesquisas de palavras: "*earthquake*" (tradução: tremor de terra ou terramoto) e "*shaking*" (tradução: tremendo).

Os resultados foram os mostrados na tabela 1.

Features	Precisão "*earthquake*"	Precisão "*shaking*"
A - estatísticas	63.64%	68.57%
B - palavras-chave	38.89%	57.41%
C - contexto	66.67%	86.36%
Todas	63.64%	65.91%

Tabela 1 – Precisão da classificação em Sakaki et al. (2001)

Com estes resultados percebemos que tanto os termos de pesquisa como as propriedades em foco na análise podem representar grandes variações nos resultados obtidos.

Como o objetivo era notificar a população de que existiria um terramoto mesmo antes de este chegar à sua localização, foram utilizados modelos probabilísticos para prever a localização e a trajetória dos terremotos no Japão. Foi utilizado um modelo temporal para aproximar o número de *tweets* de uma distribuição exponencial e obteve-se uma correlação de 87%. O modelo espacial utilizado para encontrar a localização e inferir a trajetória baseia-se em métodos de estimação, como o filtro Kalman e o filtro de partícula, muito usados na estimação de localização.

Em Pang et al. (2002) foram também usadas estas duas classes e o algoritmo *SVM* com *unigram* e *bigram*, no entanto o estudo estendeu-se para *Maximum Entropy* e *Naïve Bayes* para classificar críticas a filmes como sentimentos positivos ou negativos. Ao fazer a comparação de *performance* entre os três métodos de classificação os autores concluíram que, mesmo com pouca diferença, *Naïve Bayes* tende a ser o pior e *SVM* o melhor.

2.3. RANDOM FOREST

O algoritmo de RF, desenvolvido por Breiman, consiste num conjunto de árvores de decisão, que criam vários classificadores e agregam o resultado de todos (Breiman, 2001). Cada árvore é construída tendo em conta uma amostra *boostrap* do *dataset* de treino, ou seja, uma amostra criada aleatoriamente com reposição, do mesmo tamanho que o *dataset* original, e em cada divisão de dados (nó) é utilizado um subconjunto de variáveis escolhidas aleatoriamente para verificar qual destas tem a mais baixa impureza, pela medida de Gini (Khalilia et al, 2011), ou seja, qual é a que melhor divide e prevê o target por si só, tornando essa variável um nó da árvore (Breiman, 2001). Basicamente o algoritmo irá apurar que palavras melhor separam os dados e que melhor preveem se o *tweet* se refere ao evento da gripe. O ciclo repete-se até ao total crescimento da árvore de decisão e o treino termina quando todas as árvores estiverem contruídas.

O algoritmo de *Random Forest* pode então ser descrito da seguinte maneira (Breiman, 2001):

Dado um *dataset* de tamanho n, onde X = x1, x2, …, xi e Y = y1, y2, …, yj; dada uma RF constituída por b árvores de decisão. Para b = 1, 2, …, B:

1. Selecionar uma amostra *bootstrap* (Xb, Yb) de tamanho n;

1.1. Selecionar aleatoriamente k variáveis do total de m, onde $k \leq m$;

1.2. Usar a variável das k que melhor divide os dados tendo em conta a impureza, fazendo dela um nó da árvore;

1.3. Repetir 1.1 e 1.2 até ao total crescimento da árvore;

2. Repetir 1 para todas as b árvores de decisão;

3. Obter a classificação final tendo em conta a maioria dos votos obtidos em toda a RF.

Usar uma amostra aleatória e considerar um subconjunto aleatório de variáveis em cada nó resulta numa grande variedade de árvores fazendo com que o conjunto destas árvores seja mais efetivo do que uma por si só.

No fim do treino todas as árvores de decisão serão usadas para classificar um novo *dataset* não classificado. O *output* do classificador para cada registo é dado tendo em conta a classe que representa a maioria dos *outputs* de todas as árvores na RF para esse mesmo registo.

Como estamos a trabalhar com um algoritmo *bagging*, por selecionar amostras aleatórias com reposição existem registos que nunca são selecionados, chamados de *out-of-bag* (*OOB*), e representam cerca de um terço dos casos (Breiman, 1996). Estes registos servirão para estimar o erro sem ser necessário um *dataset* de teste e sem que seja preciso um método de validação do modelo à parte do mesmo, pois este erro é calculado ao longo do processamento da *Random Forest*. O estudo de Breiman sobre a estimativa de erro mostra evidência empírica de que as estimativas *OOB* são tão precisas quanto usar um *dataset* de teste do mesmo tamanho que o *dataset* de treino (Breiman, 1996). O erro calculado através de *OOB* pode ser utilizado para definir o número de árvores de decisão, à *posteriori* do treino. Este número pode também ser definido pelo utilizador à *priori*.

Este algoritmo foi utilizado em vários estudos para comparação com outros algoritmos tais como SVM, *Naïve Bayes* ou *Decision Tree*, mostrando obter maior precisão que os restantes. Em Koprinska et al. (2007), um estudo sobre a classificação de e-mails, foram comparados diferentes classificadores: *Random Forest, Decision Tree, SVM* e *Naïve Bayes.* Estes classificadores foram treinados para separar os e-mails por pastas, usando um conjunto de dados de cinco utilizadores diferente, e para funcionar como um filtro ao *spam*, usados três conjuntos de dados obtidos de diferentes fontes. Para comparar os diferentes classificadores foram calculadas várias medidas de performance como a *accuracy*, *recall*, *precision* e *F1-score*.

Os resultados para a filtração de *spam* e para a separação por pastas mostram que *RF* é o melhor classificador com precisão entre 98% e 99% e *F1-score* entre 97% e 99%; e precisão entre 81% e 96% e *F1-score* entre 42% e 94%, respetivamente. O autor refere que esta diferença se deve ao facto de cada utilizador ter um critério de classificação diferente o que mostra que um sistema automático irá separar muito bem os e-mails para uns utilizadores e muito mal para outros. Em relação ao pior classificador os autores concluíram que, tanto na filtração como na separação, *Naïve Bayes* se comporta pior que os restantes.

Em Khalilia et al (2011) foi usado um *dataset* não balanceado (quando uma classe contém muitos mais registos que outra) e o algoritmo *RF* para prever o risco de doença baseado no histórico clínico e posteriormente comparada a performance deste algoritmo com *SVM*, *bagging* e *boosting*.

Neste estudo foi inserido um novo método de amostragem, subamostragem aleatória repetida, onde o *dataset* de treino é repartido em subamostras onde cada uma contem os mesmos registos para todas as classes, fazendo com que todos os registos da classe com menos registos sejam selecionados e usados no treino. Foi então utilizado um *dataset* com oito milhões de registos para treinar o modelo *RF* e obtido um AUC de 89,05%, tendo em conta todas as doenças. Para fazer a comparação com os restantes algoritmos foram selecionadas oito doenças crónicas e feito o teste com subamostragem, o novo método inserido neste estudo, e sem subamostragem. Os resultados mostram que a subamostragem se comporta melhor com o *dataset* não balanceado. Em relação aos algoritmos, *RF* detém a maior precisão e o *SVM* detém a menor.

3. METODOLOGIA

3.1. VISÃO GERAL

A figura 1 contém uma visão geral da metodologia aplicada neste trabalho de projeto, com as estapas da mesma: definição de termos de pesquisa, recolha de dados, pré-processamento de dados, categorização, classificação via algoritmo e análise de correlação.

A primeira etapa é então a recolha dos dados do Twitter, de onde é recolhida uma pequena amostra para análise, tentando detetar alguma falha inicial ou algum tratamento de dados que seja necessário para ser posteriormente aplicado ao *dataset* de entrada ao modelo. Da pequena amostra de análise resultarão inputs para pré-processar e categorizar os dados.

Na etapa de categorização o conteúdo das publicações será catalogado e serão selecionadas as que são relevantes para as análises seguintes. Para que tal seja possível, o *dataset* de análise será classificado manualmente para ser utilizado no treino do classificador, e posteriormente, todo o *dataset* será classificado pelo algoritmo de classificação.

De seguida são feitas análises de correlação, comparando os dados recolhidos do Twitter e os dados recolhidos do Serviço Nacional de Saúde (SNS). Esta correlação será feita em termos de volume e tempo, ou seja, os dados serão relacionados semanalmente e será comparado o volume dos mesmos. Para além disso iremos mover um dos *datasets* temporalmente de modo a perceber com que desfasamento os dois *datasets* se ajustam melhor e têm maior correlação.

Cada etapa será explicitada ao longo dos capítulos e subcapítulos seguintes.

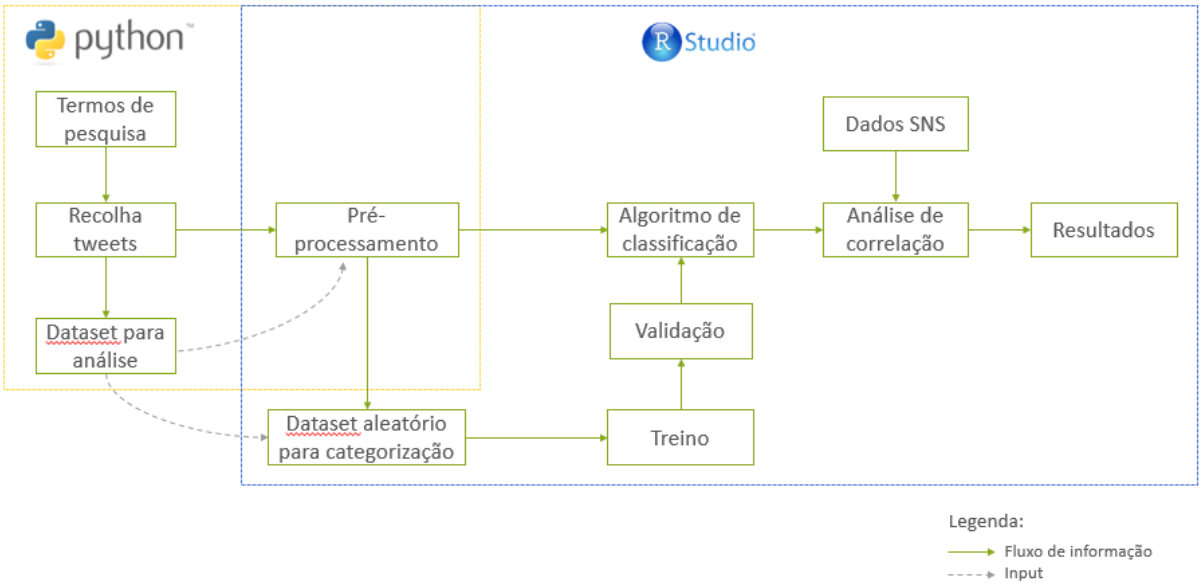

Figura 1 – Metodologia

3.2. DEFINIÇÃO DE TERMOS DE PESQUISA

Numa primeira fase foi necessária a definição dos termos associados à gripe e aos seus sintomas de forma a ser possível a filtragem de *tweet* que estejam ou não relacionados com este estudo. Estes termos englobam as diferentes variações da doença, sintomas e medicamentos.

Categoria	Termos de pesquisa
Doença	Gripe, constipação, virose, sinusite, influenza
Sintomas	Febre, dor de cabeça, espirro, calafrio, tosse, corrimento nasal, dor de garganta, frio, mau estar, nariz entupido, congestionamento, expetoração, dor muscular
Medicamentos	Antigripal, nurofen, mucolítico, expetorante, aspirina, bissolvon, mucosolvan, atossin, cegripe, brufen, ilvico, antigripine, cecrisina

Tabela 2 – Termos de pesquisa para filtragem de *tweets*

3.3. RECOLHA DE DADOS

A recolha de dados do Twitter foi feita sem interrupções entre o dia 1 de novembro de 2017 e o dia 28 de fevereiro de 2018, para poder incluir os primeiros meses em que, estatisticamente, existe maior incidência da gripe, sendo estes entre dezembro e março (George, 2006). Este processo foi feito através da utilização de uma API do Twitter que permite a recolha de informação dos servidores, em tempo-real, tais como o identificador do utilizador, as mensagens publicadas, a data de publicação (GMT +0), a língua e a localização geográfica. A ligação à API foi feita através de código em linguagem python, tendo sido este baseado em Bernardo & Henriques (2014) (script no anexo 8.1).

A recolha de *tweets* foi limitada à filtragem pelos termos definidos acima, e foram assim recolhidos 3.133.226 *tweets*, o que resulta numa média por dia de aproximadamente 26.110 *tweets*.

Para além desta recolha, foi feita outra a nível de dados oficiais acerca da incidência da gripe, através do boletim de vigilância epidemiológica da gripe, publicado semanalmente pelo Instituto Nacional de Saúde. Em cada boletim podemos encontrar a taxa de incidência semanal da síndrome gripal, número de casos confirmados de gripe por tipo de gripe (A, B), número de casos confirmados com outros agentes respiratórios, número de internamentos, número de óbitos por todas as causas, valor médio da temperatura mínima e situação internacional. Recorremos ao seguinte url para obtenção dos boletins das semanas correspondentes ao nosso estudo: http://www.insa.min-saude.pt/category/informacao-e-cultura-cientifica/publicacoes/atividade-gripal/.

3.4. PRÉ-PROCESSAMENTO DE DADOS

A fase de pré-processamento de dados será de extrema importância para o estudo, pois os dados recolhidos serão não estruturados, aos quais não devem ser aplicadas as ferramentas pretendidas.

3.4.1. Dados Twitter

Dado que estamos a lidar com redes sociais e *inputs* gerados pelo ser humano, antes do tratamento de dados foi feita uma revisão humana a uma amostra de modo a perceber que tipo de inconsistências encontramos nos dados e estudar como as podemos resolver, de maneira a que o nosso *dataset* de trabalho seja o mais consistente possível.

O processamento da informação foi feito com auxílio de linguagem python e R, e foi tido em conta o estudo prévio dos *tweets* para melhor estruturar os dados. A análise feita antes do pré-processamento mostra que parte dos *tweets* recolhidos não são de língua portuguesa, contém *links* (para notícias ou artigos relacionados com saúde) ou são *retweets*. Para além disso os *tweets* contém pontuação e palavras sem valor descritivo que não interessam para o classificador e aumentam a dimensionalidade dos documentos (Hotho, Nürnberger, Paaß, 2005).

Para além da filtragem dos *tweets* relevantes para o estudo em causa, tendo em conta os termos de pesquisa, foram então removidos os *tweets* com as seguintes características:

- *Retweets*, marcados no ínicio com "RT", que são *tweets* partilhados pelo utilizador mas publicados por outro (a inclusão destes pode levar a enviesamento da amostra pois representam uma duplicação de informação)

Esta limpeza inicial resultou num *dataset* de 1.060.923 tweets para classificação e treino.

Removemos também do texto alguns componentes, tais como:

- Pontuação e caracteres especiais, exceptuado o "#" que identifica *hashtag;*
- Espaços em branco duplicados;
- *Links* htlm, sobre os quais não há quaisquer análises a realizar;
- Números, sem qualquer valor descritivo para o estudo
- *Stop words*, palavras frequentes mas sem qualquer valor descritivo, como preposições ("com", "de", "para", "em", etc.), artigos ("a", "os", "dum", "pela", etc.), conjunções ("e", "mas", "porque", "que", etc.), etc.

O texto foi também colocado todo em minúsculas, o que ajuda na agregação dos termos, e foi feito *stemming* aos mesmos. *Stemming* converte as palavras na sua forma básica, removendo afixos e flexões das palavras (Hotho, Nürnberger, Paaß, 2005), sendo assim possível identificar palavras semelhantes, que descrevem um mesmo conceito, tendo em conta a morfologia. Pegando no exemplo mais enfatizado deste estudo, as palavras "gripe", "gripado", "engripado", "gripal" serão todas convertidas para "grip".

3.4.2. Dados oficiais

Visto que obtemos diferentes boletins por cada semana, temos de agregar a informação das diferentes semanas num só ficheiro. Posto isto, foi criado um Excel onde temos as semanas por linha e a informação pretendida acerca da gripe (taxa de incidência, internamentos, etc.) por coluna.

	B	C	D	E	F	G	H
1	Semana	Taxa de incidência por 100.000 habitantes	Internamentos	Internamento por virus A	Internamento por virus B	Temperatura mínima	Situação Internacional
2	201744	6,9	0	0	0	10,33	Baixa
3	201745	6,4	0	0	0	5,63	Baixa
4	201746	15,4	1	0	1	4,67	Baixa
5	201747	11,7	0	0	0	7,95	Baixa

Figura 2 – Estrutura de excel com os dados oficiais

3.5. MATRIZ TERM-BY-DOCUMENT

Para aplicar um algoritmo estatístico ou de *data mining* a um *dataset* de texto, é necessário primeiro convertê-lo para uma representação numérica. Para isso, iremos usar uma matriz *term-by-document* que representa uma matriz bidimensional onde de um lado temos os termos e do outro os documentos (Feinerer, Hornik, Meyer, 2008). No âmbito do nosso estudo, as linhas são os *tweets* (documentos), as colunas as palavras (termos) e cada entrada (x,y) representa a frequência de cada palavra y no *tweet* x. O *dataset* é então transformado em vetores numéricos, como podemos ver no exemplo abaixo.

	Palavra 1	Palavra 2	Palavra 3	Palavra 4	Palavra 5	…	Palavra y
Tweet 1	0	1	1	0	1	…	0
Tweet 2	2	0	2	1	0	…	1
Tweet 3	0	1	0	3	1	…	2
…	…	…	…	…	…	…	…
Tweet x	1	2	1	0	2	…	0

Tabela 3 – Exemplo de uma matriz *term-by-document*

Estas matrizes podem alcançar dimensionalidade elevada, dependendo do número de termos e documentos presentes nas mesmas (Feinerer, 2018). Tendo em conta que a nossa matriz contém mais de 1 milhão de documentos e 223.000 termos, podemos suspeitar que temos uma elevada dimensionalidade e dispersividade (uma matriz dispersa contém na sua maioria frequências nulas). Para além disso estamos dependentes da capacidade computacional para processar a matriz

pelo que devemos reduzir a dimensionalidade da mesma de modo a não termos problemas de performance. Esta redução diminui drasticamente a matriz sem perder as relações importantes contidas na mesma (Feinerer, 2018). Foram testados vários valores para remoção da dispersividade, até chegar a um valor que nos desse suficiente capacidade de processamento. Foram então removidos todos os termos que não aparecessem em pelo menos 3% dos documentos, resultando na continuidade de 54 termos para o estudo.

3.6. ANÁLISE EXPLORATÓRIA

Para termos uma maior perceção e compreensão sobre os dados deste estudo, foi feita uma pequena análise exploratória. Desta forma obtemos um entendimento básico mas necessário dos dados antes da aplicação do modelo.

3.6.1. Dados Twitter

Considerando os 1.060.923 *tweets* resultantes do pré-processamento descrito acima, obtemos uma média de aproximadamente 8840 *tweets* por dia.

Para primeira análise, verificámos a distribuição de *tweets* pelos dias de recolha (120 dias no total). Na figura 3 conseguimos verificar que existem três falhas nos dados, uma a meio de janeiro, outra no início de fevereiro e outra a meio de fevereiro. Em janeiro desde as 22 horas de dia 18 até às 9h do dia 21, e em fevereiro desde as 3h do dia 2 até às 22h do dia 4, ocorreram falhas na conexão da API aos servidores do Twitter que não foram previamente detetadas, pelo que não foram recolhidos dados nestes períodos. Em fevereiro verificamos também uma grande falha nos dados, que ocorreu devido a um erro no código de pós-processamento que fez com que os dados desde as 17h do dia 10 até às 3h do dia 16 não fossem considerados.

Analisando os quatro meses de recolha, notamos que em janeiro existe uma subida em geral do número de *tweets* por dia, acentuando-se a meio do mês de janeiro e fevereiro.

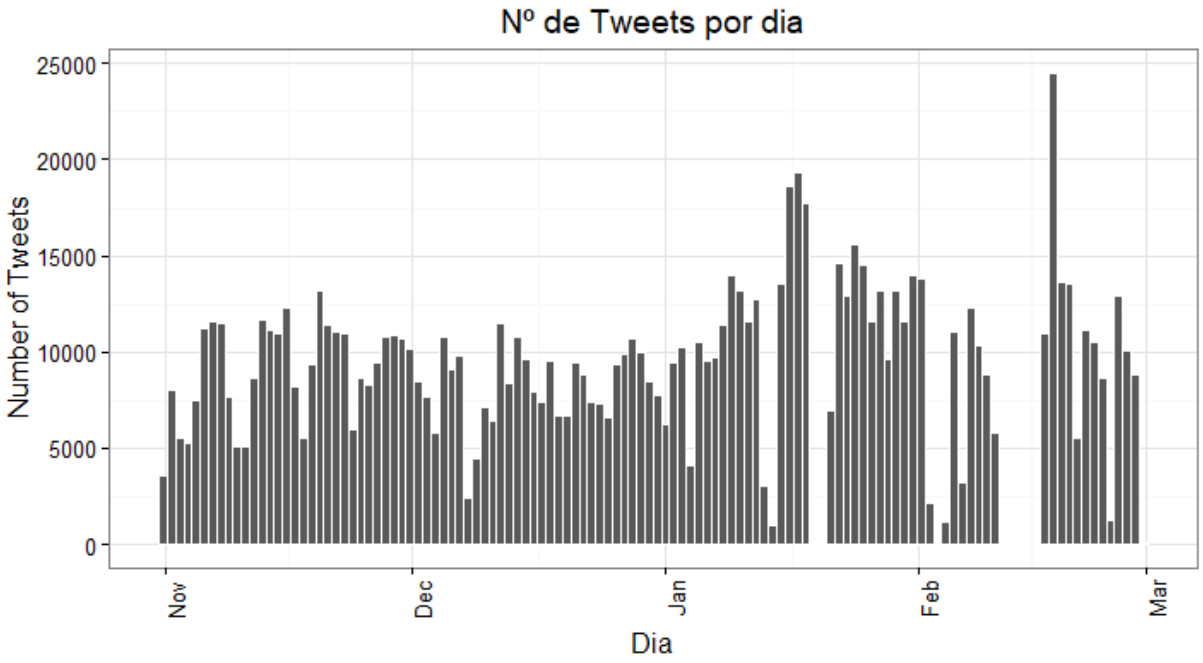

Figura 3 – Número de *tweets* por dia ao longo do estudo

Ainda sobre a análise da figura 3, verificamos que existem picos em determinada altura da semana, considerando sete dias. Analisando os restantes gráficos (figuras 4, 5, 6 e 7) que contém a distribuição dos *tweets* por dia mas detalhado ao mês, concluímos que os picos se situam durante a semana, levando a crer que durante o fim de semana, provavelmente por razões familiares e de lazer, os utilizadores não partilham tanto mensagens no Twitter sobre a gripe.

