Predicting Flu Incidence from Portuguese Tweets

José Carlos Santos and Sérgio Matos

IEETA
University of Aveiro,
Aveiro, Portugal
http://bioinformatics.ua.pt

Abstract. Social media platforms encourage people to share diverse aspects of their daily life. Among these, shared health related information might be used to infer health status and incidence rates for specific conditions or symptoms. In this work, we evaluate the use of Twitter messages and search engine query logs to estimate the incidence rate of influenza like illness in Portugal.
Based on a classified dataset of 2704 tweets from Portugal, we obtained a precision of 0.78 and an F-measure of 0.83 for a Naive Bayes classifier with 650 textual features. We obtained a Pearson's correlation ratio of 0.89 ($p < 0.001$) between health-monitoring data from the Influenzanet project and the prediction by a multiple linear regression model, using as predictors the relative frequencies estimated from the classifier output and from query logs.
Although the Portuguese community in Twitter is small, our results are comparable to previous approaches in other languages, and indicate that this approach could be used in the future to complement other measures of disease incidence rates.

Keywords: flu incidence; disease outbreak monitoring; social media; user-generated content; text classification

1 Introduction

Social media platforms such as Twitter encourage people to share their opinions, thoughts and life aspects [6]. Among these, people often share personal health related information, such as the appearance of flu symptoms or the recovery of those symptoms. Other types of user-generated content, such as Internet searches or comments to news articles, may also contain information related to some of these aspects. Thus, this information could be used to identify flu cases and estimate the influenza rate through time.

Several works regarding the retrieval of health information from social media have already been published. Aramaki et al. [1] applied SVM machine learning techniques to Twitter messages to predict influenza rates in Japan. Lampos and Cristianini [5] and Culotta [2, 3] analysed Twitter messages using regression

models, in the United Kingdom and the United States respectively, obtaining correlation rates of approximately 0.95. Different works also rely on query logs to track influenza activity. Ginsberg et al. [4] presented Google Flu Trends, which uses Google search queries and achieves an average correlation of 0.97 when compared against the ILI percentages provided by the Centers for Disease Control (CDC). The greatest advantage of these methods over traditional ones is instant feedback: while health reports are published in a weekly or monthly basis, Twitter and/or query log analyses can be obtained almost instantly. This characteristic is of extreme importance because early stage detection can reduce the impact of epidemic breakouts [1, 4]. In this work, we describe a method to estimate the occurrence rate of influenza in Portugal based on integrated information from user-generated content, namely tweets and query logs. We show results for two different validations. First, we validate the classification of tweets as 'positive' or 'negative', according to whether the text points to the occurrence of flu. Secondly, we assess how well the results of the proposed method correlate to flu incidence rates predicted by health reports. The article is organized as follows: Section 2 describes the methodology, results are given in Section 3, and a discussion of these results follows in Section 4.

2 Data and Methods

As stated before, the main objective of this work is to create a reliable predictive model to obtain instant feedback regarding the incidence of flu in Portugal. In order to assess the performance of our approach, we compare it to epidemiological results from Influenzanet[1], a health-monitoring project related to flu. Influenzanet data is collected from several participants who sign up to the project and report any influenza symptoms, such as fevers or headaches, on a weekly basis. The gold standard period considered runs from 28 November 2011 to 22 April 2012.

We used training data from nearly 14 million tweets originated in Portugal and covering the period between March 2011 and February 2012, an average of 324 thousand tweets per week. We excluded re-tweets (replies) and tweets including links. Besides Twitter, we also used around 15 million query log entries from the SAPO[2] search platform from December 2011 to May 2012, an average of 780 thousand log entries per week.

In order to predict flu incidence rates from tweets and query logs, these were expressed as weekly relative frequencies, as done in previous works. This value is calculated as the fraction of tweets (searches) considered relevant for flu prediction to the total number of tweets (searches) produced during each week. The data were time aligned with the Influenzanet results.

[1] `http://www.influenzanet.eu`
[2] `http://www.sapo.pt`

2.1 Regular Expressions

We start by applying regular expressions in order to capture tweets and queries that contain influenza related words. For the queries, we used a simple regular expression that matches "*gripe*" (influenza) word derivations: "*(en)?grip[a-z]+*". A set of 1547 searches was identified, an average of 47 searches per week.

