Using Machine Learning to Analyze Twitter for Real Time Influenza Surveillance

By Justin Cai

Abstract

Twitter has been shown in previous work that it can be a reliable data source to predict disease surveillance, such as the flu rate. With that in mind, the main motivation of this project was Twitter to predict vaccination rate. The results produced correlated 2016 data with 2015 data with a R2 of 0.96. With theses results, vaccination rates can be calculated and published monthly, as opposed to the yearly official CDC data.

Introduction

With the introduction and popularity of social media and the use of the internet in the past decade increasing, researchers have been finding ways to use the publicly available data. One application of this data is using it measure influenza trends. Tracking influenza through the government can take up to 2 weeks before statistics are published. Launched in 2008, Google Flu Trends tracks the rate of influenza on a daily basis, up to 7 to 10 days faster than the Center for Disease Control and Prevention's (CDC) FluView. It is done by tracking the volume of flu related search queries (Paul and Dredze, 2011).

In the recent years, however, Twitter has been used to study flu trends. The benefit of Twitter data is free for the public to access. Researchers used this data with machine learning algorithms to predict flu rate. One of the first ways this was done was using the supervised learning method of classification. Classifiers, algorithms that sorts input data into predefined categories, used a simple model using most common keywords or n-grams as features. Features are the values that define a piece of data and an n-gram, a phrase with n amount of words, is a type of feature (Culotta, 2010). Lamb, et al. (2013) built a more advanced classifier that could differentiate between the author of the tweets being infected and general flu awareness (e.g. "Robbie might have swine flu. I'm worried." vs. "I am getting a serious case of the flu") by the analyzing verbs, nouns, and pronouns of the sentence and using machine learning (lexicographic features) . The classifier correlated well with CDC data, an R2 value of 0.9897 for the 2009 season and an R2 value of 0.7987 for the 2011 season.

Another use researchers have found for Twitter data is looking at sentiment and networks of people. Networks of opinionated users (predominantly positive or negative compared to neutral) were found to have more information flow between users with users

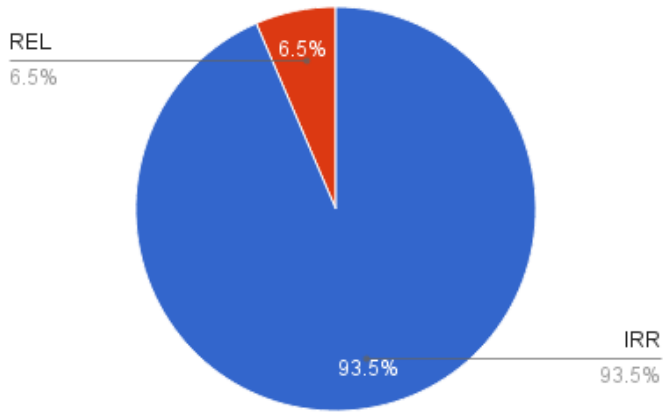
that share the sentiment compared to having different sentiments. Communities were also found to be mostly positive or negative, and not in between. Related to this work, Infectious disease outbreaks were reproduced and it was found that if groups of negative vaccine sentiment were connected to groups of unprotected individuals, the probability of outbreak increased (Salathé, and Shashank, 2011). For machine learning, one way to train the classifiers used for sentiment analysis is to create a data set using hashtags and emoticons. The accuracy of classifier depends on what features the classifier has and what data you use to train. For example, only using n-grams and hashtags and emoticons was more accurate than just hashtags but was the opposite with n-grams and lexicon features. Overall, as before, the more features, the more accurate, at the expense of time (Kouloumpis et al., 2011).