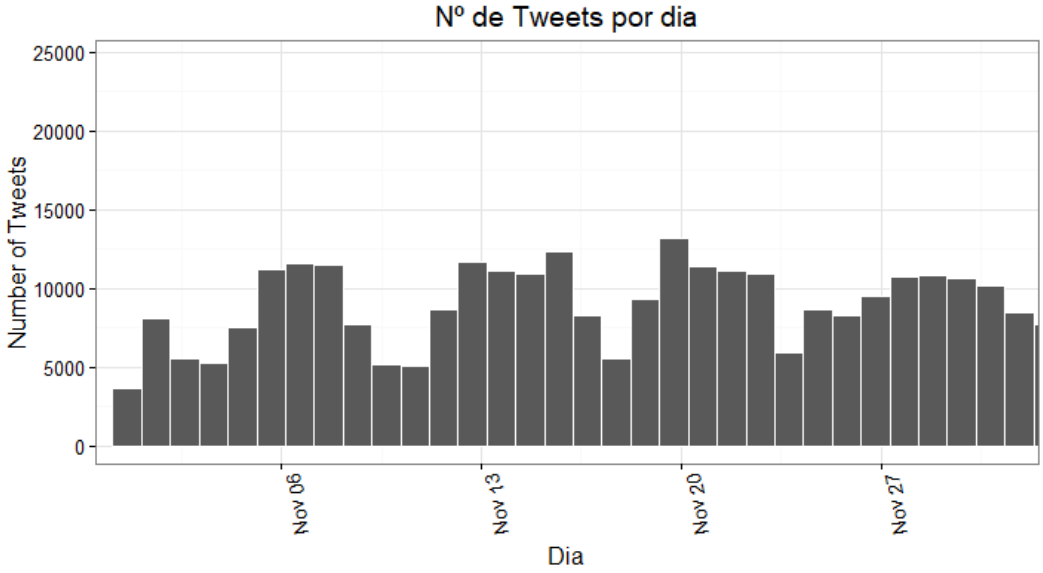

Figura 4 – Número de *tweets* por dia no mês de novembro

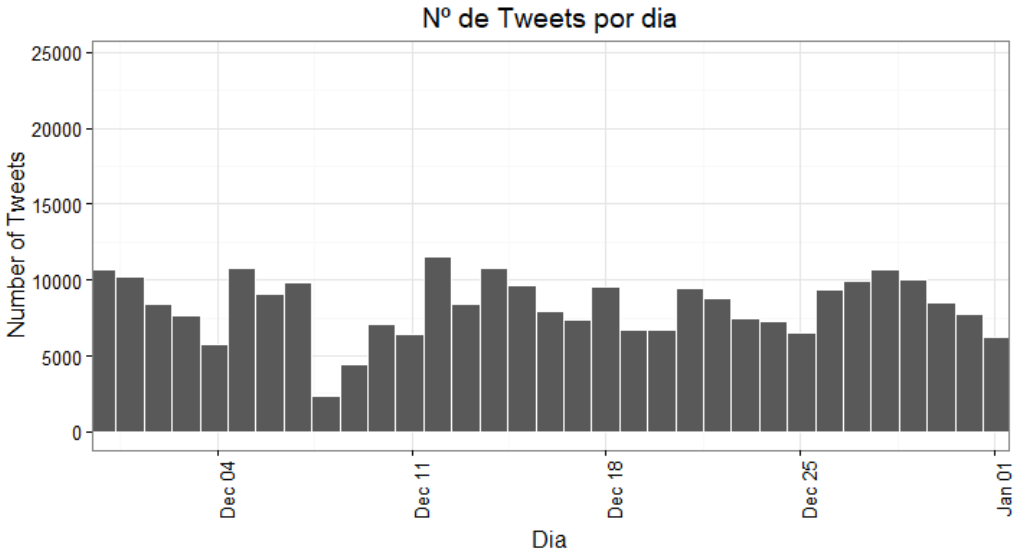

Figura 5 – Número de *tweets* por dia no mês de dezembro

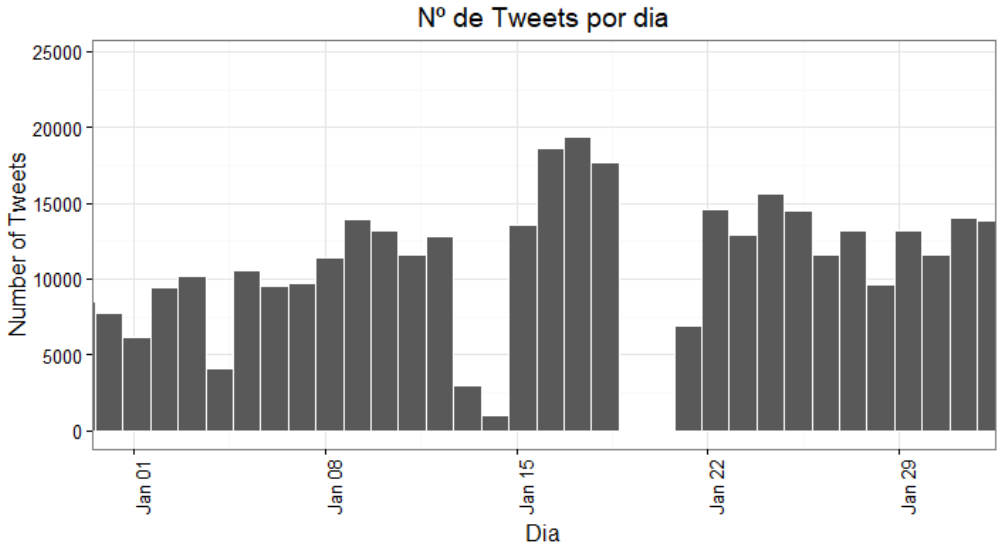

Figura 6 – Número de *tweets* por dia no mês de janeiro

É no mês de fevereiro que é atingido o máximo de número de *tweets* num só dia, chegando perto do 25.000 no dia 17 de fevereiro.

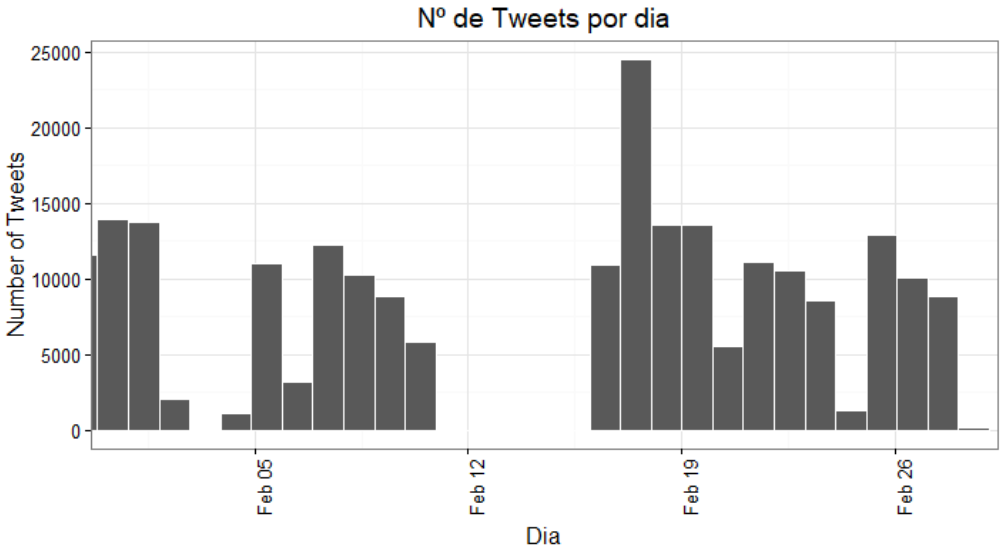

Figura 7 – Número de *tweets* por dia no mês de fevereiro

Criámos também uma nuvem de palavras para obter de uma forma visual as palavras com maior frequência em todo o *dateset,* sendo estas as palavras que se apresentam com maior tamanho na nuvem. Concluímos que as palavras "febr", "gripe" são as que mais se destacam seguidas pelas palavras "dor", "garganta", "influenza", "sinusite", "pra" e "toss". De notar que o *dataset* neste ponto do estudo já foi pré-processado, ou seja, tem o *stemming* aplicado daí termos as palavras "febr" e "toss" sem as respetivas flexões. Para além disso foram removidas as *stopwords*, no entanto temos na nuvem a palavra "pra", que é a forma reduzida da preposição "para" mas não retirada pelo pré-processamento.

Figura 8 – Nuvem de palavras em estudo

3.6.2. Dados oficiais

Sobre os dados oficiais, podemos ver a distribuição da taxa de incidência gripal ao longo de todas as semanas em estudo no gráfico abaixo. Conseguimos verificar que existem alguns picos, sendo o maior na semana 8, que corresponde à penúltima semana do ano de 2017. A partir da oitava semana vemos a taxa de incidência gripal a descer, com duas ligeiras subidas pelo meio, até à última semana do estudo. A menor taxa situa-se nas primeiras semanas de estudo e corresponde ao início do mês de novembro.

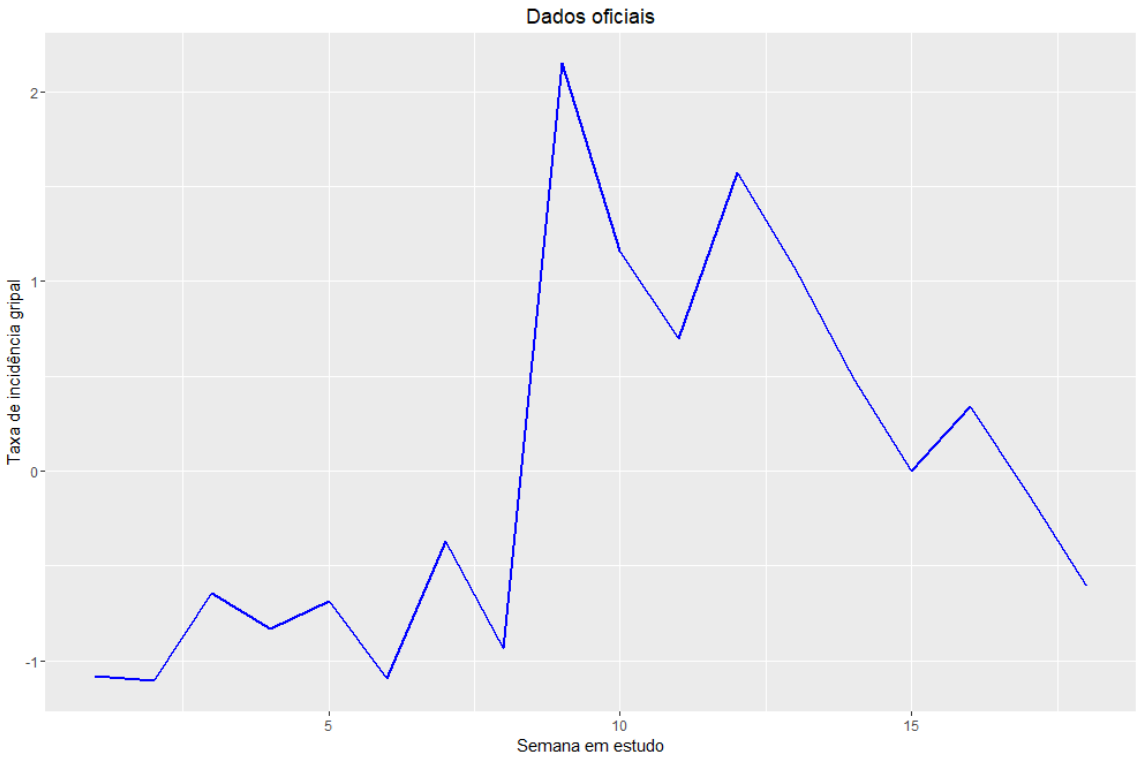

Figura 9 – Taxa de incidência gripal - dados oficiais

3.7. MATRIZ DE CORRELAÇÃO

Para perceber qual a relação entre as variáveis sem considerar as variáveis dependentes utilizámos uma matriz de correlação que mede o coeficiente de *Pearson*, possibilitando perceber que variáveis estão altamente correlacionadas. Esta análise é importante devido à colinearidade que existe quando duas variáveis estão altamente relacionadas. A presença de colinearidade é um problema pois torna-se complicado perceber qual o papel de cada uma das variáveis independentes para com a variável dependente, reduzindo a precisão dos modelos (James, Witten, Tibshirani, & Hastie, 2013).

Verificando a figura abaixo concluímos que as palavras “nariz” e “entupido” estão muito próximas de um coeficiente igual a 1, tal como as palavras “garganta” e “dor”. Posto isto decidimos retirar da matriz *term-by-document* as palavras “nariz” e “garganta”, pois as mesmas são explicadas pelas outras duas. Isto significa que se olharmos para as colunas das variáveis altamente correlacionadas, as mesmas vão ser muito semelhantes, fazendo com que o retirar de um das duas

não tenha consequências nos resultados para além de abatermos a hipótese de colinearidade. Assim reduzimos também dimensionalidade da matriz tornando o processo do treino dos dados menos pesado e facilitando a posterior análise.

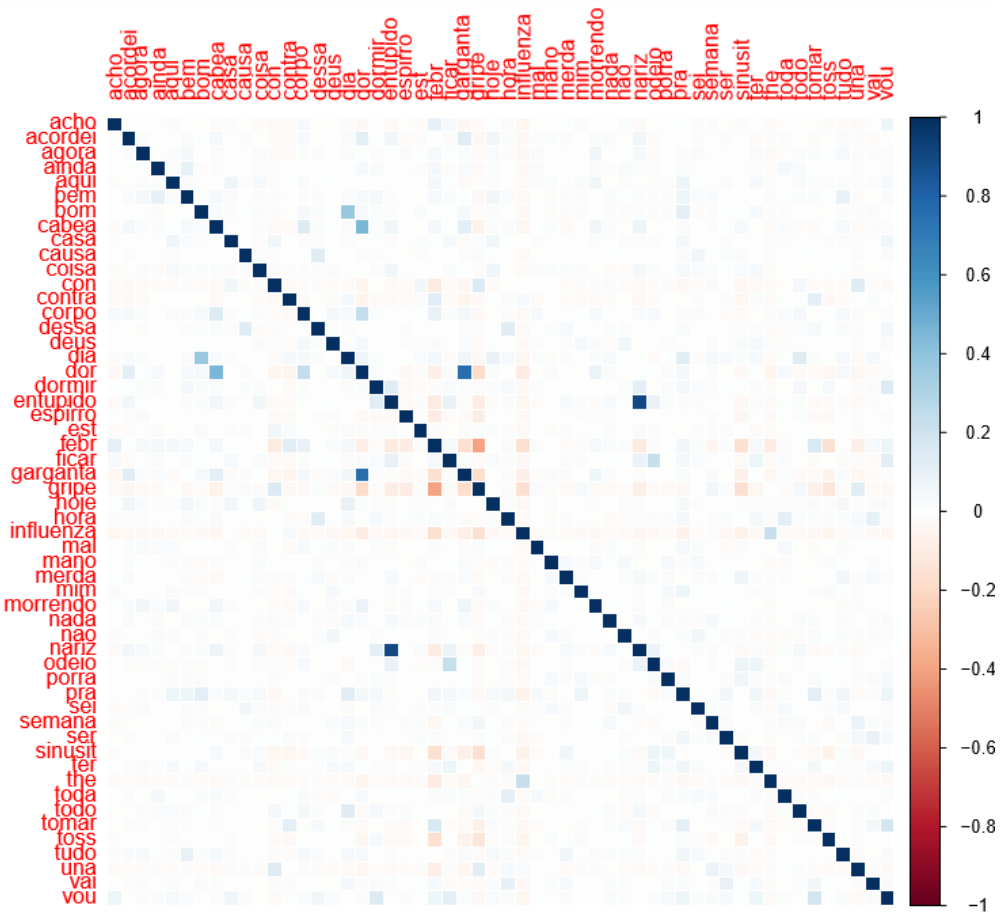

Figura 10 – Matriz de correlação das variáveis em estudo

3.8. CLASSIFICAÇÃO

Usar um *dataset* com as características do definido até aqui (filtragem e pré-processamento de dados) não é suficiente para criar um modelo preditivo. Por isso, este capítulo foca-se essencialmente na aplicação de técnicas de *machine learning* para classificar os *tweets*.

Esta classificação é de extrema importância pois, apesar dos *tweets* conterem pelo menos uma das palavras-chave definidas no início da metodologia, o mesmo pode não estar relacionado como objeto em estudo. Vejamos os exemplos:

"Depois do jogo de futebol fiquei com uma dor muscular na perna."

e

"Não se aguenta pessoas com a febre pelo Justin Bieber."

e

"Espero que a gripe não me ataque este inverno."

Estes *tweets* contêm de facto palavras-chave, no entanto os dois primeiros não se referem à gripe, e o terceiro refere-se a gripe mas a pessoa que o publicou não está doente.

A classificação está então dividida em duas partes: a primeira onde um conjunto aleatório de 10.000 *tweets* é classificado manualmente por oito pessoa diferentes, constituindo o conjunto de treino; a segunda onde o restante *dataset* é classificado por um algoritmo de classificação. Os 10.000 *tweets* foram classificados por diferentes pessoas para o processo ser mais rápido e para retirar o enviesamento de interpretação por uma só pessoa.

3.8.1. Criação do treino

Como é inviável classificar manualmente todos os *tweets* recolhidos, iremos utilizar os 10.000 *tweets* classificados manualmente para ensinar o modelo a classificar os restantes 1.050.923. Assim, no fim do processo de classificação, iremos obter todo o *dataset* classificado como sendo positivo, caso o evento gripe se verifique, ou negativo caso contrário. Para além desta classificação, será feita outra baseada em Paul & Dredze (2012) com cinco classes: doente, saúde, não relacionado, não português e ambíguo. A descrição de cada classe destas duas taxonomias encontra-se nas tabelas 4 e 5.

Categoria	Descrição
Positivo	Tweet contendo acontecimentos sobre a gripe, do próprio ou de terceiros
Negativo	Tweet sem qualquer conteúdo relacionado com gripe, ou relacionado e em formato de notícia

Tabela 4 – Categorias da taxonomia de duas classes e sua definição

Categoria	Descrição
Doente	Tweet contendo acontecimentos do próprio indivíduo sobre a gripe
Saúde	Tweet contendo acontecimentos de terceiros sobre a gripe
Não relacionado	Tweet sem qualquer relação a gripe
Não português	Tweet escrito em língua que não a portuguesa
Ambíguo	Tweet do qual nada se consegue concluir

Tabela 5 – Categorias da taxonomia de cinco classes e sua definição

A utilização de duas taxonomias deve-se ao facto de perceber qual delas se adequa ao evento em estudo, sendo que esta comparação não foi feita em nenhuma obra das citadas neste trabalho. Para além pretendemos perceber se existem diferenças ou melhorias significativas pelo uso de taxonomias diferentes.

3.8.2. Algoritmo de classificação

Para que todo o *dataset* de mais de 1 milhão de *tweets* seja classificado automaticamente é necessário um algoritmo capaz de lidar com um grande volume de dados. Neste trabalho de projeto foram testados três algoritmos, *Random Forest*, (*RF*) *Naïve Bayes (NB)* e Regressão Logística, no entanto foram obtidos melhores resultados no RF para ambas as classes, pelo que será este algoritmo o descrito e analisado mais detalhadamente.

3.8.2.1. Random Forest

Neste estudo foram construídas 100 árvores de decisão de treino para cada RF, uma considerando a taxonomia de duas classes e a outra considerando a taxonomia de cinco classes. Foi testado para cada RF qual o número de variáveis selecionadas aleatoriamente que nos dá maior precisão no modelo. Concluímos então que para a RF de duas classes o valor *default* dado pelo algoritmo é o melhor, sendo que em cada nó o algoritmo irá selecionar 7 variáveis aleatórias para escolher a que melhor reparte os dados. Para a RF de cinco classes concluímos que 6 variáveis permitem obter uma melhor precisão no modelo.

4. ANÁLISE DE RESULTADOS

4.1. ALGORITMO DE CLASSIFICAÇÃO

As matrizes de confusão abaixo comparam os valoreis reais com os valores que foram previstos pelo algoritmo para duas e cinco classes, respetivamente, dando-nos o erro de classificação em cada classe:

		Valores previstos		
		0 - Negativo	1 - Positivo	Erro de classificação
Valores amostrais	0 - Negativo	3537	1983	0,3592
	1 - Positivo	1083	3390	0,2421

Tabela 6 – Matriz de confusão para a RF de duas classes

		Valores previstos					
		1 - Doente	2 - Saúde	3 - Não relacionado	4 - Não português	5 - Ambíguo	Erro de classificação
Valores amostrais	1 - Doente	3193	2	160	660	12	0,2071
	2 - Saúde	549	0	88	81	3	1
	3 - Não relacionado	1043	4	528	466	8	0,7423
	4 - Não português	108	0	26	2102	0	0,0599
	5 - Ambíguo	698	1	101	160	0	1

Tabela 7 – Matriz de confusão para a RF de cinco classes

O algoritmo usando duas classes tem uma precisão de 69,32% mostrando que se comporta melhor a prever valores positivos. O algoritmo usando cinco classes tem uma precisão de 58,27%, mostrando que se comporta melhor a prever as classes “doente” e “não português” mas que não consegue prever as classes “saúde” e “ambíguo”, pois erram em 100% dos casos.

Para além da precisão do modelo, existem outras medidas para medir a performance do mesmo, é o caso do *F-score*. Esta medida torna-se extremamente importante quando no *dataset* já classificado não existe uma divisão igual do número de registos por classes (Hotho, Nürnberger, & Paaß, 2005), que o nosso estudo tem em consideração mesmo não sendo um caso extremo (a amostra de 10.000 registos tem 55% de registos negativos e 45% de registos positivos). Como o processo de classificação assume que a distribuição é a mesma do *dataset* de treino, quando estamos na presença de *datasets* não balanceados podemos obter resultados enviesados (Hotho, Nürnberger, & Paaß, 2005; Khalilia et al 2011). Por exemplo, assumindo que nos 10.000 registos pré-classificados temos 1.000 de classe negativa e 9000 de classe positiva e que a matriz de confusão nos mostra que apenas 200 são registos previstos como negativos e 8.800 previstos como positivos. Com este exemplo obteríamos uma precisão de 90%, no entanto a classe negativa tem um erro de 80%.

Esta medida tem em conta *precision*, que quantifica o rácio das registos que são de efetivamente relevantes, ou seja o rácio entre os registos corretamente previstos como positivos e o total de registos previstos como positivos, e *recall*, que quantifica o rácio dos registos que são efetivamente relevantes e que são selecionados, ou seja o rácio entre os registos corretamente previstos como positivos e o total de registos positivos na amostra (Hotho, Nürnberger, & Paaß, 2005). Neste estudo e tendo em conta a RF de duas classes, obtemos um F-score de 68,86%, que não é muito distante da medida de precisão de 69,32%.