For filtering the Twitter data we used a more complex expression, since tweets may contain a more descriptive account of someone's health status. The regular expression was built according to common insights about how people describe flu and flu-like symptoms, and can be divided into three groups, as described in Table 1: "*gripe*" (influenza) word derivations, "*constipação*" (cold) word derivations and flu related symptoms, such as body/throat pains, headache and fever. Using this regular expression, a set of 3183 tweets was identified, an average of 67 tweets per week.

Table 1. Regular expressions to filter influenza related tweets.

Theme	RegEx
Flu	*(en)?grip[a-z]+*
Cold	*constip[a-z]+*
Flu Symptoms	*(febre .* grau(s)?) | (grau(s)? .* febre) | (**bodypains** .* febre) | (febre .* **bodypains**)*

bodypains: *do(r(es)?\|i-me)\s*(no\|na\|de\|o\|a)?\s*(corpo\|cabeca\|garganta)*

2.2 Classification Methods

Using filtering based on regular expressions as described above is not sufficient, as many tweets that contain words related to flu do not imply that the person writing the text has the flu. Tweets like "*Hoping the flu doesn't strike me again this winter*" contain the keyword flu but do not tell us that this person has the flu. To solve this problem, we applied machine-learning techniques to classify each tweet as "positive" or "negative" according to its content.

User Annotations In order to create the predictive models, we asked a group of 37 independent annotators to manually classify a set of tweets. During the annotation task, each annotator was repeatedly assigned a random tweet, with the following restrictions: each tweet had to be labelled by up to three annotators, and each annotator could not label the same tweet more than once. Annotators were instructed to consider a tweet as positive if it revealed that the person who wrote it was with the flu, was having flu symptoms or was recently ill with the flu. A third category was also used, to indicate tweets referring to "cold". We explore the inclusion of these tweets as positive or negative in the results section. To reduce incorrect answers, the annotators could also label the tweet

as unknown. Tweets with inconsistent or insufficient labelling information did not receive a final label and were not included in the dataset.

We gathered a total of 7132 annotations, resulting in 2704 labelled tweets of which 949 were positive for flu. The three top annotators contributed to the labelling of 56% of the dataset, and the top ten annotators contributed to 90% of the final labels. To validate the data obtained from the annotators, the majority voted labels of 500 random tweets were verified by one of the authors, leading to an annotator accuracy of 95.2%.

Feature Extraction and Selection Tweets were represented by a bag-of-words (BOW) model. The Natural Language Processing Toolkit [8] (NLTK) was used to tokenize the text, remove portuguese stopwords and stem all remaining words in each tweet. Char bigrams for each word were also generated, making up a total of 5106 features.

We applied feature selection techniques for defining the best set of features to use. For this, each feature was compared to the true class label to obtain the mutual information (MI) value. The higher a feature's MI score, the more it is related to the true class label, meaning that the feature contains discriminative information to decide if that tweet should be classified as positive or negative. We selected the optimal number of features empirically, by selecting features with MI value above different threshold values.

Machine Learning Methods Several machine learning techniques (SVM, Naïve Bayes, Random Forest, Decision Tree, Nearest Neighbour) were tested in order to evaluate which would produce better results. We used the SVM-light [7] implementation of SVMs. The remaining classifiers were trained using the Scikit-learn toolkit [9].

2.3 Linear Regression Models

We used linear regression models to estimate the flu incidence rate, using the Influenzanet data to train and validate the regression. We trained both single and multiple linear regressions, combining the predicted values obtained from the different classifiers, query logs and regular expressions. As the input to the regressions, we used the weekly relative frequencies obtained after applying the regular expressions to the web queries and to tweets, and after classifying the tweets with the various classifiers tested. Also, instead of using the number of positive predictions from the classifiers to calculate the weekly relative frequencies, we summed over the classification probabilities of the positive predicted documents in each week, similarly to what was proposed by Culotta [2, 3]:

$$f_i = \frac{\sum_{d_j \epsilon D_i} p(y_j = 1|x_j)}{|D_i|}, \quad p(y_j = 1|x_j) > t \tag{1}$$

where f_i is the predicted incidence in week i, D_i is the set of documents for week i, $p(y_j = 1|x_j)$ is the probability for classifying document x_j as positive, $|D_i|$ is the number of documents in week i, and t is the classification threshold.