Research Question

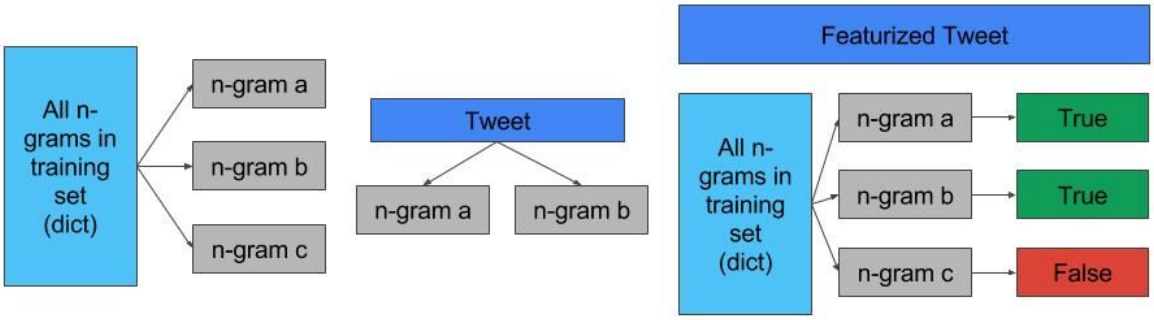
Based off of previous works, it was shown that Twitter can be a reliable source for tracking influenza rate. One topic about influenza surveillance not yet done, based off previous research, is vaccination rates on Twitter. The formal research question was: Can Twitter be used to create a model that accurately predicts vaccination rates in the US?

Methods and Analysis

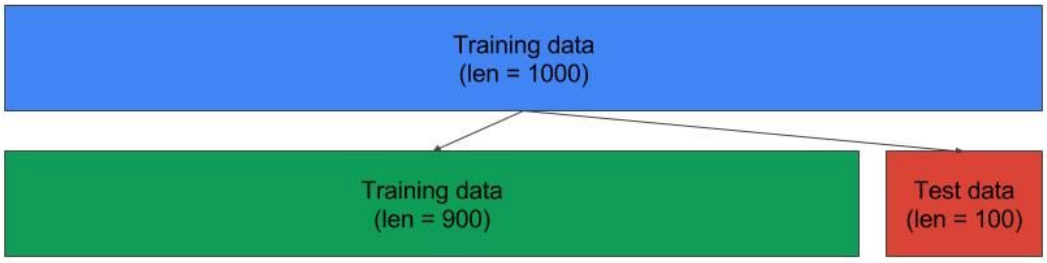
To create a model that predicts vaccination rates, a machine learning alogrthim needed be decided upon. In order to decide, experimentation was done on three different types of classifiers, logistic regression, Naive Bayes, and support vector classification (SVC), three different types of classifiers that are supported in the Python machine learning library, scikit-learn. Since those are all supervised algorithms, a training set needs to be created. From a dataset of tweets collect through Twitter's Streaming API between September 2015 - April 2016, a random sample 1000 tweets were manually labeled. After the labeling process, each tweet consisted of the text and the label. All tweets with the author expressing that they want or need a flu shot were marked as relevant (REL) and all others were irrelevant (IRR).