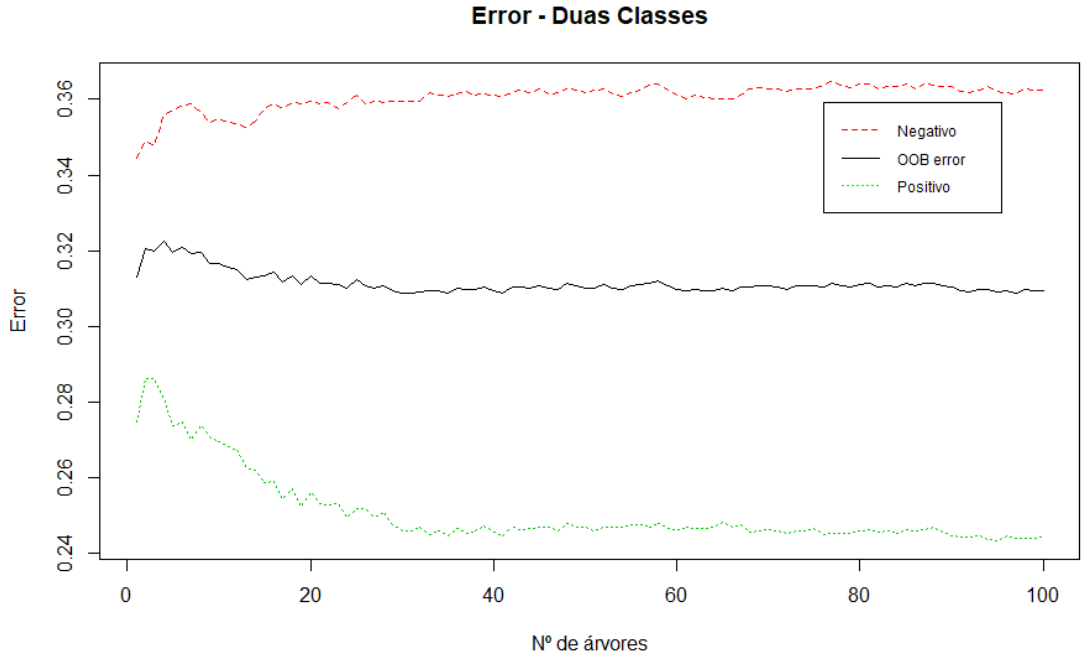

Figura 11 – Distribuição do erro para duas classes

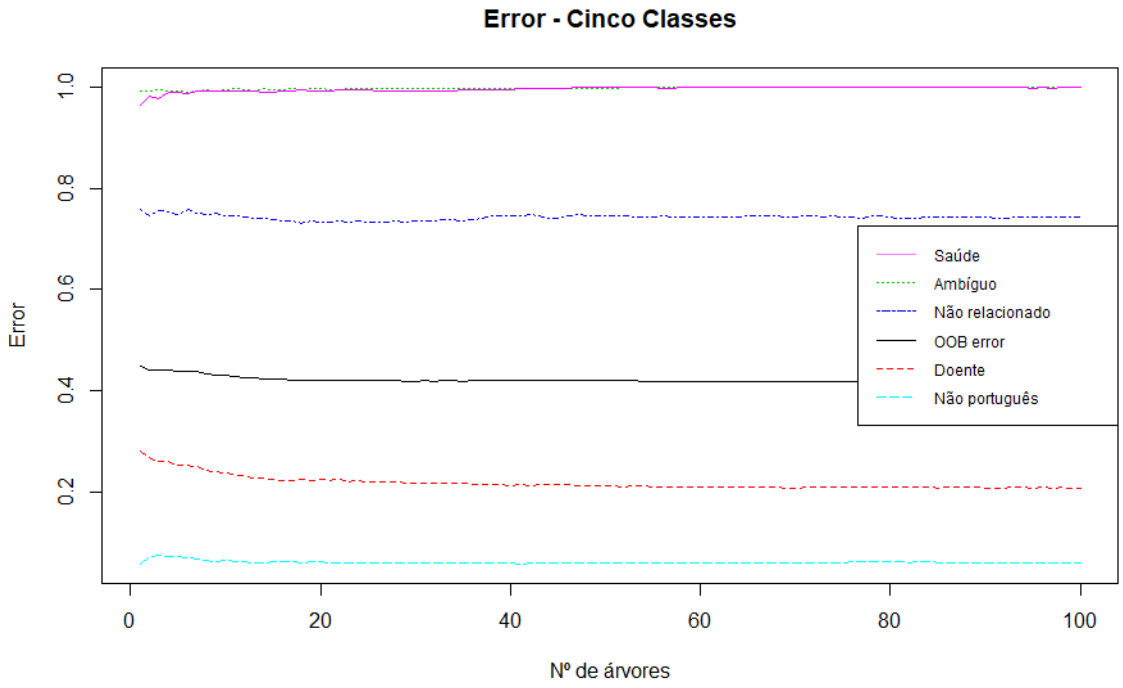

Figura 12 – Distribuição do erro para cinco classes

Pelas figuras 11 e 12 podemos no início o erro *OOB* era maior, no entanto, à medida que o número de árvores aumenta o erro baixa até estagnar entre as 50 e 60 árvores no caso das duas classes, e até estagnar entre as 5 e 10 árvores no caso das cinco classes. Este gráfico mostra que as 100 árvores de decisão usadas foram suficientes para treinar o modelo e para além disso mostra que o algoritmo não faz *overfit.* Tal como vimos no estudo de Breiman, isto deve-se à Lei dos Grandes Números que diz que se a mesma experiência for repetida várias vezes esta tende a aproximar-se do valor real (Breiman, 2001).

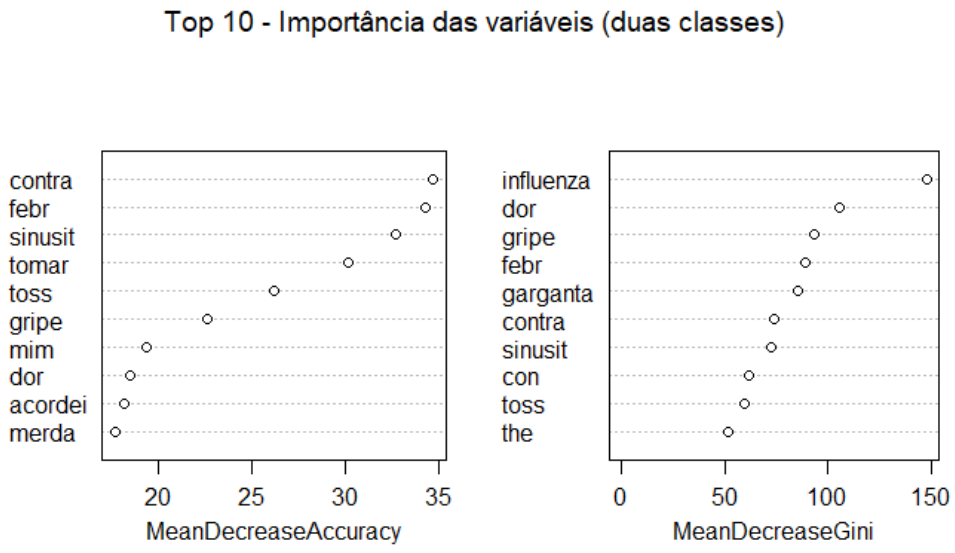

Figura 13 – Top 10 das variáveis mais importantes na RF de duas classes

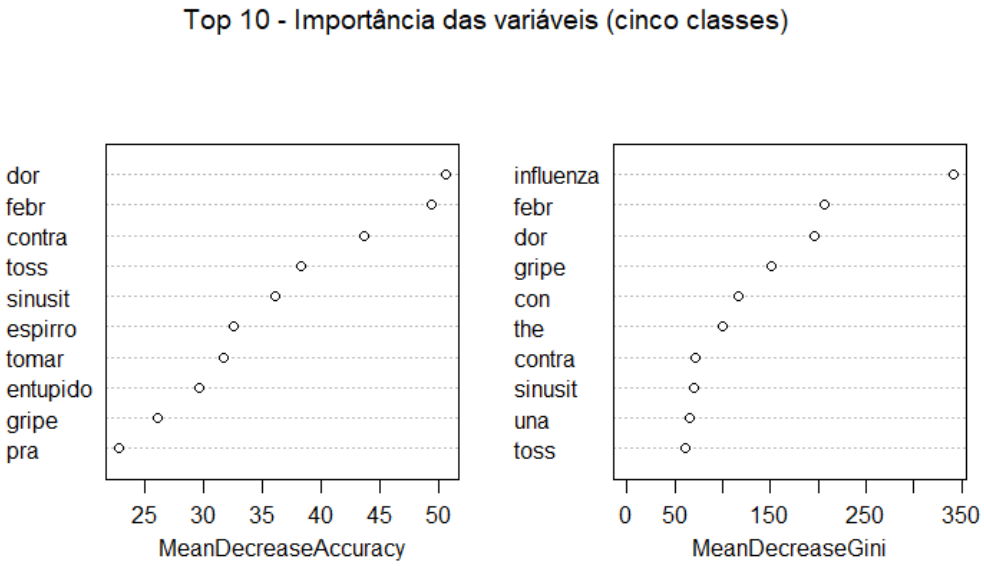

Figura 14 – Top 10 das variáveis mais importantes na RF de cinco classes

Uma das vantagens que RF nos dá é o cálculo da importância das variáveis, que mede o quão importante e relacionada está cada variável com os resultados da classificação dados pelo modelo. O cálculo do *mean decrease accuracy* é feito através de permutações nos dados *OOB*. Os registos *OOB*, ao percorrerem cada árvore, calculam a taxa de erro, seguida da permutação dos valores de cada variável preditiva. A taxa de erro é de novo calculada e obtida a média da diferença entre as duas taxas tendo em conta todas as árvores (Breiman, L., & Cutler, A., 2018; Khalilia et al, 2011). O cálculo do *mean decrease gini* é feito através do *Gini Index*, e é obtido através da média entre todas as árvores da diminuição da impureza no nó devido à divisão pela variável (Breiman, L., & Cutler, A., 2018). Nas figuras 13 e 14 podemos ver as dez variáveis que maior importância têm no modelo. Os gráficos "*MeanDecreaseAccuracy"* mostram que variáveis são mais importantes para a precisão do modelo e os gráficos "*MeanDecreaseGini"* mostram as variáveis que mais puramente dividem as folhas das árvores tendo em conta as classes em estudo. Por exemplo, considerando o "*MeanDecreaseAccuracy"* na RF de cinco classes, se retirássemos a variável "dor" do modelo, o mesmo veria a sua precisão descer em 50%.

4.2. COMPARAÇÃO DOS DADOS

Neste subcapítulo serão comparados os dados oficiais com os dados extraídos do Twitter tentando perceber se existe relação entre eles, ou seja, se um número maior de *tweets* relacionados com a gripe se relaciona com uma maior incidência da doença ou vice-versa.

O *dataset* dos dados oficiais dá-nos a incidência gripal por semana, pelo que tivemos de agregar os nossos dados do Twitter pelas 18 semanas em estudo e obter a média na classe em cada uma delas. Foi então aplicada a mesma escala aos dados previstos e aos dados oficiais e colocados os dois *datasets* num gráfico para comparação visual.

Em relação à taxonomia de duas classes obtemos o seguinte:

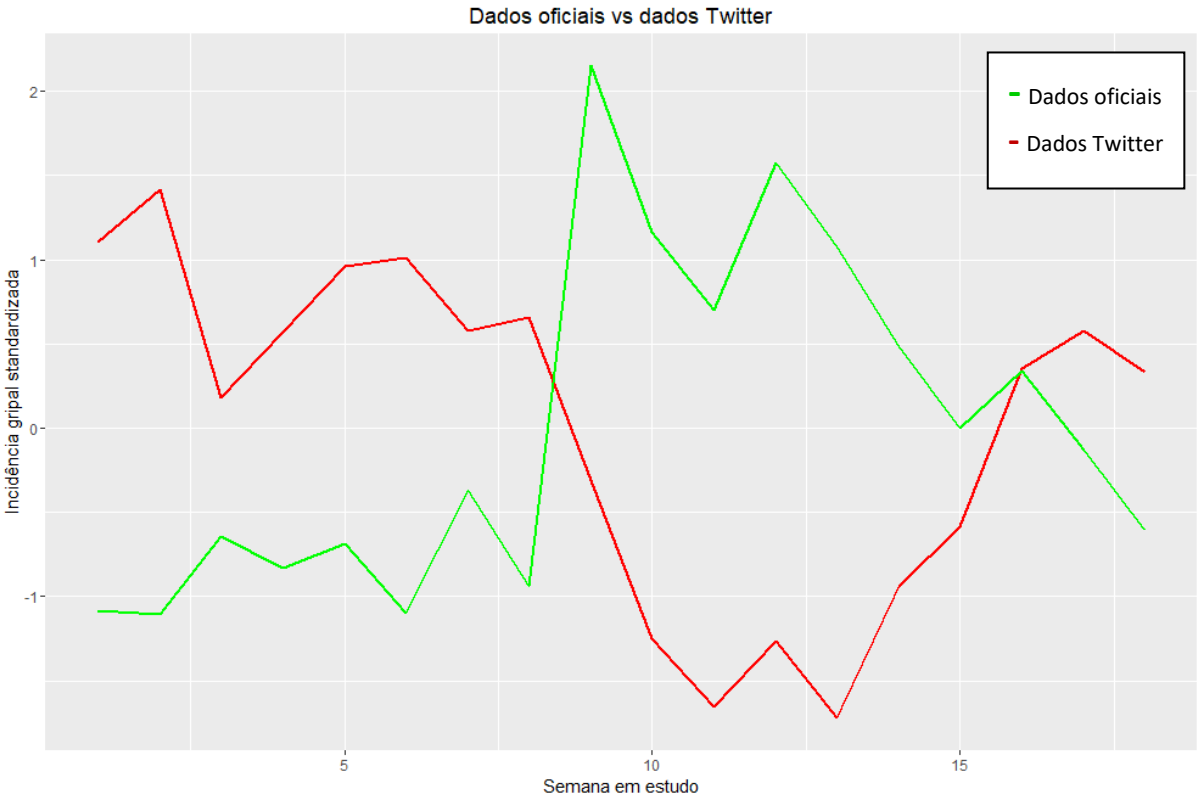

Figura 15 – Dados oficiais vs dados Twitter, modelo de duas classes

Visualmente, verificamos que pode existir uma relação entre os dois *dataset*, considerando um certo desfasamento no tempo. Se por um lado os dados aparentam estar inversamente correlacionados com um *lag* = 0 (isto é, sem aplicar desfasamento numa das séries de dados) por outro, os picos parecem durar aproximadamente o mesmo número de semanas nas duas séries. Assim iremos aplicar *lag* aos dados do Twitter, fazendo com que os mesmos avancem no tempo. Os gráficos obtidos com os diferentes *lag* podem ser vistos no anexo 8.2.

Para obter algo mais preciso do que apenas algo visual, foi aplicada uma regressão linear sobre os dados e obtidos os seguintes resultados:

Modelo	Coeficiente de determinação (R^2)	Rank
Lag = 0	0.6448	1
Lag = 1	0.3624	3
Lag = 2	0.1982	5
Lag = 3	0.02587	10
Lag = 4	0.009157	11
Lag = 5	0.05633	8

Lag = 6	0.04197	9
Lag = 7	0.2782	4
Lag = 8	0.5249	2
Lag = 9	0.07448	7
Lag = 10	0.08894	6

Tabela 8 – Estatísticas das regressões lineares com e sem *lag* para o modelo de duas classes

Analisando a tabela acima, conseguimos perceber que o maior coeficiente de determinação se obtém quando não aplicamos *lag* aos dados. Analisando o gráfico, podemos concluir que tal se deve ao facto de que quando os dados oficiais aumentam a taxa de incidência, os dados do Twitter diminuem a sua taxa e vice-versa. Se considerarmos o *lag* = 8, ou seja, avançamos os dados do Twitter oito semanas, obtemos também um grande coeficiente de determinação, mostrando que é este o segundo melhor modelo que ajusta a relação de previsão entre os dois *datasets*. Um coeficiente próximo de 1 indica que existe uma grande proporção de variância explicada, logo, que os dois modelos se correlacionam (James, Witten, Tibshirani, & Hastie, 2013).

Em relação às cinco classes, e considerando que apenas as classes "doente" e "saúde" podem ser consideradas relacionadas com a gripe, iremos considerar cada classe como um *dataset* distinto. Assim, tomando como exemplo a primeira classe, sempre que o registo for classificado como "doente" iremos assignar o valor 1 e caso contrário o valor 0, tornando-se um modelo de duas classes e sendo assim possível de ser aplicada a regressão linear.

Observando o gráfico de relação entre os dados oficiais e os dados da classe "doente", verificamos que é bastante semelhante ao anterior gráfico de relação entre os dados oficiais e os dados de duas classes, pelo que também se pode verificar uma relação entre os dois *dataset*, considerando o desfasamento no tempo. Os gráficos obtidos com os diferentes *lag* podem ser vistos no anexo 8.2.

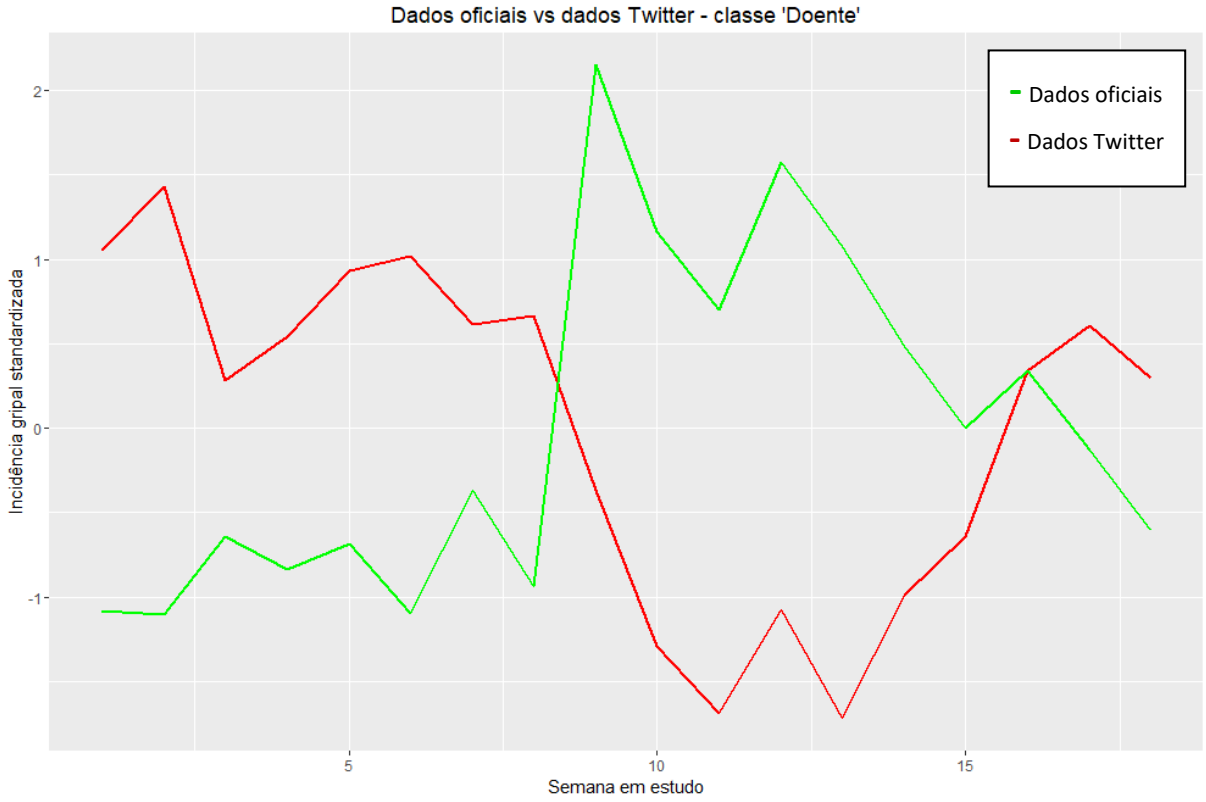

Figura 16 – Dados oficiais vs dados Twitter, classe "doente" do modelo de cinco classes

Aplicada a regressão linear obtemos os seguintes resultados:

Modelo	Coeficiente de determinação (R^2)	Rank
Lag = 0	0.6399	1
Lag = 1	0.3626	3
Lag = 2	0.1979	5
Lag = 3	0.02794	10
Lag = 4	0.008891	11
Lag = 5	0.05143	8
Lag = 6	0.045	9
Lag = 7	0.2941	4
Lag = 8	0.5039	2
Lag = 9	0.09799	7

Lag = 10	0.1218	6

Tabela 9 – Estatísticas das regressões lineares com e sem *lag* para o modelo de cinco classes, classe "doente"

Tal como no modelo de duas classes, obtemos um maior coeficiente de determinação com *lag* = 0, sendo que o segundo melhor modelo seria obtido avançando os dados do Twitter sete semanas.

Por último, e analisando o gráfico de relação entre os dados oficiais e os dados da classe "saúde", verificamos que não é tão fácil afirmar que existe uma correlação dos *datasets* em toda a extensão temporal. Teremos então de analisar os diferentes desfasamentos temporais. Os gráficos obtidos com os diferentes *lag* podem ser vistos no anexo 8.2.