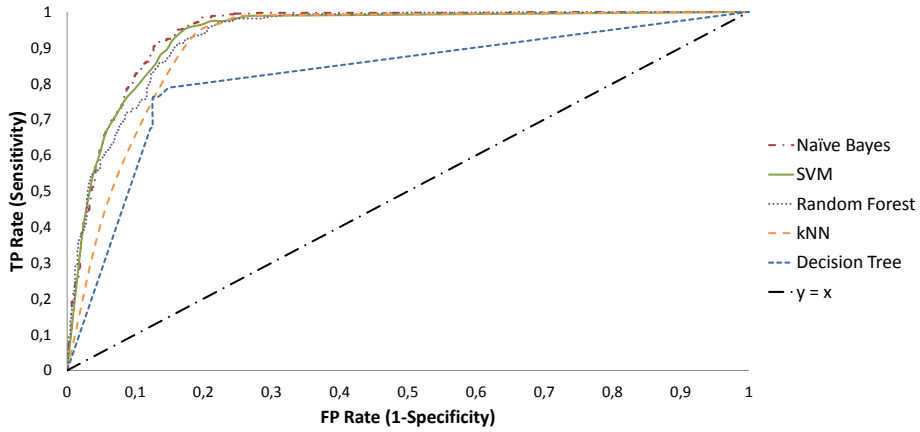

Fig. 1. Receiver operating characteristic (ROC) analysis for each classifier, after feature selection.

3 Results

3.1 Binary Classification of Twitter Messages

The performance of the different classifiers was compared through 5×2-fold cross validation using the entire dataset of 2704 tweets, covering the period from May 2011 to February 2012. Using the full set of features, the best results were obtained with the SVM classifier, with an F-measure of 0.75. After feature selection, the best overall results were obtained for a set of 650 features, achieving an F-measure of 0.83 with both SVM and Naïve Bayes classifiers (Table 2).

Table 2. Classifier results for the reduced set of features after feature selection by mutual information (650 features)

Classifier	F-Measure	Precision	Recall	AUC
Naïve Bayes	0.83	0.78	0.90	0.941
SVM	0.83	0.78	0.88	0.939
Random Forest	0.81	0.76	0.86	0.934
kNN	0.80	0.72	0.92	0.896
Decision Tree	0.75	0.75	0.75	0.780

Applying a simple linear regression between the predictions of these classifiers and Influenzanet data resulted in an average correlation ratio of 0.76 for both SVM and Naïve Bayes. When the classifiers were trained with tweets marked as "cold" treated as positive data, the results improved considerably for the Naïve Bayes classifier (0.82) but only a slight change was obtained with the SVM classifier (0.77).

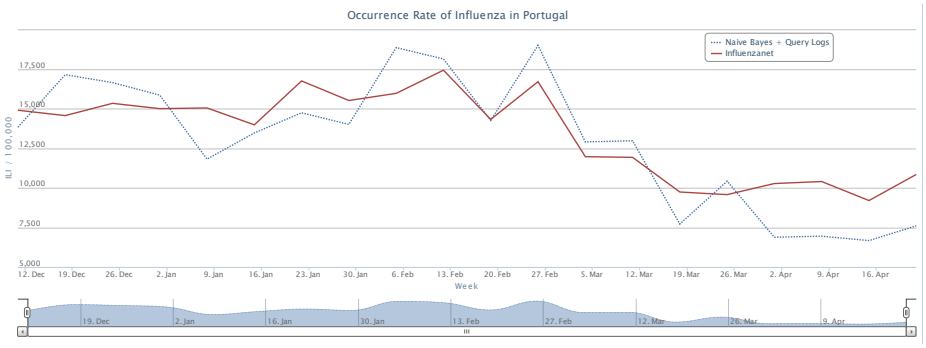

Fig. 2. Prediction results for the linear regression model using the Naïve Bayes Bayes classifier and queries.

3.2 Flu Trend Prediction

For flu trend prediction, we tested linear regression models with the relative frequencies calculated from the classification results, query logs and regular expressions as the predictors. The prediction results were obtained for data in the period from December 2011 to April 2012 (20 weeks). To avoid overlapping between training and test data, the models for tweet classification were trained with a subset of 1728 manually annotated tweets, covering the period from March 2011 to November 2011.

For each classifier, we selected from the receiver operating characteristic (ROC) analysis (Fig. 1) and based on the training data only, an operating point that maximized the classification precision without a severe loss on the classifier recall. Although operating points with higher F-measure values could have been selected, these would represent higher recall, at the expense of a lower precision. We therefore chose the more stringent models, in order to reduce the amount of false positive hits, and consequently, the amount of noise present in the final results.