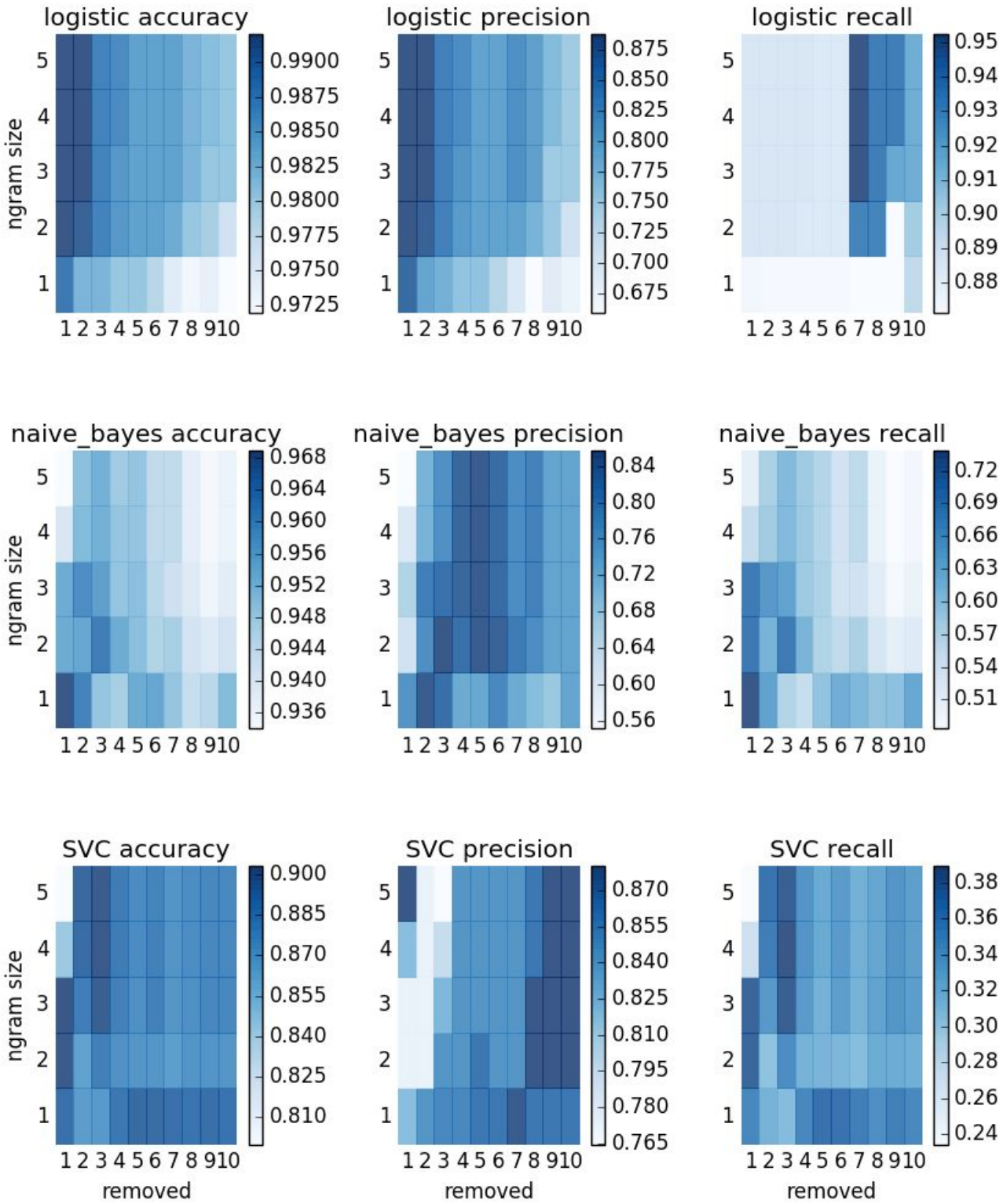


After the training set was created, the tweets needed to be featurized in a way that is a valid format for the classifiers to use. The text itself isn't really a feature, but looking at individuals n-grams is a to featurize text. In essence, text featurization works by collecting all the n-grams (dict) in the given corpus (corpus: a collection of text) and for each n-gram in the dict, check if it is in the text of interest. If is it, assign a one, and if it isn't assign a 0. For example, if there are 1000 unique words, then the feature space has 1000 dimensions, with each training example having 1000 parameters.



After the featurizing the training set with NLTK, a Python natural language processing library, each of the classifiers was tested with a 10 fold cross validation. With each type of classifier, the number of features varied through two ways: n-gram size, and removing words that didn't appear more than a certain amount of time.