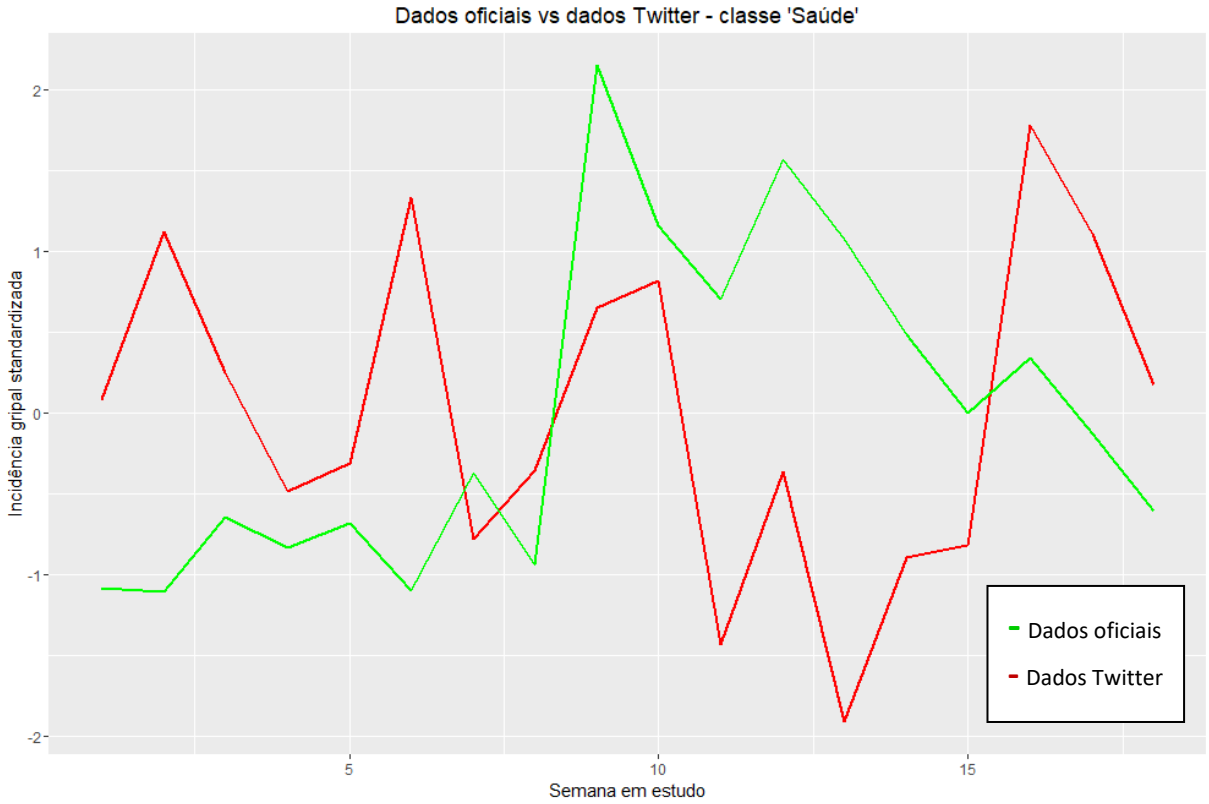

Figura 17 – Dados oficiais vs dados Twitter, classe "saúde" do modelo de cinco classes

Aplicada a regressão linear obtemos os seguintes resultados:

Modelo	Coeficiente de determinação (R^2)	Rank
Lag = 0	0.03276	8
Lag = 1	0.06883	5

Lag = 2	0.1083	3
Lag = 3	0.04877	4
Lag = 4	0.01003	10
Lag = 5	0.01364	9
Lag = 6	0.0005818	11
Lag = 7	0.1808	2
Lag = 8	0.03808	6
Lag = 9	0.03642	7
Lag = 10	0.3693	1

Tabela 10 – Estatísticas das regressões lineares com e sem *lag* para o modelo de cinco classes, classe "saúde"

Os resultados do coeficiente de determinação neste caso não nos dão grande evidência de que os *datasets* estejam correlacionados, visto que o máximo que obtemos é de 0.3693 com um *lag* = 10, ou seja avançando os dados do Twitter dez semanas.

5. CONCLUSÕES

Analisando os resultados obtidos no capítulo anterior podemos concluir que existe de facto uma capacidade de previsão da taxa de incidência gripal através dos dados do Twitter, estando esta previsão dependente do tipo de classificação feita aos dados e do desfasamento temporal de um dos *datasets.*

Considerando a taxonomia de duas classes, onde obtivemos uma boa precisão de classificação do modelo de 69,32%, verificamos que o coeficiente de determinação varia bastante ao longo dos diferentes *lags* aplicados. O coeficiente começa com um *lag* bastante elevado, vai diminuindo até às quatro semanas de avanço temporal, voltamos depois a aumentar até ao *lag* = 8 e diminuindo nos restantes. Encontramos aqui dois picos de correlação, um sem qualquer desfasamento, outro com um avanço de oito semanas dos dados do Twitter. Não podemos, no entanto, afirmar que a taxa de incidência gripal possa ser prevista oito semanas antes de ser detetada oficialmente pois a gripe cura-se em uma ou duas semanas (George, 2006), o que torna o desfasamento de oito semanas exagerado. Se considerarmos o caso onde obtemos maior coeficiente de determinação, que é sem qualquer avanço temporal, podemos concluir que quando a taxa de incidência gripal oficial aumenta, a taxa de incidência gripal calculada através dos dados do Twitter diminui. Esta é uma conclusão plausível se considerarmos que quando existe um pico da gripe, as pessoas acabam por andar mais doentes e não ligar tanto às redes sociais nem estão tão ativas a publicar mensagens no Twitter. Podemos também suspeitar que quando se dá a maior incidência gripal é quando existem casos de maior gravidade, fazendo com que os dados do Twitter não acompanhem a subida da taxa oficial. Estas conclusões podem ser extrapoladas para a análise à classe "Doente" da taxonomia de cinco classes, visto apresentar valores bastante semelhantes à taxonomia de duas classes.

Considerando a classe "Saúde" na taxonomia de cinco classes, verificamos que o coeficiente de determinação varia muito menos, sendo que o seu máximo não alcança os 0.4. Concluímos então que não existe evidência de que os dados do Twitter consigam prever a taxa de incidência gripal considerando apenas esta classe.

O modelo de cinco classes obteve uma precisão de apenas 58,27%, mas estando nós a trabalhar apenas com duas das classes, podemos considerar a precisão individual de cada uma. Assim podemos afirmar que a classe "Doente" tem uma precisão de 79,29% (considerando o erro de classe de 20,71%) e a classe "Sáude" tem uma precisão de 0% (considerando o erro de classe de 100%). Tendo em conta estes valores, podemos afirmar que tanto o modelo de duas classes como o modelo de cinco considerando a classe "Doente" são bons modelos para classificação, com boas medidas de precisão. Sobre o modelo de cinco classes considerando a classe "Saúde" podemos afirmar sem dúvida alguma que não é um bom modelo, visto que erra a classificação dos dados em 100% dos casos. Este resultado pode ser devido ao enviesamento e pouca coerência na classificação manual feita pelas oito pessoas, pois foi a classe que mais dúvidas suscitou ao longo do processo de classificação.

De um modo geral, a aplicação das diferentes taxonomias não nos trouxe melhorias ao estudo, visto que os resultados da classe "Doente" são muito semelhantes à taxonomia de duas

classes e os resultados da classe “Saúde” não corroboram a hipótese inicial do estudo dos dados do Twitter preverem a taxa de incidência gripal.

Acerca dos objetivos propostos inicialmente neste trabalho, os mesmos foram alcançados. Decerto que no fim deste trabalho de projeto aprofundámos conhecimentos teóricos e práticos acerca da gripe, de *Text* e *Data Mining,* dos algoritmos estudados e das ferramentas utilizadas. Conseguimos também tirar algumas conclusões interessantes tendo em conta as diferentes taxonomias e desfasamentos temporais. Basta apenas identificar outras redes sociais nas quais seja possível aplicar a mesma metodologia.

Segundo a Agência Lusa (2016) as redes sociais mais utilizadas em Portugal são Facebook, Youtube, Google+, LinkedIn, Instagram e Twitter. Analisando o Instagram, podemos recolher os dados da mesma forma, através de uma API, no entanto, esta rede social é bastante utilizada para partilha de fotos. Poderíamos analisar a descrição de cada foto mas a rede social, sendo o seu foco a fotografia, não é tanto usada para partilha de sentimentos do dia-a-dia como o Twitter. O LinkedIn é uma rede social virada para o mundo profissional e não tanto para partilha da parte pessoal pelo que não seria tão interessante realizar um estudo sobre esta rede social. O Google+ prende-se muito pela partilha de qualquer conteúdo principalmente fotografias e *links* para outras páginas, como notícias ou artigos sobre os mais variados temas. O Youtube é uma rede social de partilha de vídeos, pelo que a nossa metodologia de *Text Mining* não se consegue aplicar ao mesmo. O Facebook, é bastante utilizado para partilha de outro tipo de conteúdo que não o textual. Para além disso, as publicações textuais não têm qualquer limitação do número de caracteres, o que iria dificultar ainda mais o pré-processamento de dados e seria necessária uma grande capacidade computacional para tratamento dos dados. Posto isto, concluímos que o Twitter é a melhor rede social para aplicar técnicas de *Text Mining* pois é de facto bastante utilizado para partilha de publicações escritas. Para além do Twitter, seria interessante utilizar o Facebook, no entanto e devido às justificações dadas anteriormente, seria necessário um grande trabalho e capacidade computacional para o tratamento de dados.

Por último e comparando o nosso estudo com alguns dos estudos referidos na revisão de literatura, como o estudo de Szomszor, Kostkova, & De Quincey, 2011, onde os autores mostram que o surto de gripe suína de 2009 podia ter sido previsto uma semana antes, podemos afirmar que a nossa aposta no desfasamento de oito semanas, apesar de excessiva, é um resultado interessante, tendo em conta que estamos a trabalhar com dados estatísticos. Para além disso, se considerarmos o *lag* = 1, obtemos o terceiro maior coeficiente de determinação o que também é um resultado interessante.

Analisando outros estudos que concluíram que existe correlação entre os dados do Twitter e a taxa de incidência gripal (Corley et al., 2009), doenças que afetam a saúde pública (Paul & Dredze, 2012), e sismos (Sakaki et al., 2010) podemos afirmar que o Twitter é realmente uma ferramenta através da qual se consegue monitorizar em tempo real um dado evento. Podemos então considerar que obtemos resultados válidos no nosso estudo se não considerarmos desfasamento temporal.

É importante salientar que esses resultados demonstram a aplicabilidade das redes sociais como complemento aos tradicionais mecanismos de vigilância epidemiológica, de maneira a tomar ações antecipadamente e reduzir os impactos sobre a população.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Ao longo do desenvolvimento deste trabalho de projeto fomos encontrando algumas dificuldades. Uma delas e a que mais influenciou no progresso foi o tratamento dos dados devido ao grande número de registos e a todas as inconsistências encontradas em diferentes análises feitas a diferentes amostras, o que fez com que o este processo se tornasse bastante iterativo e demorado. Para além disso a elevada volumetria exigiu uma grande capacidade computacional.

Também na recolha de dados encontrámos alguns impasses, como o facto de termos de monitorizar a API constantemente visto que de vez em quando a ligação aos servidores do Twitter falha. Duas destas falhas não foram prontamente detetadas o que levou a uma pequena falha nos dados recolhidos. No entanto a maior falha nos dados deu-se devido a um erro em código R, na preparação do *dataset* final para treino, que fez com que um dos muitos ficheiros de recolha de dados não fosse considerado. Este erro foi apenas detetado durante a análise de resultados.

Como os dados recolhidos foram filtrados previamente segundo alguns termos de pesquisa, é de esperar que diferentes termos nos tragam diferentes resultados. Para trabalhos futuros recomendamos uma análise mais cuidada e mais alargada dos termos de pesquisa de modo a obter melhores resultados acerca da gripe.

Recomendamos também a consideração da localização geográfica e não apenas da língua portuguesa, pois o nosso *dataset* obtido do Twitter contem muitos *tweets* brasileiros o que acaba por influenciar os resultados, visto que estamos a estudar apenas a taxa de incidência gripal portuguesa.

Como última recomendação deixamos a recolha e análise de dados considerando um ano inteiro e não apenas os meses de maior incidência, podendo assim ser possível verificar todas as variações da taxa de incidência gripal.

7. BIBLIOGRAFIA

Agência Lusa (2016, June 29). Uso das redes sociais em Portugal triplicou em sete anos, mas empresas utilizam-nas pouco. *Observador*. Retrieved from http://observador.pt/2016/06/29/uso-das-redes-sociais-em-portugal-triplicou-em-sete-anos-mas-empresas-utilizam-nas-pouco/

Barreto, A.M. (2011). Uma visão sobre a evolução da relação entre marcas e consumidores após a emergência da Web 2.0. Prisma.com, Porto, Portugal, n. 15, 2011. Retrived from http://revistas.ua.pt/index.php/prismacom/article/view/1088

Bernardo, I., & Henriques, R. (2014). A Era de um mercado social: A relação entre o Twitter e o Mercado Accionista.

Breiman, L. (1996). Out-of-Bag Estimation. Technical Report, 1–13. https://doi.org/10.1016/j.patcog.2009.05.010

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (2018). Breiman and Cutler's random forests for classification and regression. Retrieved from https://CRAN.R-project.org/package=randomForest

Centers for Disease Control and Prevention (2010, June 16). *The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010*. Retrieved from https://www.cdc.gov/h1n1flu/cdcresponse.htm

Corley, C. D., Mikler, A. R., Singh, K. P., & Cook, D. J. (2009). Monitoring Influenza Trends through Mining Social Media. International Conference on Bioinformatics and Computational Biology (BIOCOMP), 1–7. Retrieved from http://www.eecs.wsu.edu/~cook/pubs/biocomp09.pdf

Dixon, M. (1997). An overview of document mining technology. Computer Based Learning Unit, University of Leeds. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:An+Overview+of+Document+Mining+Technology#0

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25(5).* https://doi.org/10.18637/jss.v025.i05

Feinerer, I. (2018). Text Mining Package. Retrieved from https://cran.r-project.org/web/packages/tm/

George, F. (2006, February 21). Introdução ao estudo da gripe. *DGS.* Retrieved from http://www.dgs.pt/documentos-e-publicacoes/introducao-ao-estudo-da-gripe.aspx

Gerschenfeld, A. (2013, November 27). Pandemia de gripe de 2009 matou muito mais pessoas do que se pensava*. Público.* Retrieved from https://www.publico.pt/2013/11/27/ciencia/noticia/pandemia-de-gripe-de-2009-matou-muito-mais-pessoas-do-que-se-pensava-1614123

Guiomar, R., Costa, I., Cristovão, P., Pechirra, P., Rodrigues, A., & Nunes, B. (2015). Programa Nacional de Vigilância da Gripe: relatório da época 2014/2015. Instituto Nacional de Saúde Doutor Ricardo Jorge, IP. Retrived from http://hdl.handle.net/10400.18/3175

Guiomar, R., Pechirra, P., Cristovão, P., Costa, I., Conde, P., Rodrigues, A., Silva, S., Machado, A., & Nunes, B. (2016). Programa Nacional de Vigilância da Gripe: relatório da época 2015/2016. Instituto Nacional de Saúde Doutor Ricardo Jorge, IP. Retrived from http://hdl.handle.net/10400.18/4044

Pechirra, P., Cristovão, P., Costa, I., Conde, P., Guiomar, R., Rodrigues, A., Silva, S., Torres, A., & Machado, A. (2018). Programa Nacional de Vigilância da Gripe: relatório da época 2017/2018. Instituto Nacional de Saúde Doutor Ricardo Jorge, IP. Retrived from http://repositorio.insa.pt/handle/10400.18/5619

Huberman, B. A., Romero, D. M., & Wu, F. (2008, December 5). Social Networks that Matter: Twitter under the Microscope. *First Monday*, 14(1). https://doi.org/10.2139/ssrn.1313405

Hotho, A., Nürnberger, A., Paaß, G. (2005, May 13). A Brief Survey of Text Mining. *Ldv Forum,* 20(1), 19-62.

James, G., Witten, D., Tibshirani, R., & Hastie, T. (2013). An Introduction to Statistical Learning with Applications in R. 59-71. https://doi.org/10.1007/978-1-4614-7138-7

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making, 11(1). https://doi.org/10.1186/1472-6947-11-51

Koprinska, I., Poon, J., Clark, J., & Chan, J. (2007). Learning to classify e-mail. Information Sciences, 177(10), 2167–2187. https://doi.org/10.1016/j.ins.2006.12.005

Leal, A., Ferreira, J., P. & Carvalho, R. (2015, April). Repórter TVI na íntegra: caos nas urgências mesmo depois da gripe. Retrieved from http:// http://www.tvi24.iol.pt/sociedade/reportagem/reporter-tvi-na-integra-caos-nas-urgencias-mesmo-depois-da-gripe

Lerman, K., & Ghosh, R. (2010). Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 90–97. https://doi.org/10.1146/annurev.an.03.100174.001431

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. a. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, (May), 122–129. https://doi.org/citeulike-article-id:7044833

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of Empirical Methods in Natural Language Processing, (July), 79–86. https://doi.org/10.3115/1118693.1118704

Paul, M. J., & Dredze, M. (2012). A model for mining public health topics from Twitter. Health, 11(May 2009), 16. https://doi.org/10.1371/journal.pone.0083672

Recuero, R. (2005). Redes Sociais na Internet: Considerações Iniciais. *Ecompós*, 2.

Russell, M. A. (2013). Mining the social web (2nd ed.). Sebastopol, CA: O'Reilly Media, Inc.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. Proceedings of the 19th International Conference on World Wide Web, 851–860. https://doi.org/10.1145/1772690.1772777

Sonnier, P. (2013, August 15). *What is Digital Health? by Paul Sonnier* [Video file]. Retrieved from https://youtu.be/oqnmpg2JmzM

St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *BMJ*, 344, 1-3. https://doi.org/10.1136/bmj.e2353

Szomszor, M., Kostkova, P., & De Quincey, E. (2011). #Swineflu: Twitter predicts swine flu outbreak in 2009. In Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (Vol. 69 LNICST, pp. 18–26). https://doi.org/10.1007/978-3-642-23635-8_3

Tavares, P. (2017, January 5). Urgências crescem 10% no SNS e até 20% nos privados. Diário de Notícias. Retrieved from http://www.dn.pt/portugal/interior/urgencias-crescem-10-no-sns-e-ate-20-nos-privados-5586619.html

8. ANEXOS

8.1. SCRIPT PARA EXTRAÇÃO DE DADOS

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
import time

ckey = 'A7ohivnrOk7VfZzL76BiPjSo3'
csecret = 'X7h1W5U6IK8AvV0aakoWVlP2hGIZ5AmHUS4gORwMg6nnqd5mDJ'
atoken= '2944441234-ujwkducALDSNR0NQPjy27mPtXiDhvx2TIbKJbQs'
asecret= 'IpBPyYALHWaMhgwNs9Km5yVe0mgnEz1lMvEA0vcdTuofn'

class listener(StreamListener):

    def on_data(self, data):
        try:

            bef_tweet = data.split(',"text":',1)[0]
            tweet = data.split(',"text":',1)[1].split(',"source":',1)[0]
            aft_tweet = data.split(',"source":',1)[1]

            tcomma = tweet.replace(",", "")

            data_tcomma = bef_tweet + ',"text":' + tcomma + ',"source":' + aft_tweet

            #tirar vírgula da localização e colocar ponto e vírgula
            bef_location = data_tcomma.split(',"location":',1)[0]
            location = data_tcomma.split(',"location":',1)[1].split(',"url":',1)[0]
            aft_location = data_tcomma.split(',"url":',1)[1]

            lcomma = location.replace(",",";")

            data_lcomma = bef_location + ',"location":' + lcomma + ',"url":' + aft_location

            #tirar vítgula do place full_name e colocar ponto e vírgula
            place_null = data_lcomma.split(',"place":',1)[1].split(',"contributors":',1)[0]

            if place_null == 'null':
                data_wcomma = data_lcomma
            else:
                bef_place = data_lcomma.split(',"full_name":',1)[0]
                place = data_lcomma.split(',"full_name":',1)[1].split(',"country_code":',1)[0]
                aft_place = data_lcomma.split(',"country_code":',1)[1]

                wcomma = place.replace(",",";")

                data_wcomma = bef_place + ',"full_name":' + wcomma + ',"country_code":' +
aft_place
```

```
                print data_wcomma

                saveFile = open('RecolhaTweets3.json','a')
                saveFile.write(data_wcomma)
                saveFile.write('\n')
                saveFile.close()
                return True
            except BaseException, e:
                print 'failed ondata,',str(e)
                time.sleep(5)

        def on_error(self, status):
            print status

auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
twitterStream.filter(track=["gripe","constipacao","virose","sinusite","influenza",
                    "febre","espirro","calafrio","tosse","corrimento nasal","dor de garganta","nariz
entupido","congestionamento","expetoracao",