Since we had limited amounts of data, we randomly partitioned the data into a training set and a test set, each covering ten data points (weeks). This was repeated ten times, and the Pearson's correlation coefficient between the predictor output and the Influenzanet rates, used as gold standard, was calculated for each partition. The average results are shown in Table 3. In these results, the Naïve Bayes classifier was used to classify the tweets. For comparison, the best regression obtained with the SVM classifier resulted in a correlation ratio of 0.849. This result was obtained with the same setting as the best result for the Naïve Bayes classifier as shown in Table 3.

Fig. 2 shows the comparison between the multiple linear regression result and the Influenzanet data for the 20 weeks considered. The regression model was trained on the initial ten weeks and applied to the entire sequence.

Table 3. Pearson's correlation ratios between linear regression estimates and Influenzanet data. *flu* and *expanded* regular expressions correspond to the pattern for "gripe" (flu) word derivations and the complete pattern as shown in Table 1, respectively. The weekly relative frequency was calculated based on the number of positively classified tweets (counts) or on the probabilities given by the classifier (Eq. 1). Tweets referring to "*cold*" were used either as positive or negative data when training the classifier.

		"*cold*" negative	"*cold*" positive
flu reg exp		0.837	
expanded reg exp		0.801	
Queries		0.571	
Queries + *flu* reg exp		0.849	
Queries + *expanded* reg exp		0.885	
Naïve Bayes	Counts	0.761	0.825
	Probabilities	0.804	0.824
Naïve Bayes + Queries	Counts	0.781	0.872
	Probabilities	0.794	**0.886**
Naïve Bayes + *expanded* reg exp	Counts	0.770	0.703
	Probabilities	0.807	0.791
Naïve Bayes + *expanded* reg exp + Queries	Counts	0.873	0.836
	Probabilities	0.867	0.801

4 Discussion

To the best of our knowledge this is the first work in this subject done specifically for the Portuguese language. Although most of the used methods are similar and applicable across languages, the amount of available data in languages other than English, as well as language specificities, may influence the final results obtained.

Despite Twitter being a largely used social web platform, it is not very popular in Portugal, which limited the size of our dataset when compared to similar works. As a comparison, we had access to around 14 million tweets, with a daily average of nearly 40,000 tweets, from which 1728 were used to train the binary classifiers. Aramaki et al. [1] used 300 million tweets, from which 5,000 were used for training. Cullota [2] used a total of 500,000 messages, selecting 206 of those messages to train a model. Due to the limited amount of used data, overfitting problems are reported in that work.

Another important novelty of our work is the integration of results from the analysis of tweets with results from user searches on a web search engine, through multiple linear regression models. This contributed to a better approximation to health monitoring results used as gold-standard in this work. A possible extension to this would be to use other sources of user-generated content, such as blog posts and comments on web pages.

The best result reached a Pearson's correlation ratio, between the estimated incidence rate and the Influenzanet data, of 0.89 ($p < 0.001$). This result indicates that this method can be used to complement other measures of disease

incidence rates. Unfortunately, the amount of data available for validating the prediction model was reduced, which may limit the relevance of the results. Further studies should be performed to validate these results and also to assess if prediction models trained in a flu season can be applied in the following season. Another important aspect to consider in further studies is whether it is possible to predict, with some advance, an increase in the incidence of flu (or other illnesses) in order to optimize the response by the health authorities.

Acknowledgements This research work was partially funded by Fundação para a Ciência e a Tecnologia (FCT), ref. PTDC/EIA-CCO/100541/2008 and by Labs Sapo, project "SPotTED - Social web Public healTh Event Detection". S. Matos is funded by FCT under the Ciência2007 programme.

References

1. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 1568-1576 (2011)
2. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the First Workshop on Social Media Analytics, ACM, 115-122 (2010)
3. Culotta, A.: Detecting influenza outbreaks by analyzing Twitter messages. arXiv:1007.4748 [cs.IR] (2010)
4. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature, 457, 1012-4 (2009)
5. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. 2010 2nd International Workshop on Cognitive Information Processing (CIP), 411-416 (2010)
6. Paul, M., Dredze, M.: You are what you tweet: Analyzing Twitter for public health. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, 265-272 (2011)
7. Joachims, T.: Making large Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press (1999)
8. Loper, E., Bird, S., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc (2009)
9. Pedregosa, F. and Varoquaux, G., Gramfort, A., Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-30 (2011)