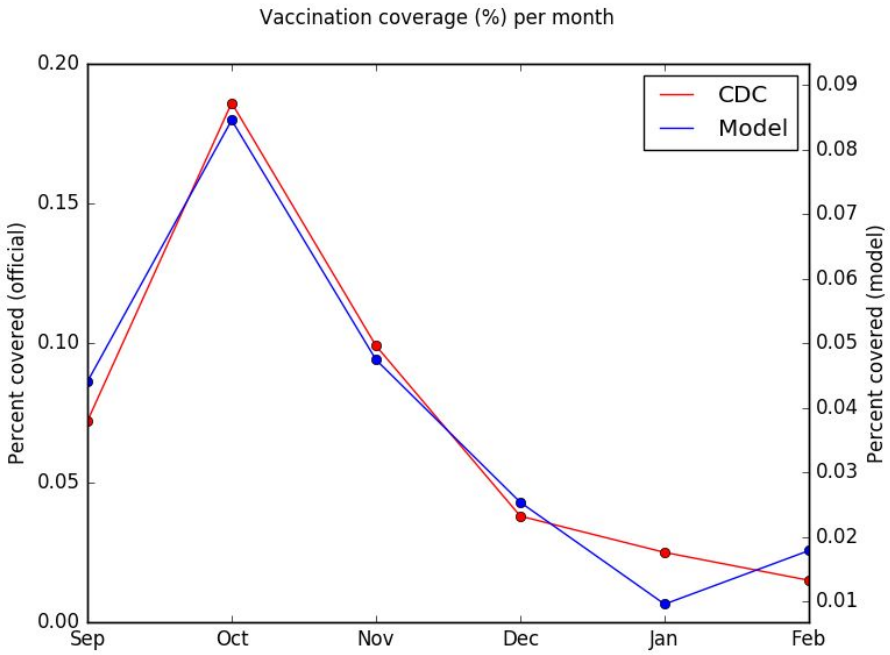

logistic accuracy
logistic precision
logistic recall
naive_bayes accuracy
naive_bayes precision
naive_bayes recall
SVC accuracy
SVC precision
SVC recall
ngram size
removed
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5
0.9900 0.9875 0.9850 0.9825 0.9800 0.9775 0.9750 0.9725
0.875 0.850 0.825 0.800 0.775 0.750 0.725 0.700 0.675
0.95 0.94 0.93 0.92 0.91 0.90 0.89 0.88
0.968 0.964 0.960 0.956 0.952 0.948 0.944 0.940 0.936
0.84 0.80 0.76 0.72 0.68 0.64 0.60 0.56
0.72 0.69 0.66 0.63 0.60 0.57 0.54 0.51
0.900 0.885 0.870 0.855 0.840 0.825 0.810
0.870 0.855 0.840 0.825 0.810 0.795 0.780 0.765
0.38 0.36 0.34 0.32 0.30 0.28 0.26 0.24

The results from the classifiers turned out well, but varied from inital predictions. Based of previous research, SVC would have the highest accuracy, logistic regression second, and Naive Bayes last. One reason why SVC could of not scored as well is because of the imbalanced data set. Only 6.5% of the training set was relevant, so having a near 50-50 split of relevant and irrelevant could make SVC perform better. Logistic regression performed very well. An explanation for that is that logistic regression is a generalized linear model, and text categorization is linearly separable, therefore, it performed well (Joachims, 1998).

From these results, it shows that logistic regression overall has the best scores. From this, logistic regressions with feature parameters n-grams: 4, removed: 7 was used to classifier the rest of the data. Then for each month, this expression was calculated:

$$\frac{REL}{REL+IRR}$$

which is the percentage covered per month. The data from the CDC comes total coverage per month, so the difference between was calculated between months to get coverage per month. Finally, as mentioned before, vaccination data is only release yearly, so the correlation is relation with the 2015 data. The results are figured below. The model had an R2 value of 0.9634 with the official data. When the axis of the model is scaled to match that of the official data, the correlation is clear.



Once further testing and similar results are produced, the conclusion that Twitter can be used to track flu vaccination rates can be made. The results produced in this project are promising. In the future, acquiring datasets of tweets from pasts years and classifying them well help back these results up. Once the CDC releases data for the 2015-2016 season, the 2016 data can be correlated to the model's prediction of the 2016 data. Other future endeavors on this project are to see what results different classifiers produce. Another would be use another scaling features by inverse document frequency, instead of using 1's and 0's.

Acknowledgements

I would like to thank Dr. Michael Paul at the University of Colorado Boulder for helping me get started with this project and teaching me about machine learning. I would also like to acknowledge all the wonderful Python libraries used in this project that helped achieve my goal in this project: NumPy, SciPy, matplotlib, scikit-learn, and NLTK.

Works Cited

Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." Proceedings of the first workshop on social media analytics. ACM, 2010.

Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer Berlin Heidelberg, 1998.

Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11 (2011): 538-541.

Lamb, Alex, Michael J. Paul, and Mark Dredze. "Separating Fact from Fear: Tracking Flu Infections on Twitter." HLT-NAACL. 2013.

Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." ICWSM. 2011.

Salathé, Marcel, and Shashank Khandelwal. "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control." PLoS Comput Biol 7.10 (2011): e1002199.