"antigripal","mucolitico","expetorante","bissolvon","mucosolvan","atossin","cegripe","ilvico","antigri
pine","cecrisina"])
```

8.2. GRÁFICOS DE LAG

8.2.1. Duas classes

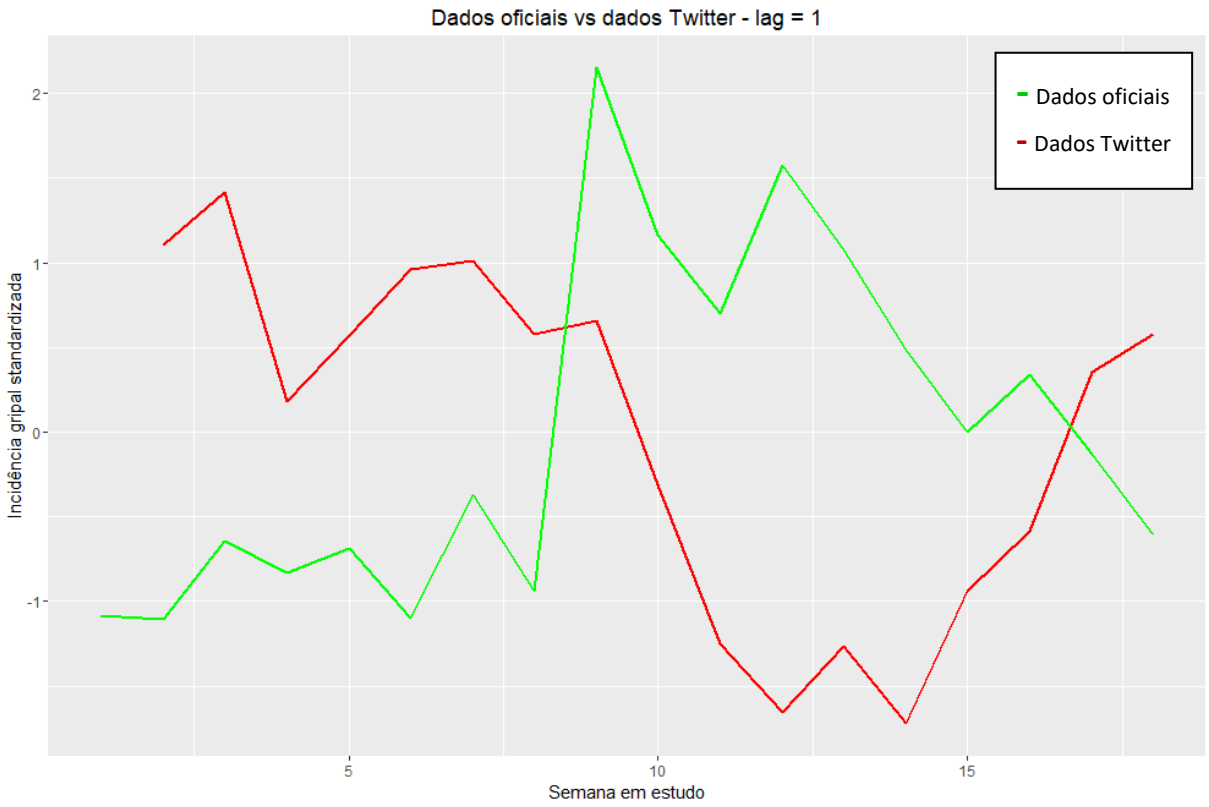

Figura 18 – Dados oficiais vs dados Twitter com *lag* = 1

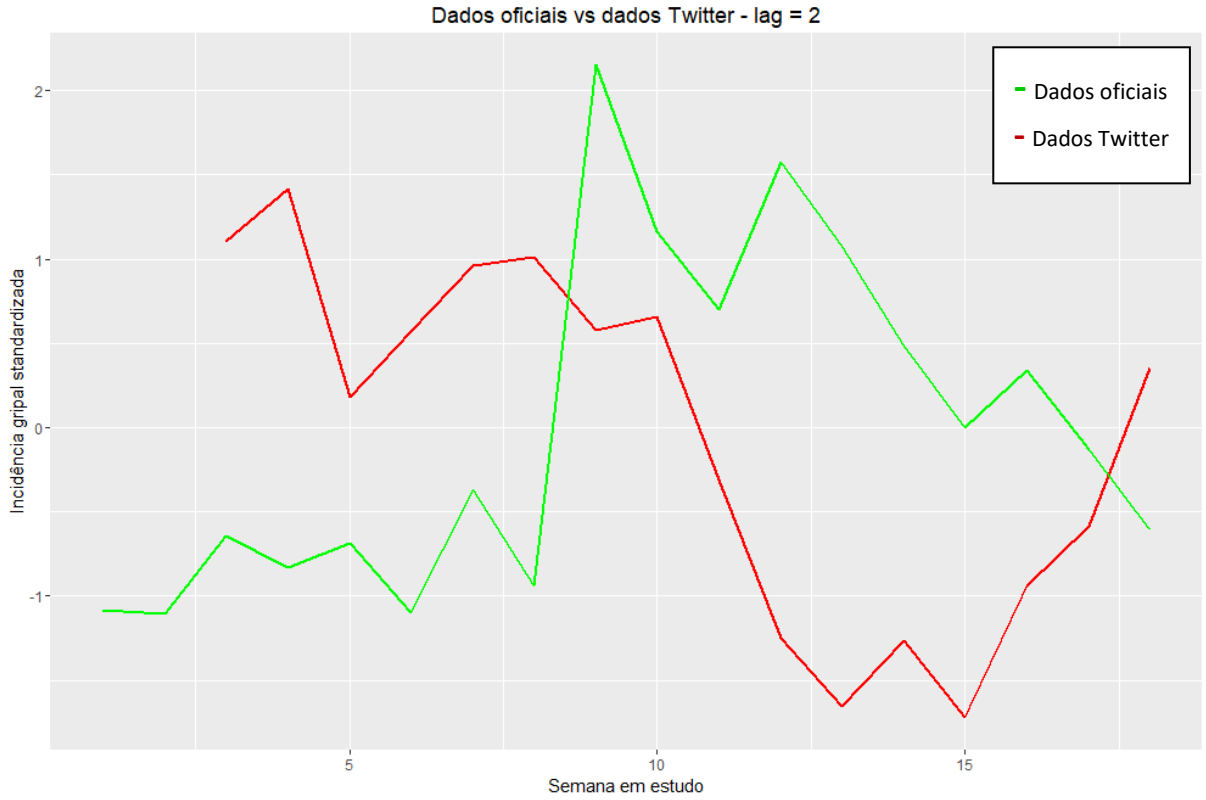

Figura 19 – Dados oficiais vs dados Twitter com *lag* = 2

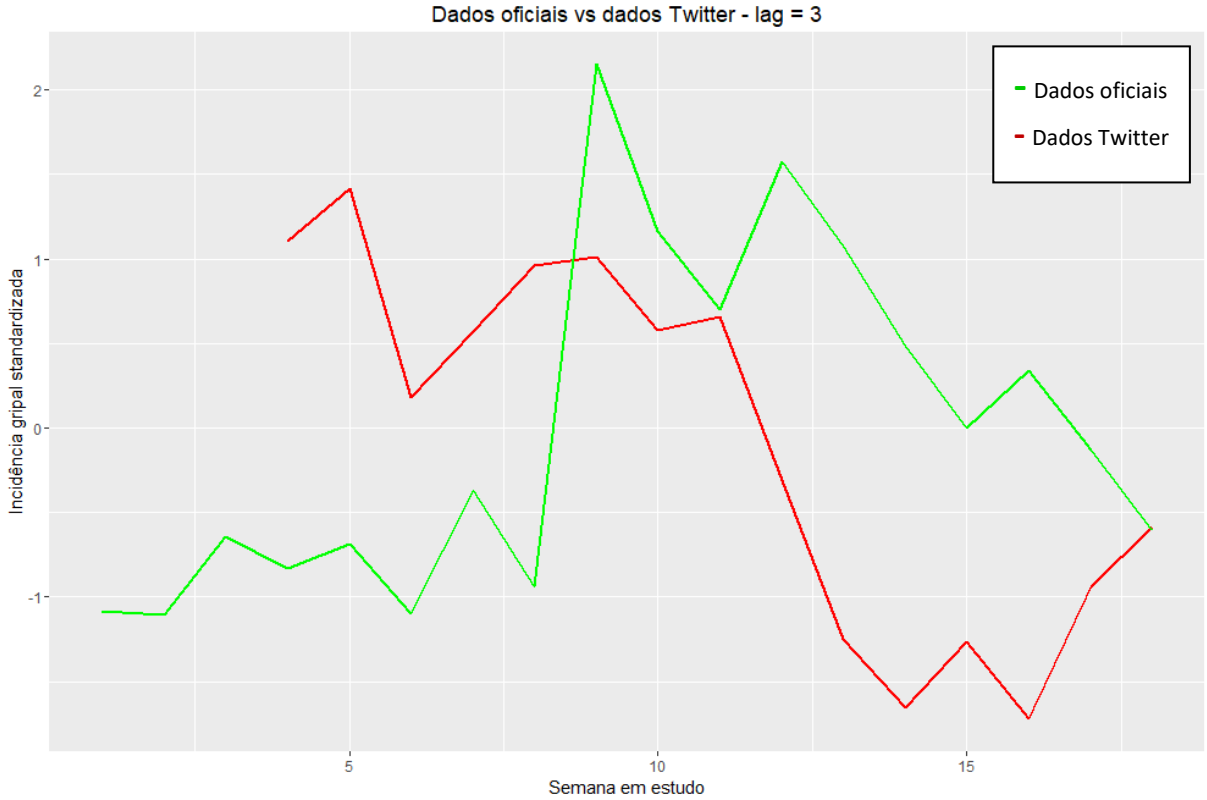

Figura 20 – Dados oficiais vs dados Twitter com *lag* = 3

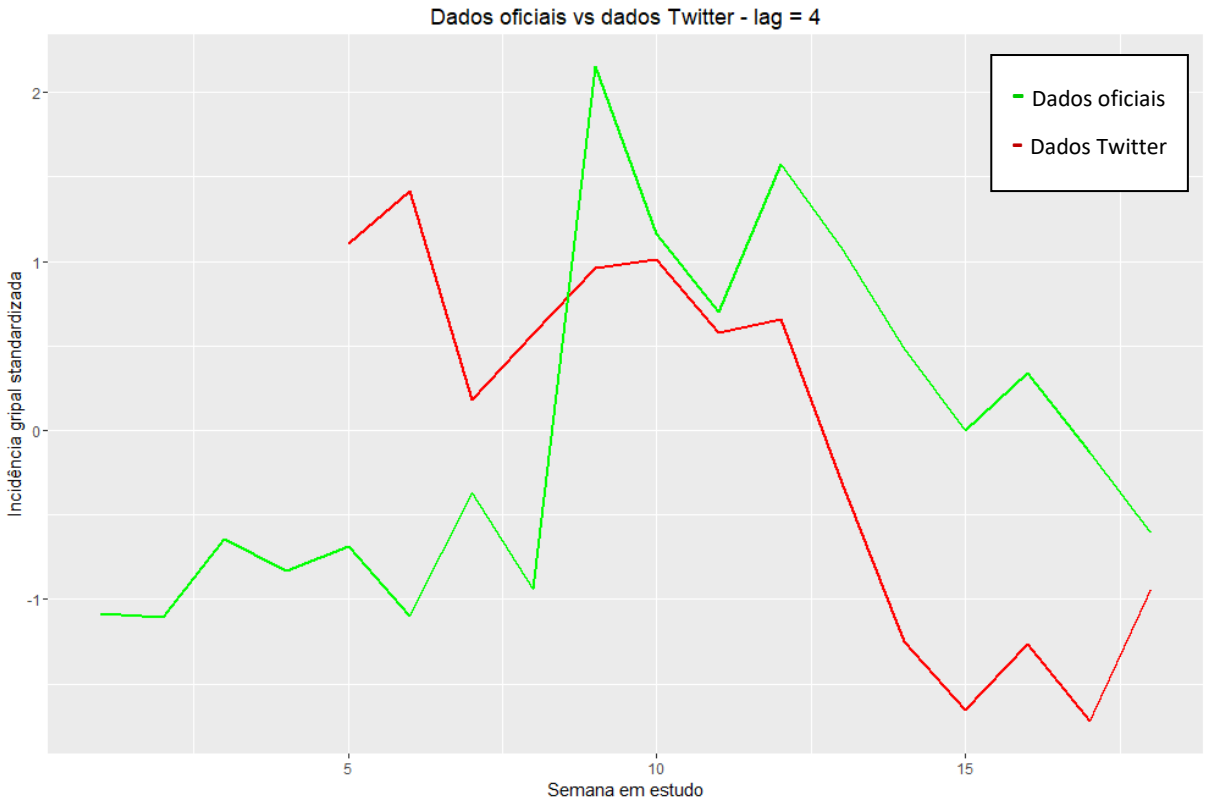

Figura 21 – Dados oficiais vs dados Twitter com *lag* = 4

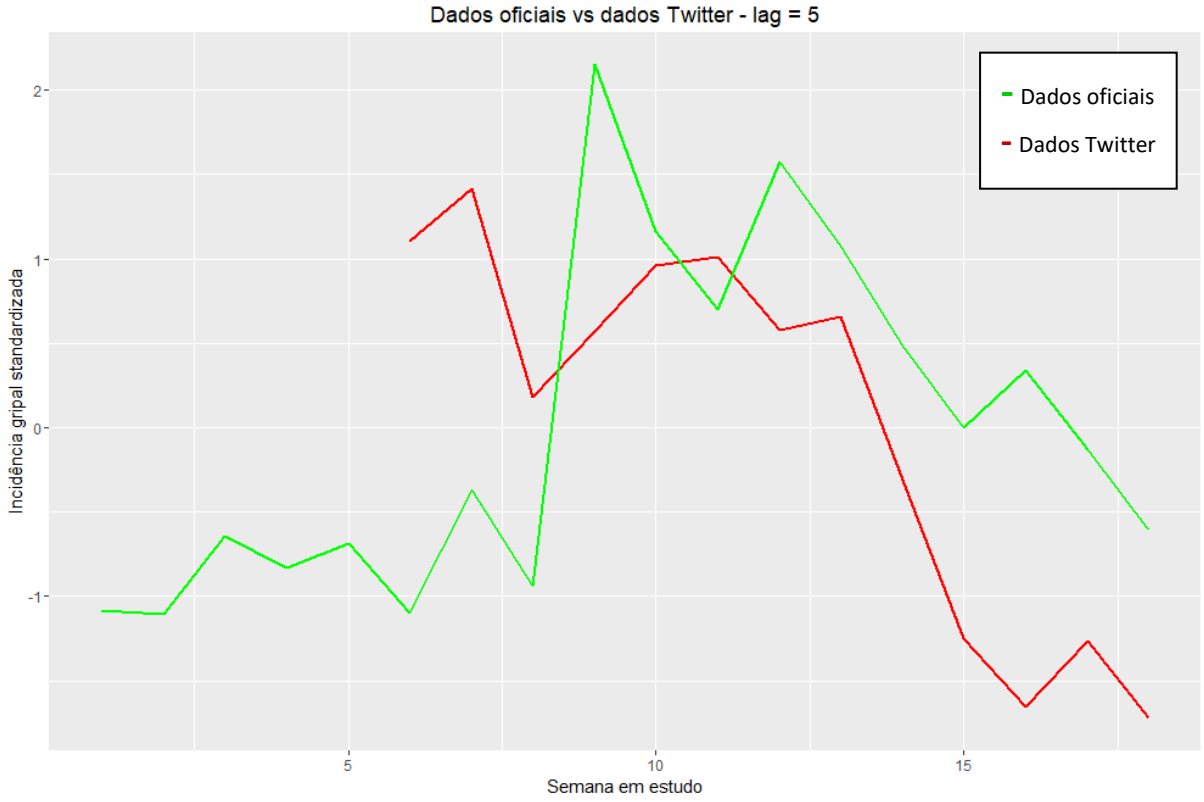

Figura 22 – Dados oficiais vs dados Twitter com *lag* = 5

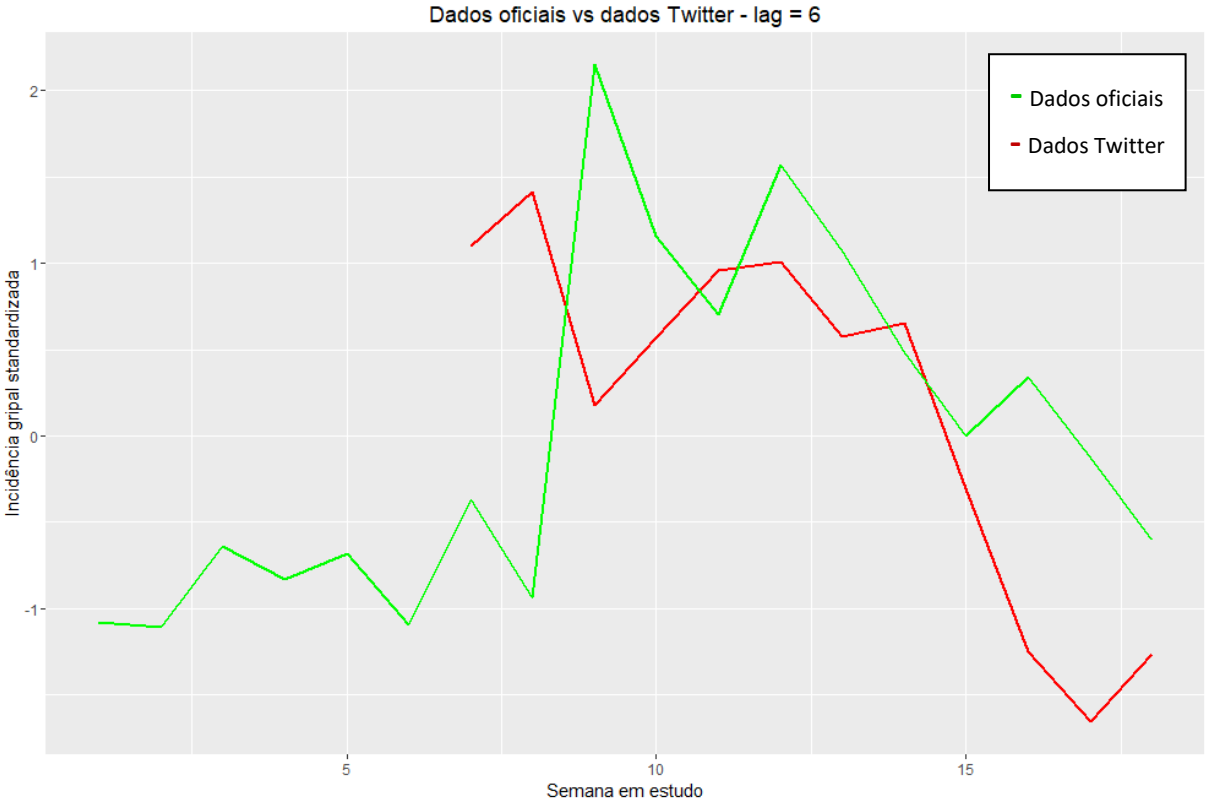

Figura 23 – Dados oficiais vs dados Twitter com *lag* = 6

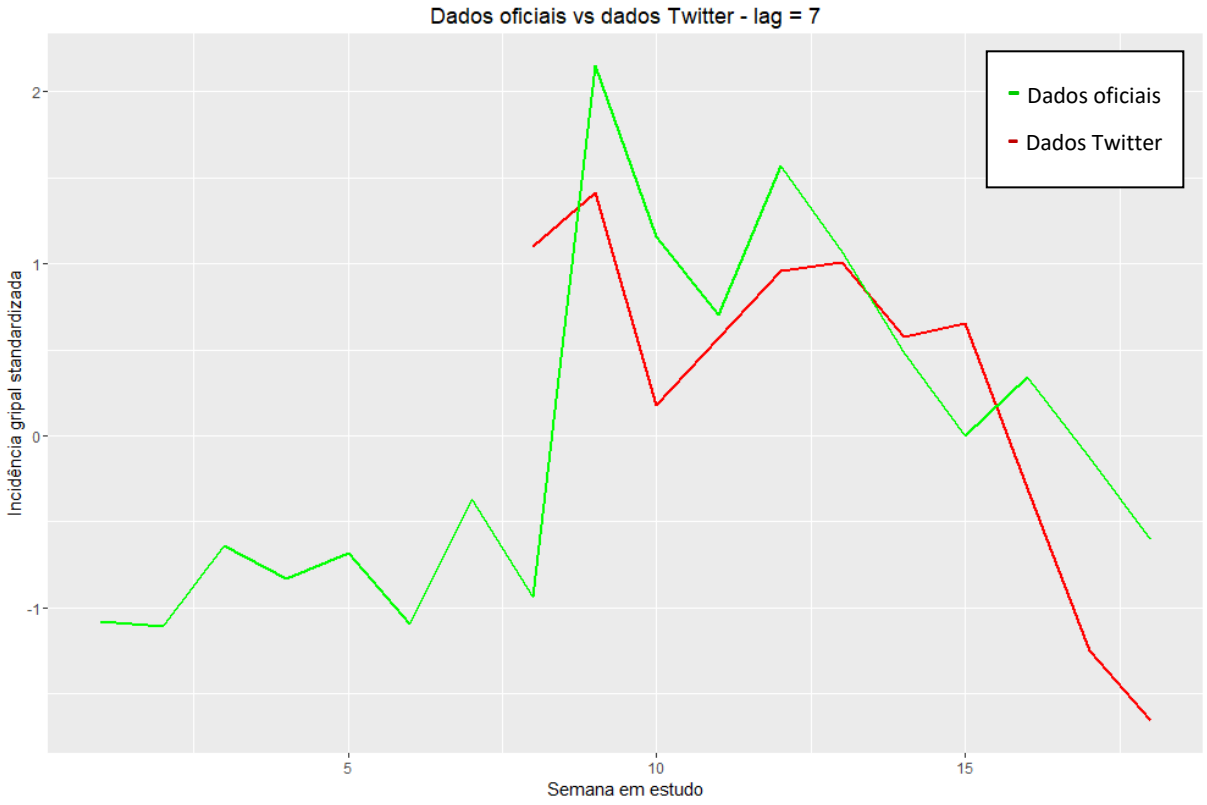

Figura 24 – Dados oficiais vs dados Twitter com *lag* = 7

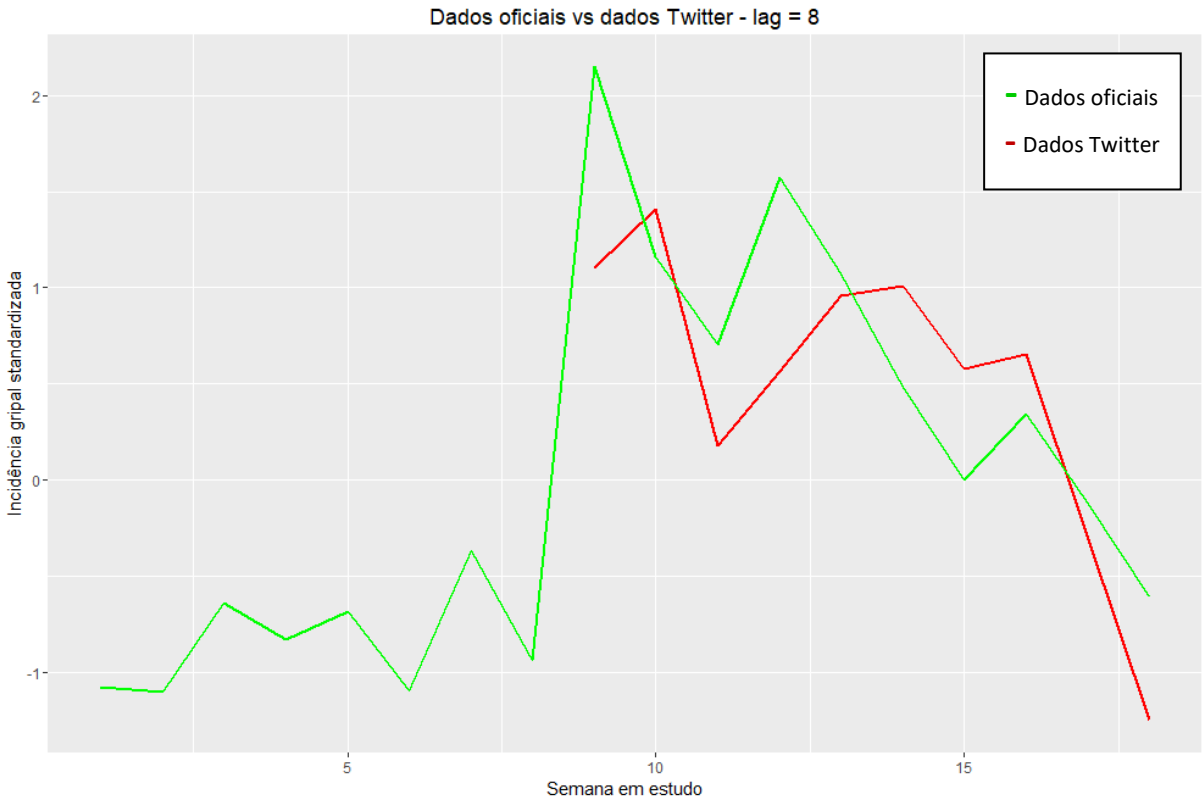

Figura 25 – Dados oficiais vs dados Twitter com *lag* = 8

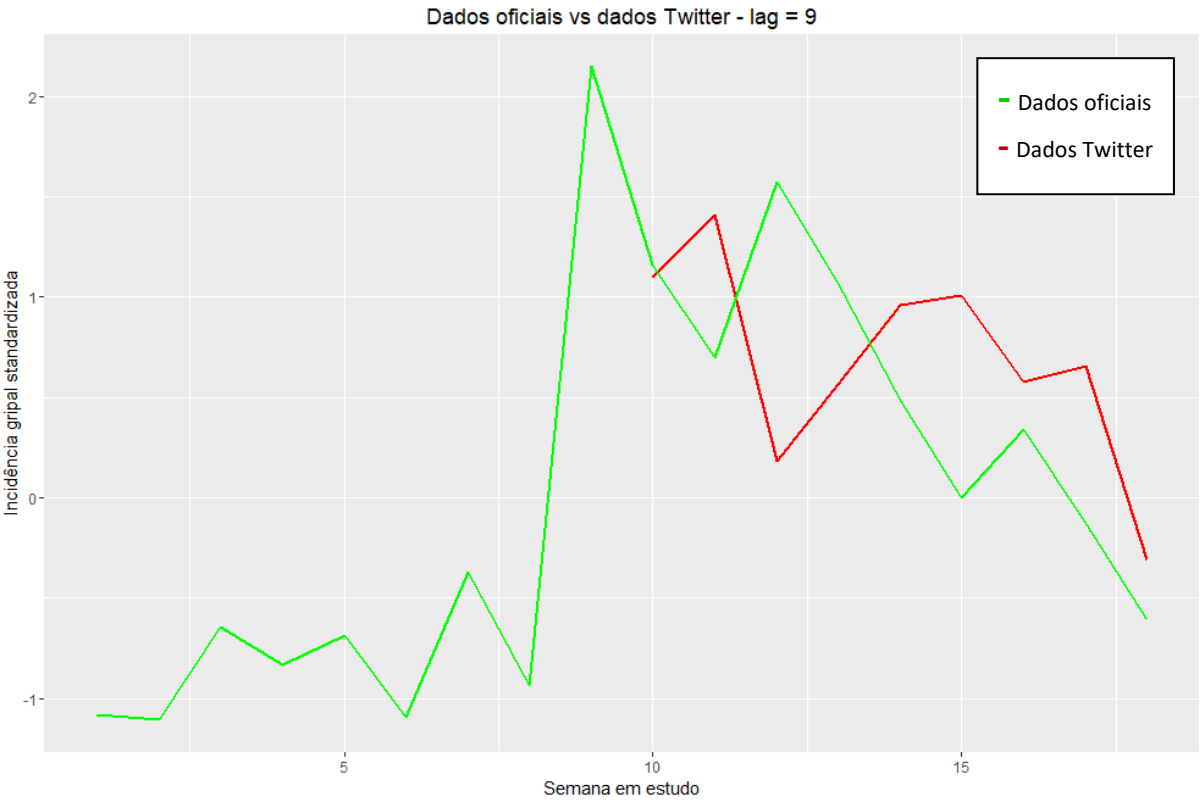

Figura 26 – Dados oficiais vs dados Twitter com *lag* = 9

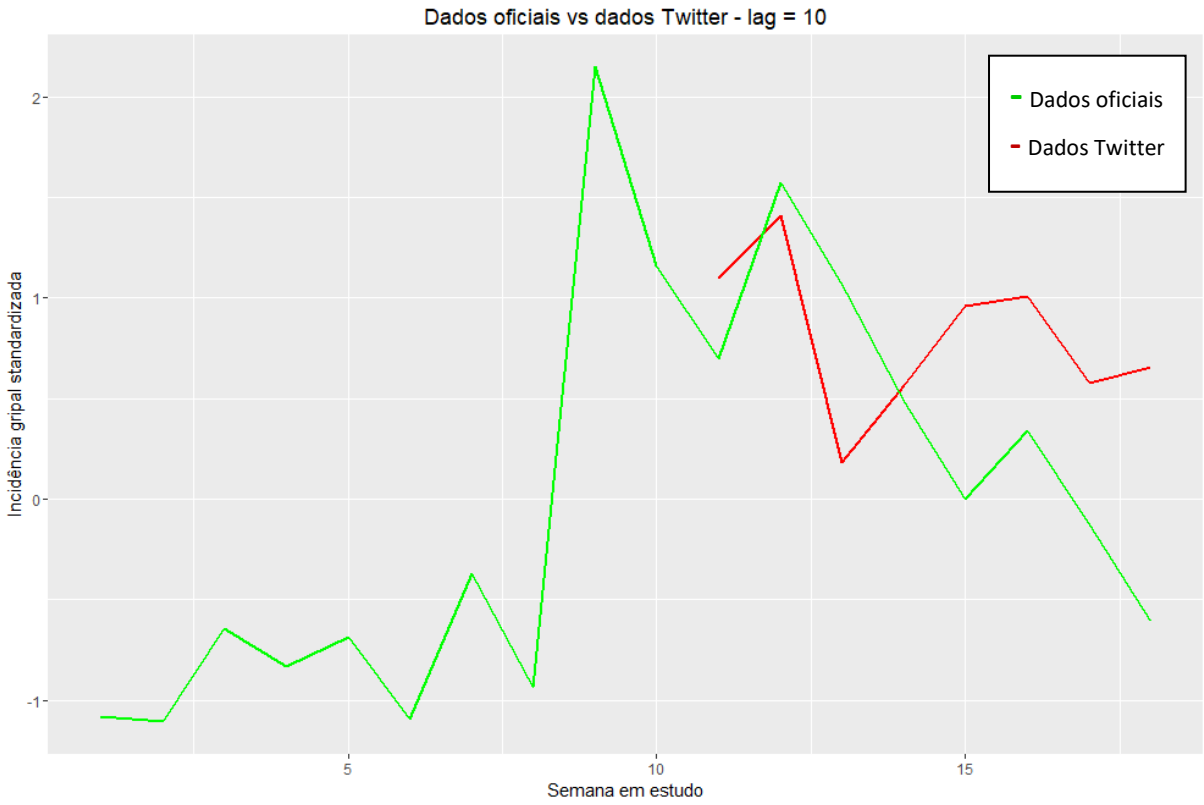

Figura 27 – Dados oficiais vs dados Twitter com *lag* = 10

8.2.2. Cinco classes

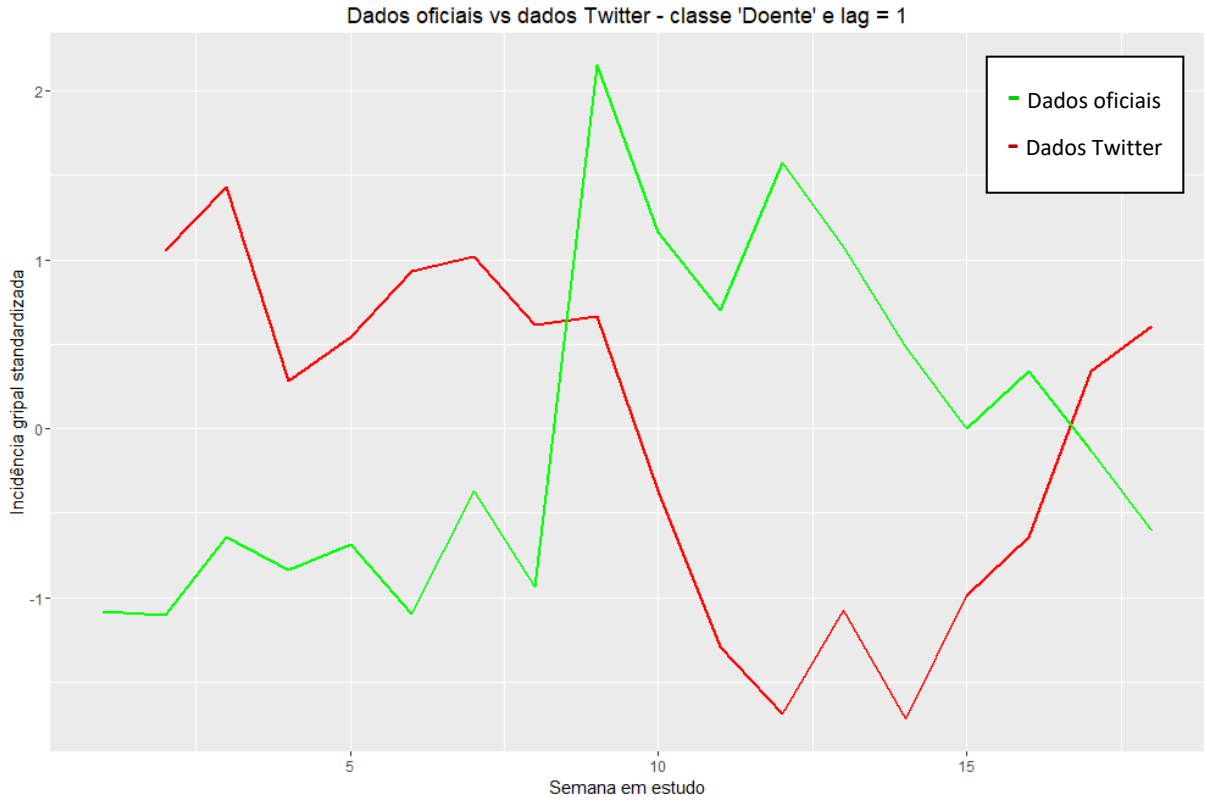

Figura 28 – Dados oficiais vs dados Twitter - classe "Doente" e com *lag* = 1

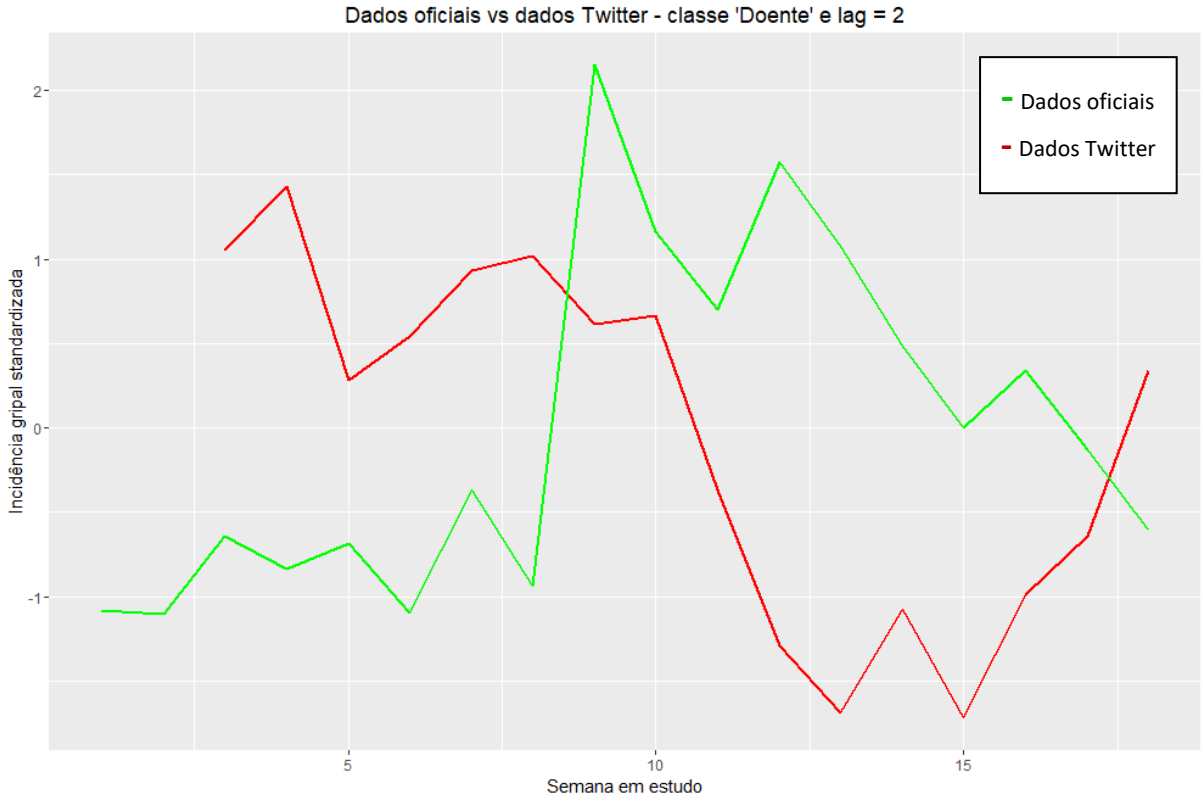

Figura 29 – Dados oficiais vs dados Twitter - classe "Doente" e com *lag* = 2

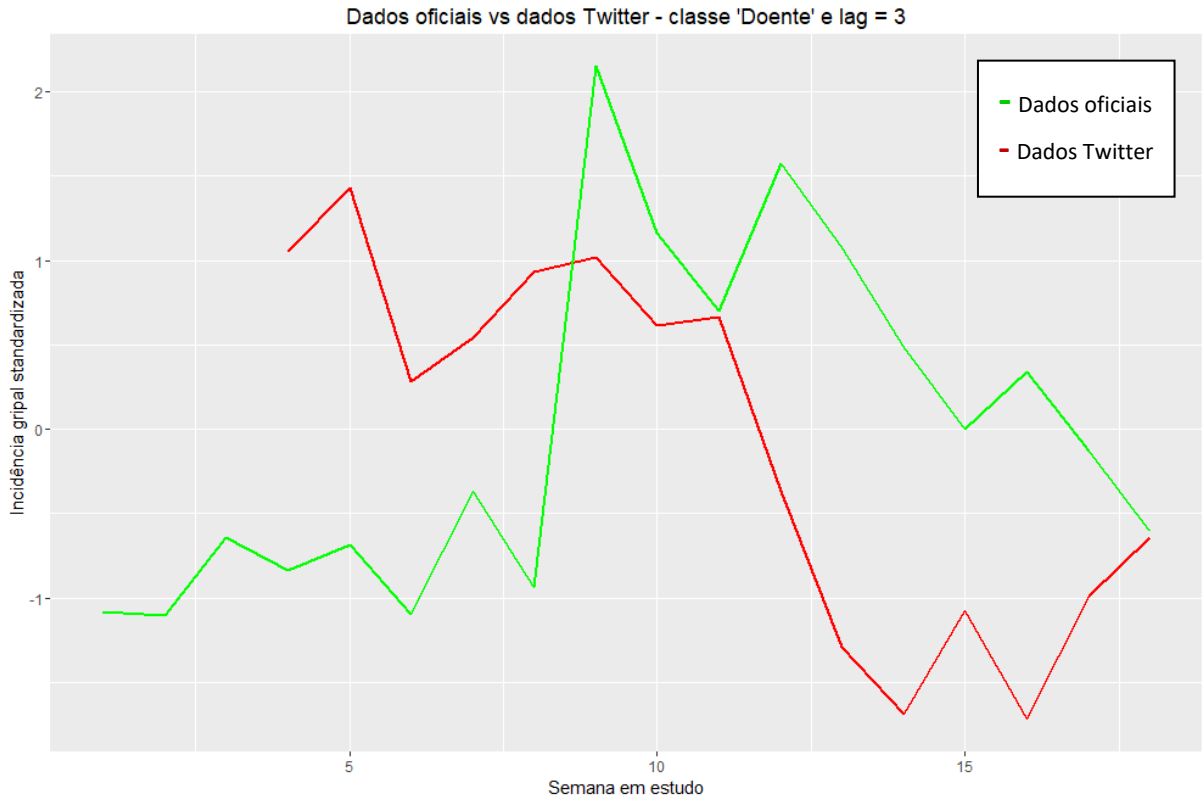

Figura 30 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 3

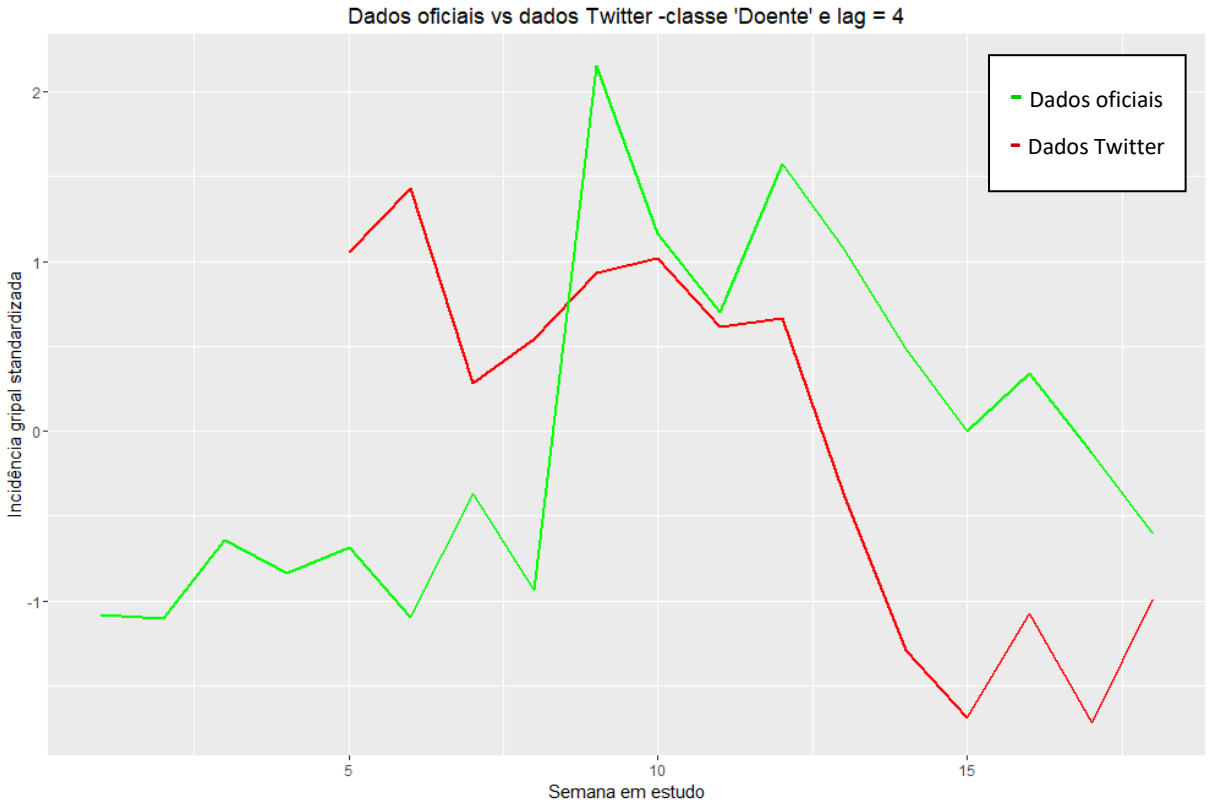

Figura 31 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 4

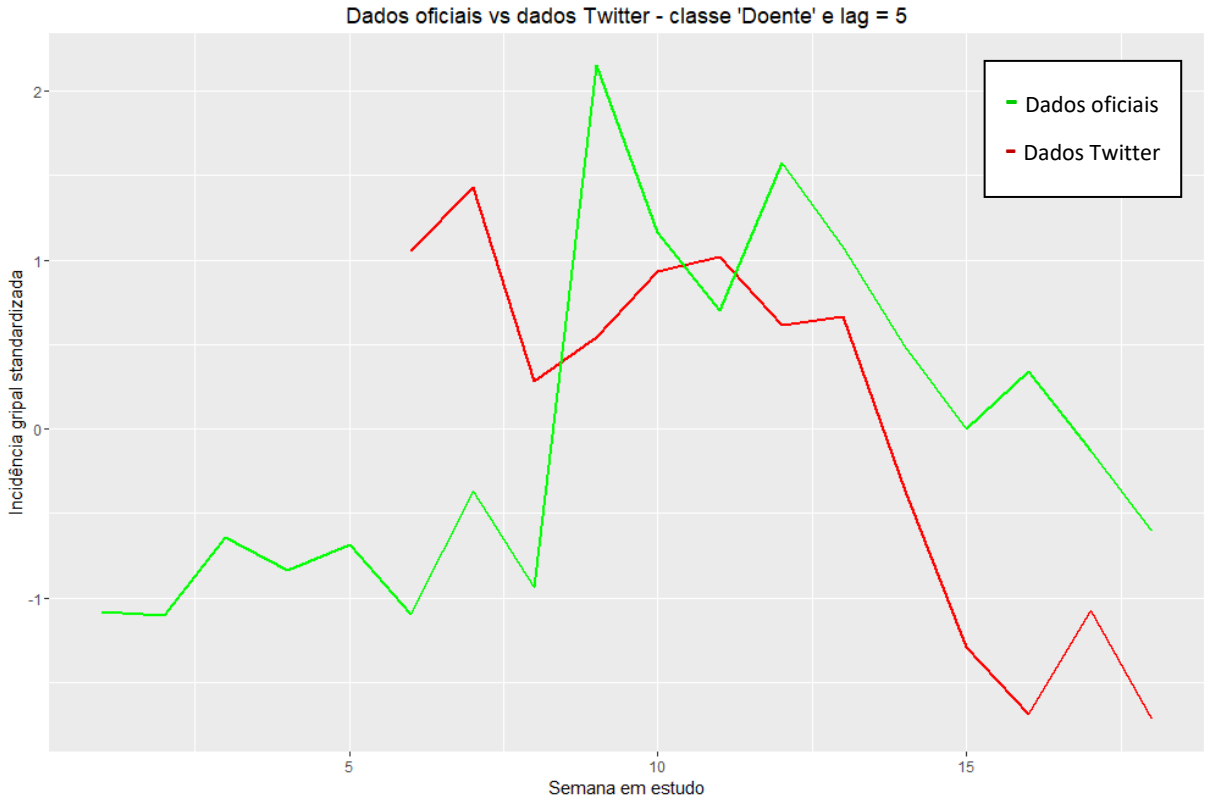

Figura 32 – Dados oficiais vs dados Twitter - classe "Doente" e com *lag* = 5

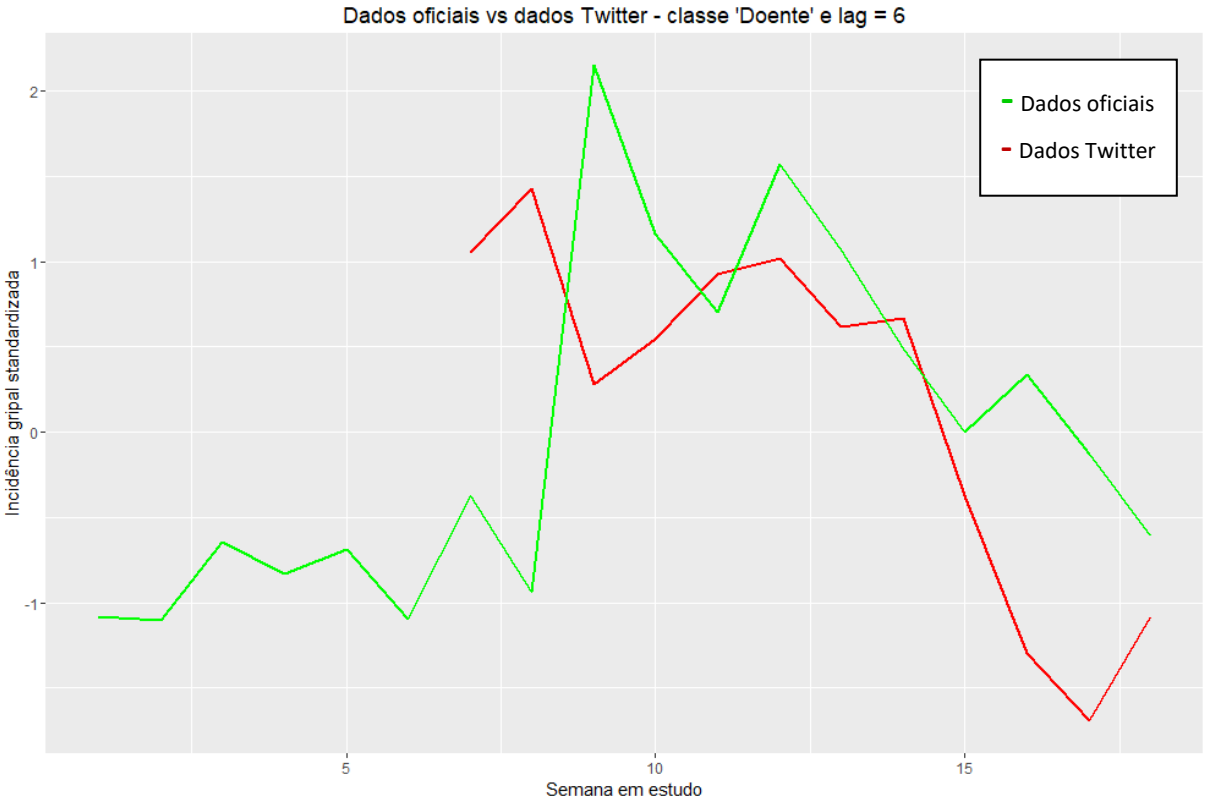

Figura 33 – Dados oficiais vs dados Twitter - classe "Doente" e com *lag* = 6

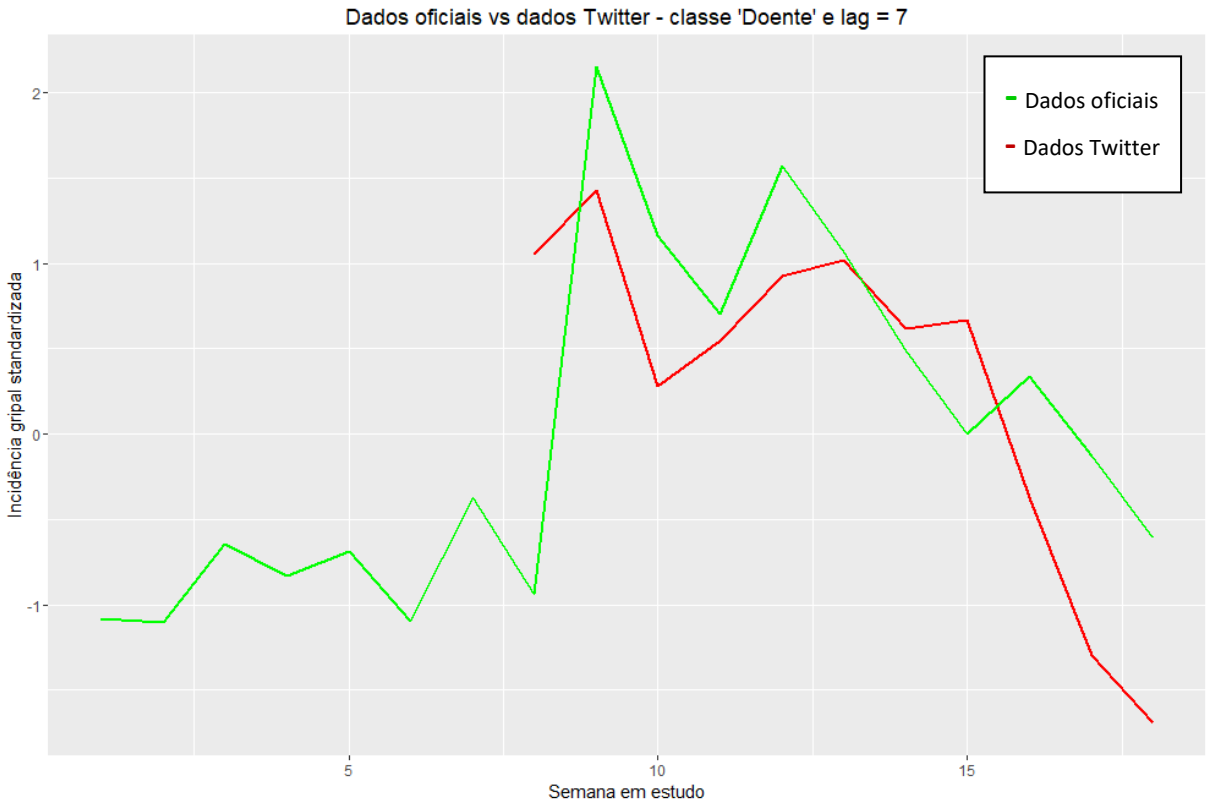

Figura 34 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 7

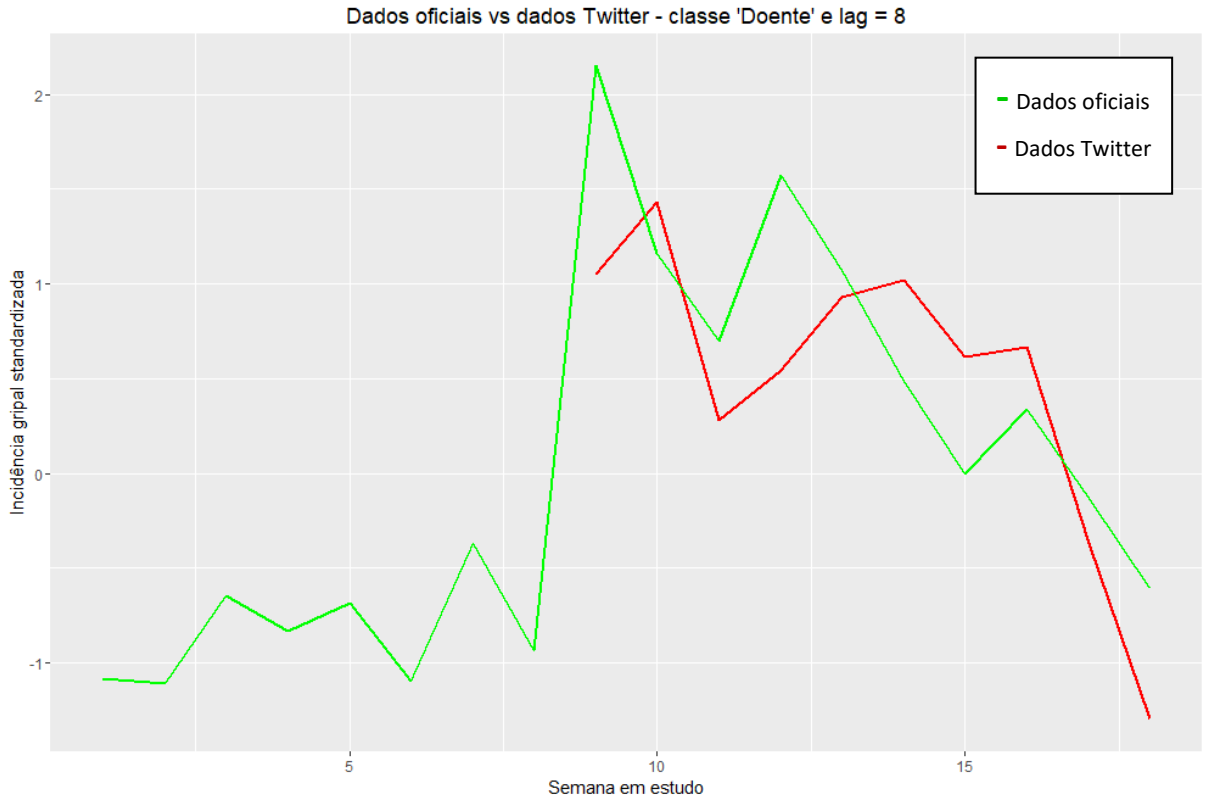

Figura 35 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 8

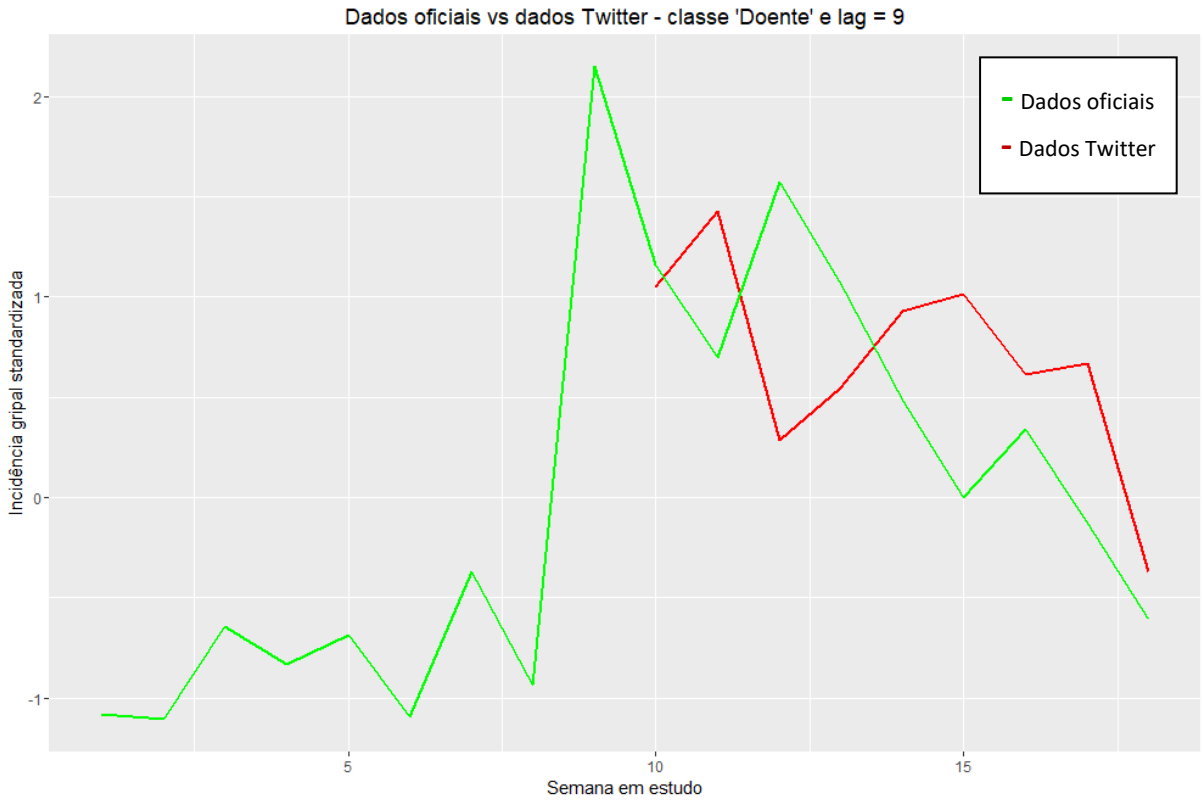

Figura 36 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 9

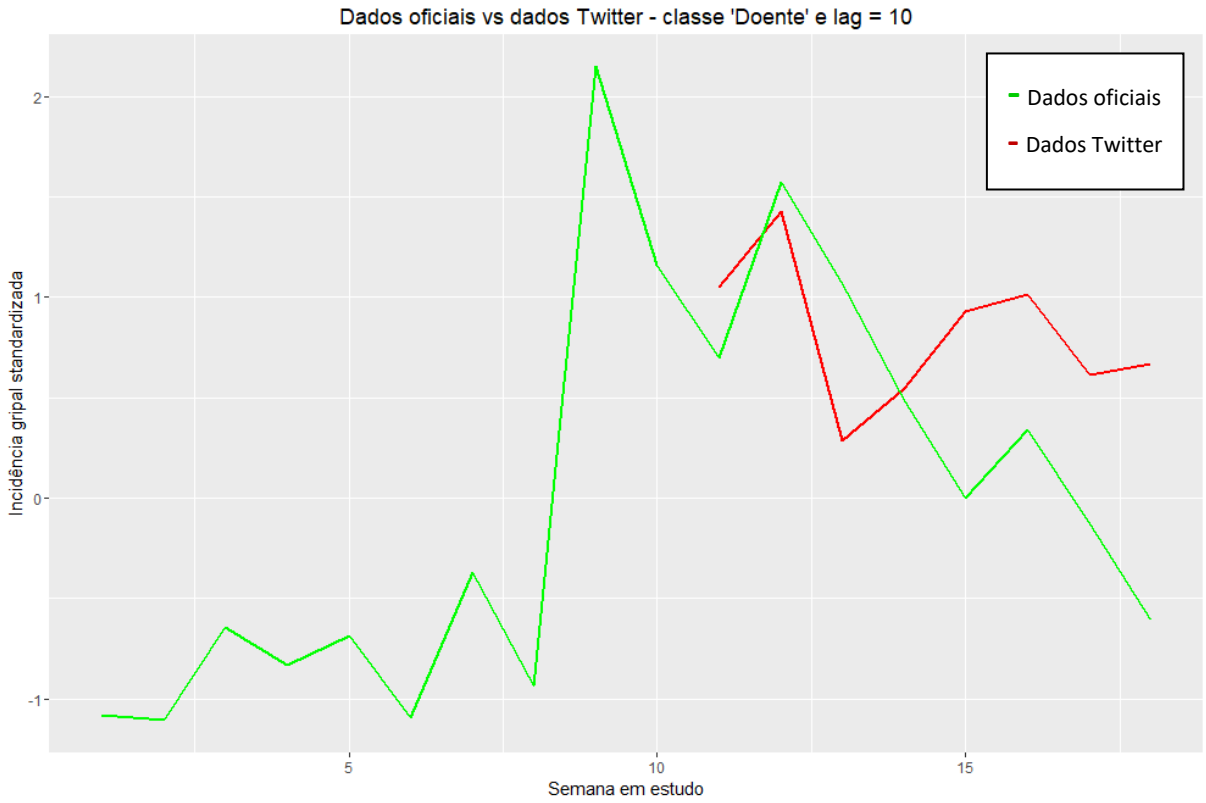

Figura 37 – Dados oficiais vs dados Twitter - classe “Doente” e com *lag* = 10

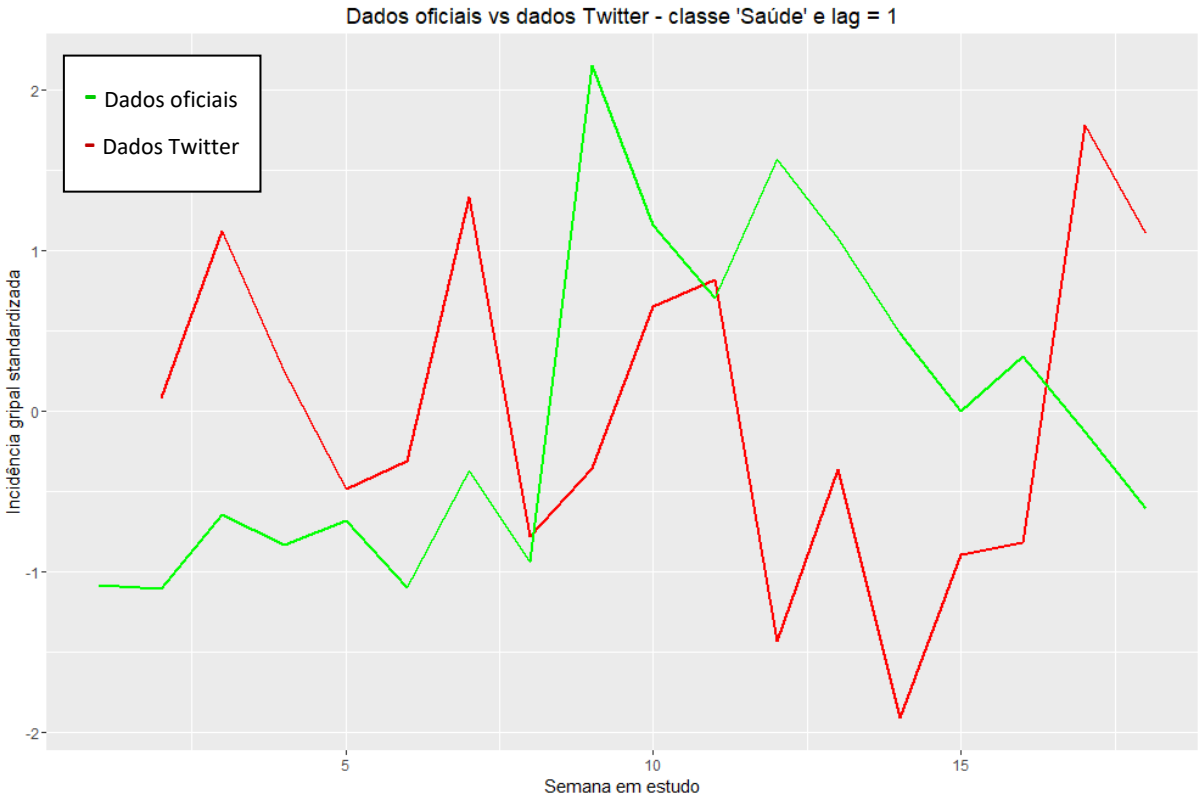

Figura 38 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 1

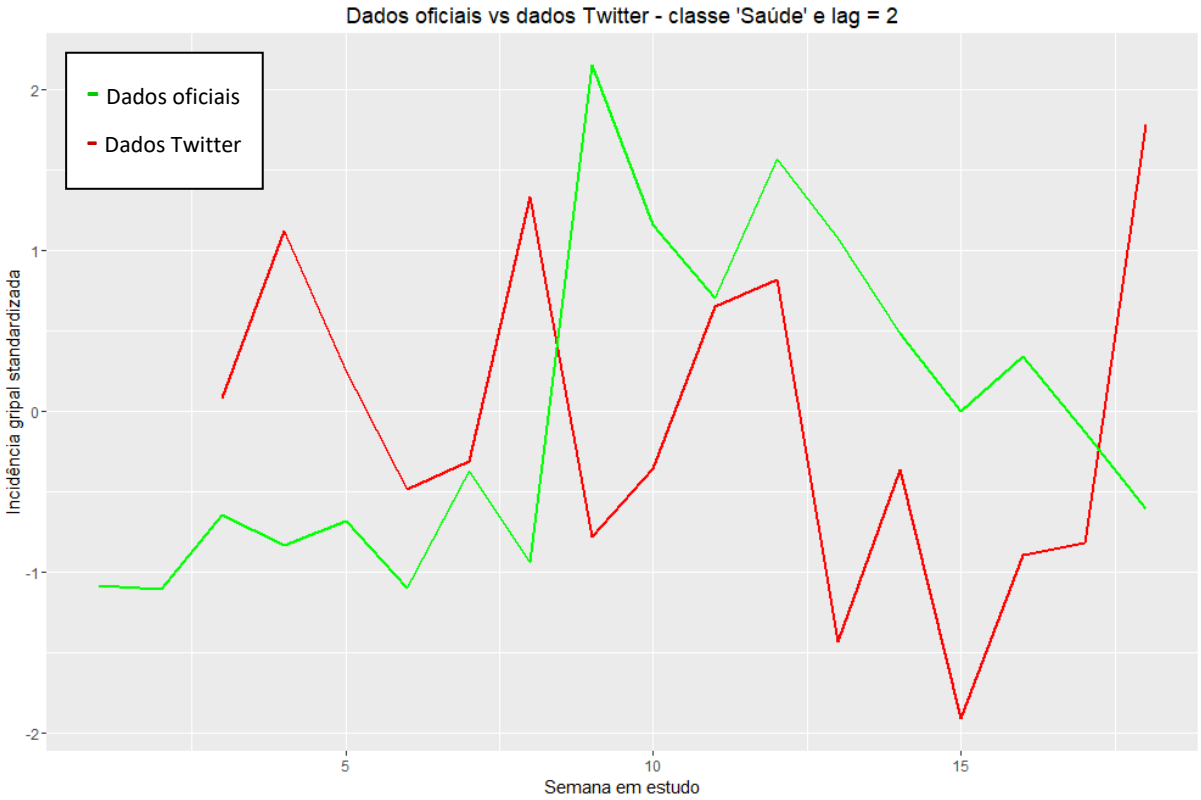

Figura 39 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 2

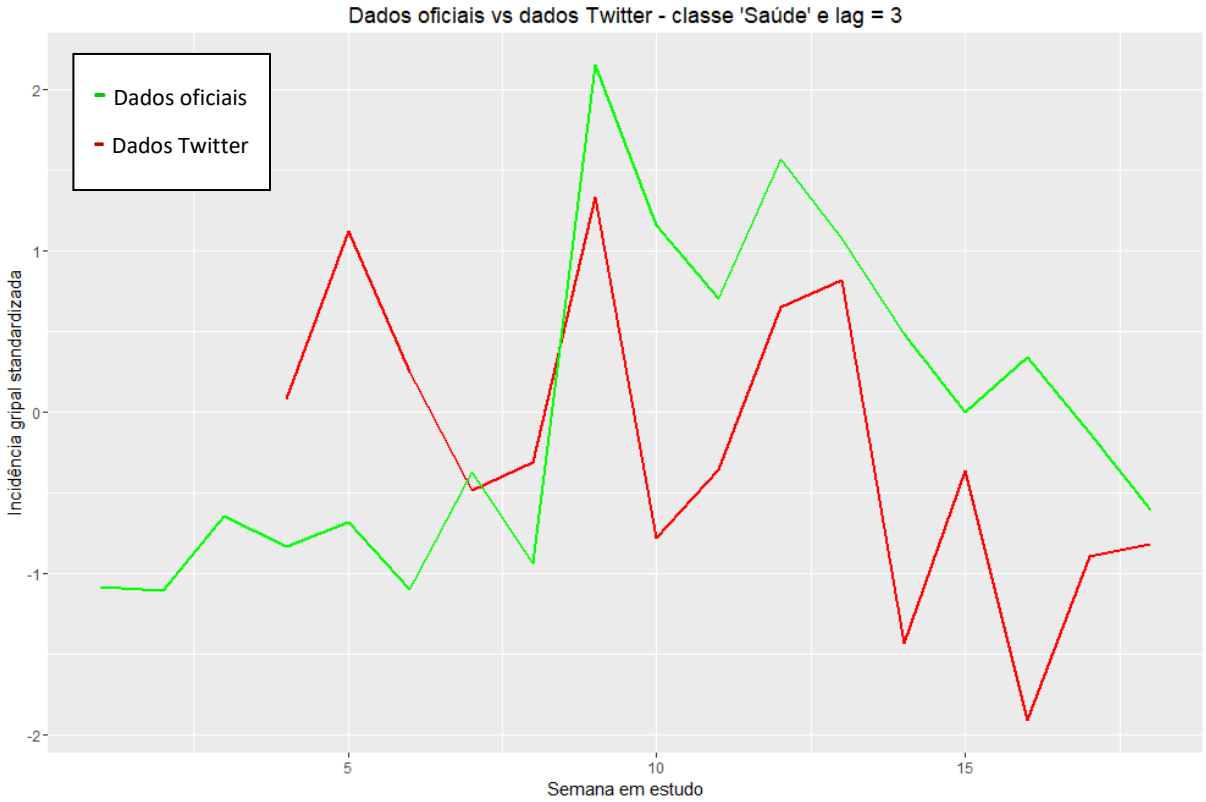

Figura 40 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 3

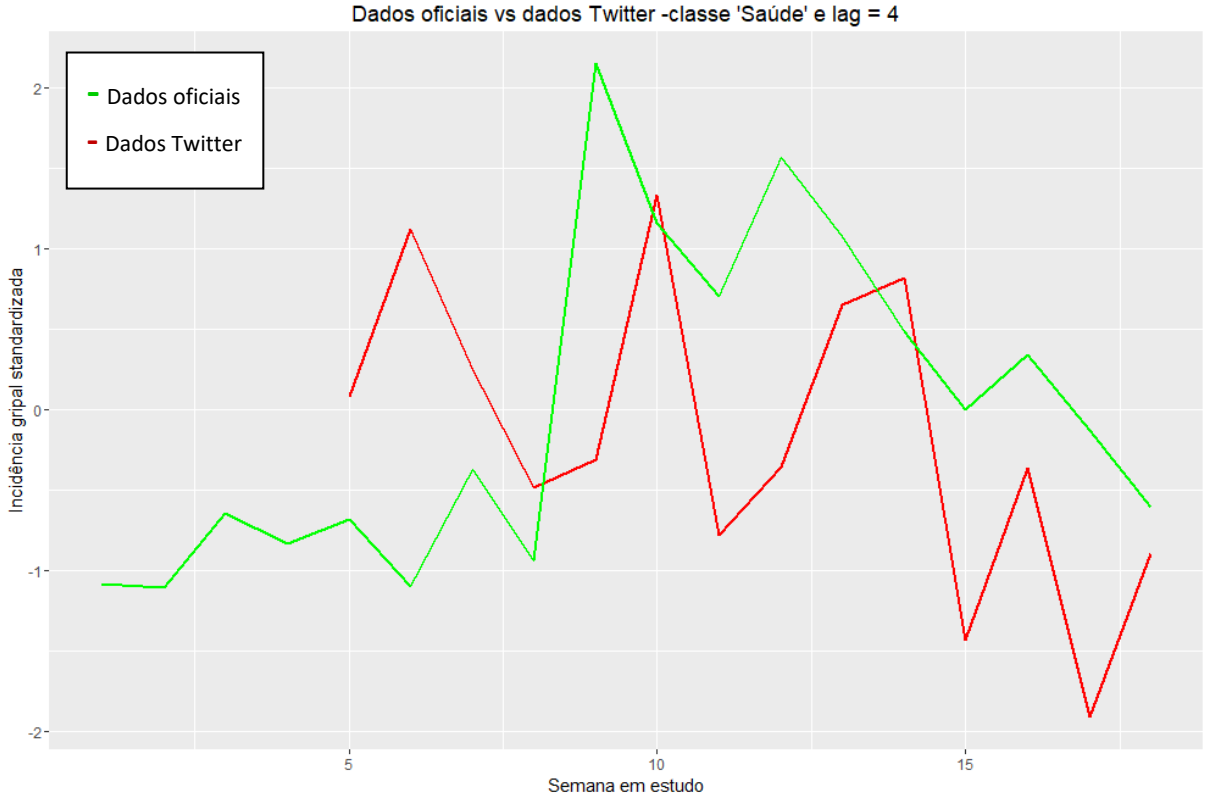

Figura 41 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 4

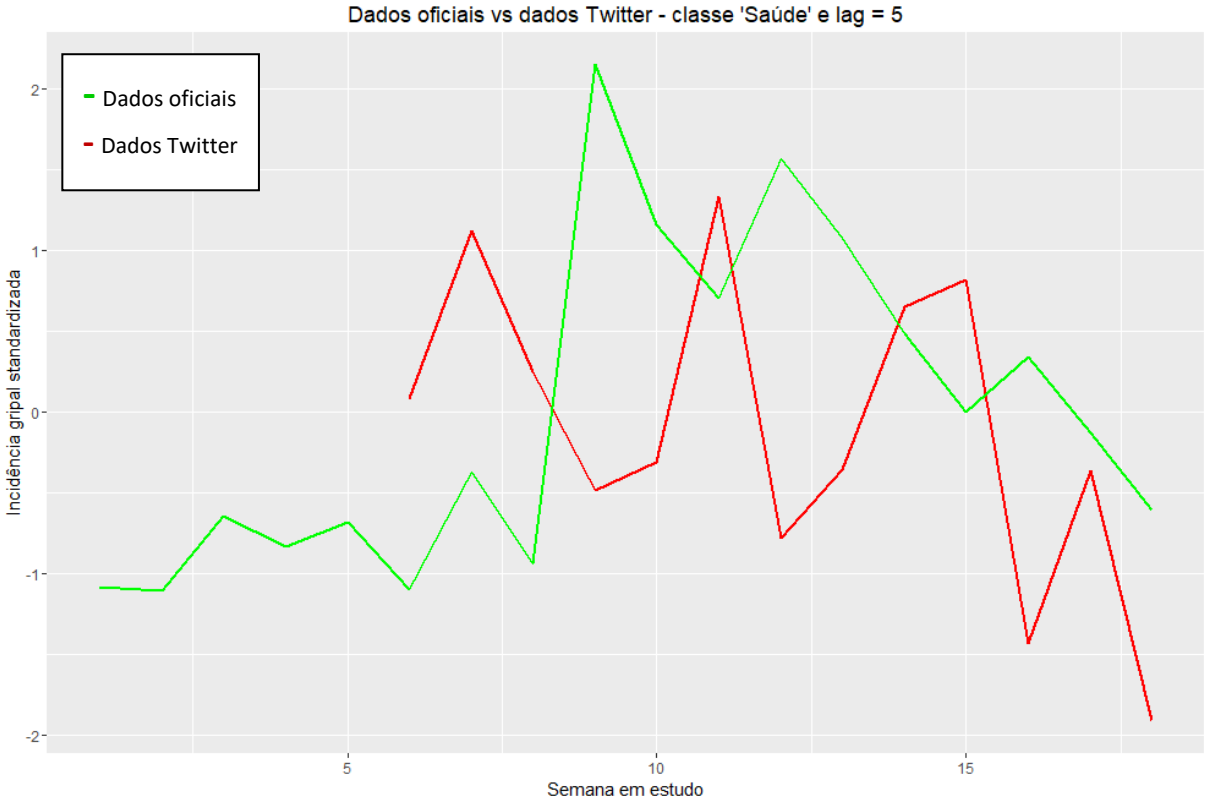

Figura 42 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 5

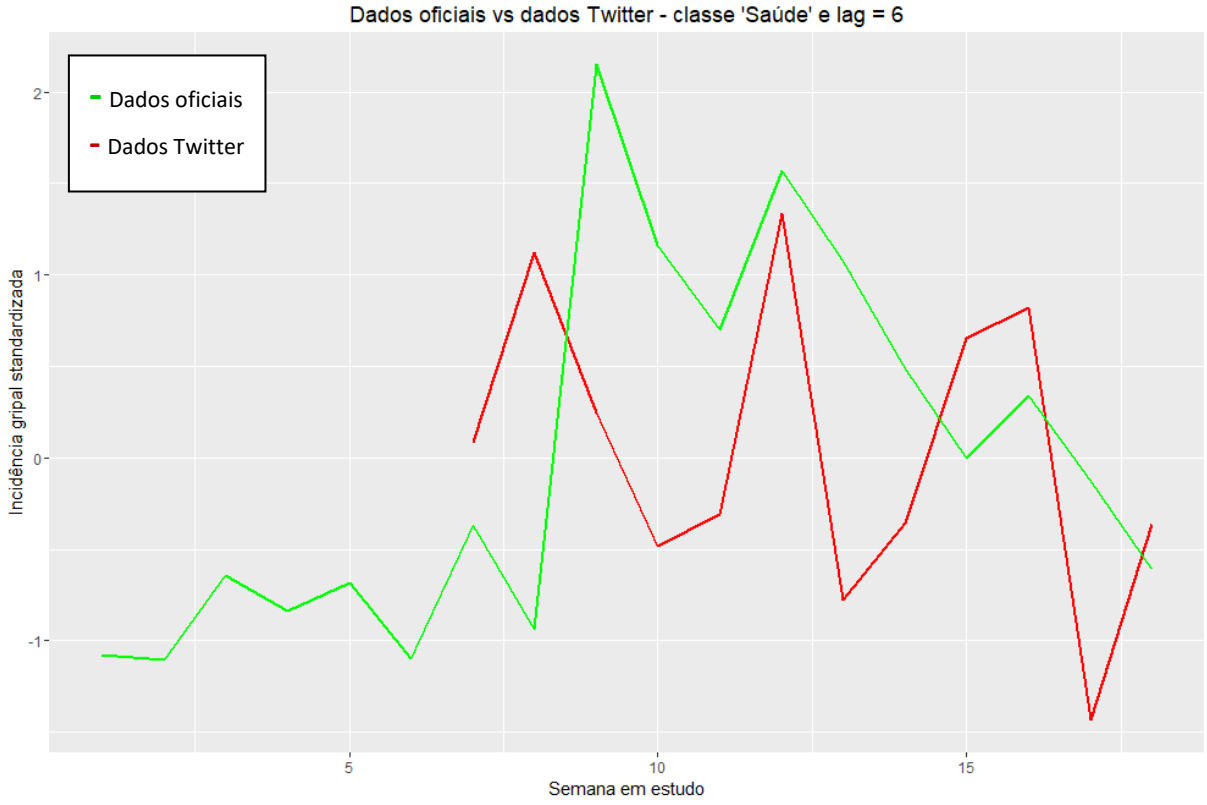

Figura 43 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 6

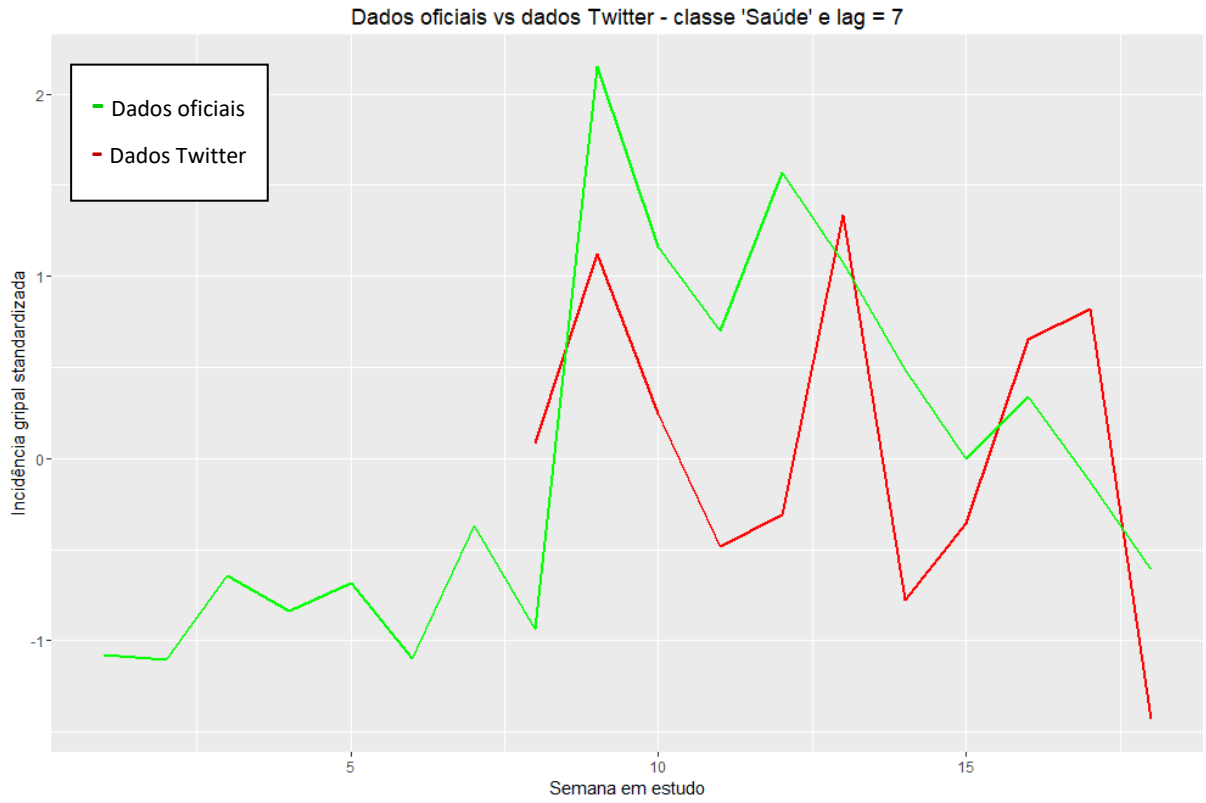

Figura 44 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 7

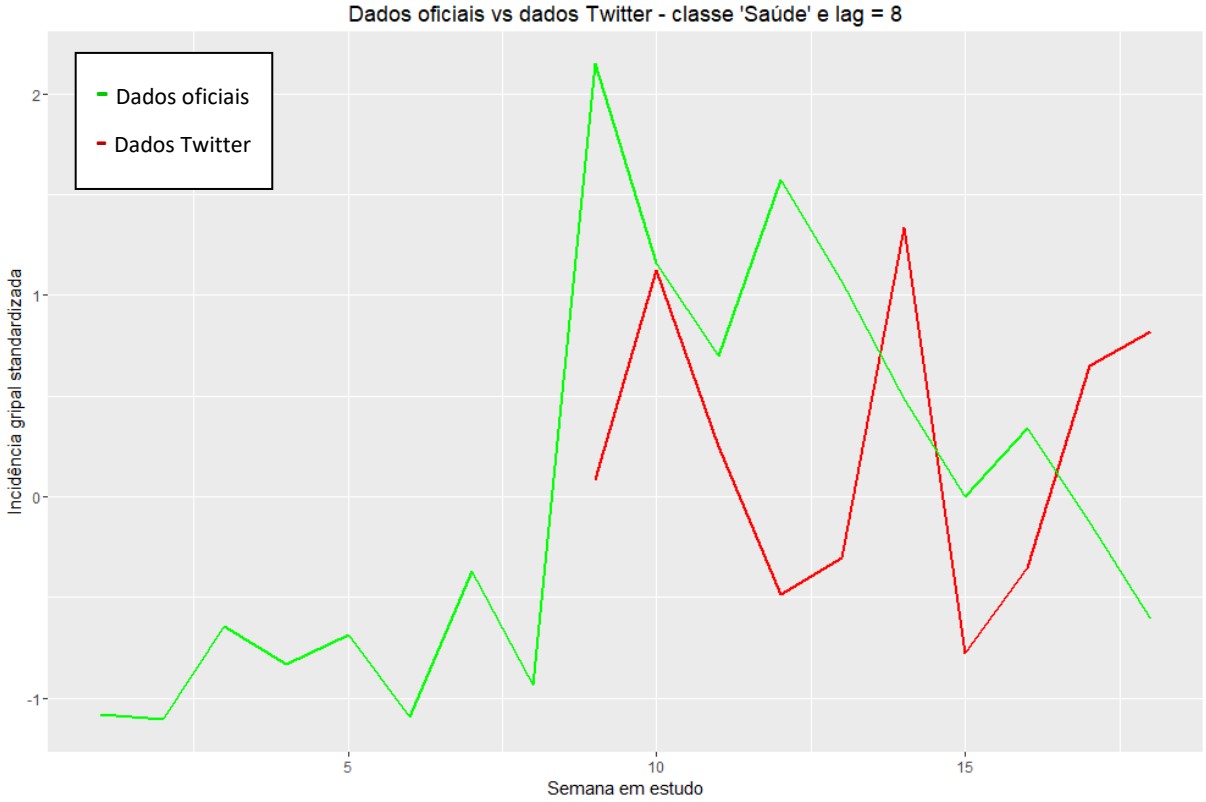

Figura 45 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 8

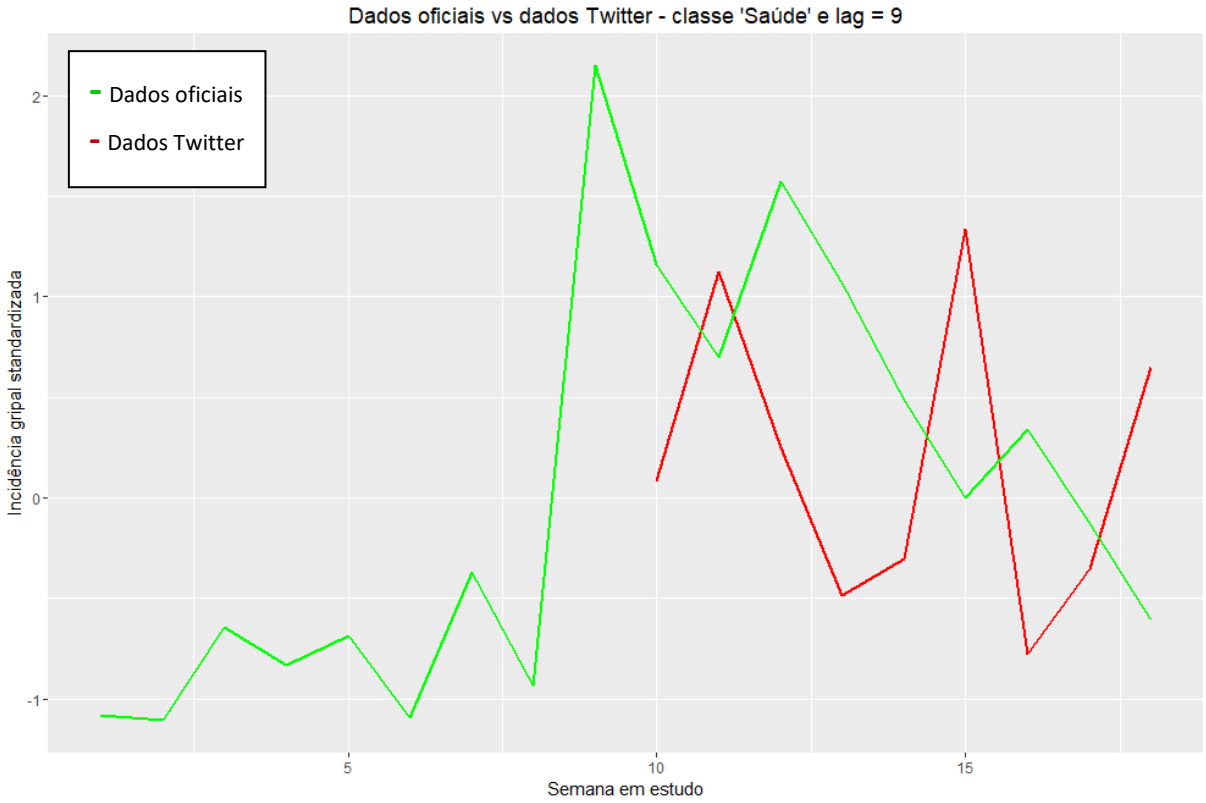

Figura 46 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 9

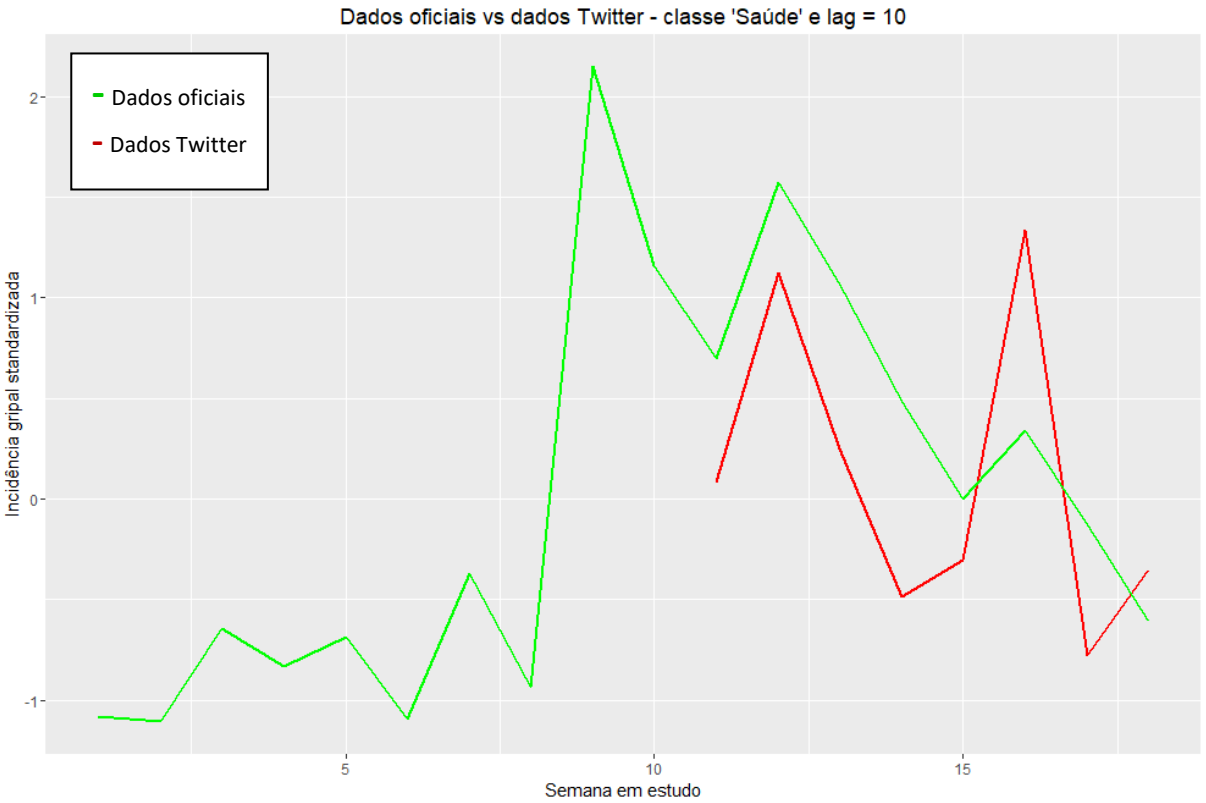

Figura 47 – Dados oficiais vs dados Twitter - classe “Saúde” e com *lag* = 10