Implementation of Text Classifiers Using Learning by Induction Approach - Case Study of Twitter Data

Falade Adesola, Odusote Babafemi, and Isaac Odun-Ayo, *Member, IAENG*

***Abstract*— The amount of data currently residing on social media is not sufficiently tapped and is certainly limitless as millions of people are constantly posting one message or the other to these public forums on the Internet. Twitter is one of the largest social media network with over 320 million monthly active users which has proven to be a fertile ground for harvesting opinion from several people to influence decision-making process within organizations and institutions. Based on a thorough review of literature and past work in the area of text mining and twitter sentiment analysis, a system was developed which applied three different supervised machine learning algorithms to a dataset curated by graduate students at Stanford University in order to accurately classify tweets into either positive or negative sentiment based on its content. The result showed that Maximum Entropy has the highest accuracy of 83.5% among the three algorithms.**

Based on further analysis and research it was discovered that the classifiers could be improved upon. Using this as a basis, the authors then implemented a system that learns from wrong classification as corrected by the users. This paper presents the results from this research.

***Index Terms*—Machine Learning, Learning by Induction, Supervised Learning Algorithms, Twitter, Naïve Bayes, Support Vector Machine, Maximum entropy.**

I. Introduction

DATA is generated by every aspect of our daily lives, and it is equally shared with or without knowing it. From sharing that post on Facebook, to ordering that item from Amazon or even just visiting a website, data is collected and stored but what exactly that is done with data is what actually matters.

The measure of data in reality increases at an exponential rate, and can be seen in all parts of day by day lives, from data about buyers gathered by organizations to posts by people on online networking [1]. It is imperative that the availability of this data optimally maximized. A lot of work and research has gone into extracting meaningful information from twitter data, for forecasting of events by tweet data mining. Few of such work was carried out by [2], where he analyzed quantitative characteristics of frequent sets and association rules in tweets relating to various events, to using Twitter to predict football outcomes in [3] by building predictive models based on tweets about the English Premier League over a period of three months. Others include predicting events such as outcomes of elections in [4], stock market prices prediction in [5], box-offices revenues of movies before release in [6] and the spread of diseases in [7], among others.

The introduction of sentiment analysis of tweet data to any process of twitter data mining has proven to provide a much more accurate representation of data as discovered in a research in [8] in which they were able to successfully predict the outcome of the US Primary elections in [20] based on twitter data. Previous methods for opinion/sentiment analysis have done little in implementing supervised learning algorithms with learning by induction in Twitter data analysis.

II. Related Work

A. You are what you Tweet: Analyzing Twitter for Public Health

Following the recent work which evaluated Twitter messages with influenza rates in the United States, Michael and Mark aim to consider a wider range of public health applications for Twitter. They applied the recently presented Ailment Topic Aspect Model to more than 1.5 million wellbeing related tweets and discussed over twelve diseases, including hypersensitivities, weight and sleeping disorder. They introduced expansions to incorporate earlier learning into this model and applied it to a few assignments: following diseases over times (syndromic observation), measuring behavioral danger components, limiting diseases by geographic district, and breaking down indications and medicine use. They demonstrated quantitative relationships with general wellbeing information and subjective assessments of model yield. The outcomes recommend that Twitter has wide relevance for general wellbeing research [7].

B. Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers

The researchers in [9] noticed that competition existed

Manuscript received October, 2018; revised, 2018. This work was supported in part by Center of ICT/ICE Research, CUCRID Building, Covenant University, Ota, Nigeria

A. Falade Adesola is with the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria (phone: +2348033275556; e-mail: adesola.falade@covenantuniversity.edu.ng).

B. Odusote Babafemi is with the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria (e-mail: odusote.babafemi@covenantuniversity.edu.ng).

I. Odun-Ayo is with the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria (e-mail:isaac.odun-ayo@covenantuniversity.edu.ng)

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

among Indonesian mobile phone providers to gain new clients through adverts, particularly on social networks. The issue arose on the most capable strategy for determining the brand reputation of these suppliers taking into account singular response on their service quality.

The paper addressed this issue by measuring brand reputation considering purchaser devotion through client's sentiment analysis from Twitter data. Test model was collected and extricated from 10,000 unrefined tweets from January to March 2015 from the three major mobile phone suppliers in Indonesia namely: PT XL Axiata Tbk, PT Telkomsel Tbk, and PT Indosat Tbk.. They analyzed a couple of features extractions, algorithms, and the classification schemes. After data cleaning and data adjusting, the sentiments were grouped and analyzed utilizing three unique algorithms: Naïve Bayes, Support Vector Machine, and Decision Tree classifier method.

They were able to measure consumer loyalty on five items: 3G, 4G, Short Messaging, Voice and Internet services. In addition, the paper considered some associated business bits of information in the telecom administrations industry. In perspective of the general connection of these four things namely: the NBR scores for PT XL, Axiata Tbk, PT Telkomsel Tbk, and PT Indosat Tbk were 32.3%, 19.0%, and 10.9% respectively.

C. Using Twitter to Predict Football Outcomes

The reason for this research was to determine if data mined from Twitter can be utilized for the purpose of designing a predictive model for football (Soccer). They were able to develop a set of predictive models for the result of football games in the English Premier League for a three month period. In view of the tweets, they considered whether these blueprint can overcome predictive models which utilize just historical data and straightforward football insights.

In addition, joint models were built utilizing both Twitter and historical data. The last results showed that information mined from Twitter can be a helpful hotspot for anticipating games in the Premier League. The last Twitter-based model performs significantly better when measured by Cohen's kappa and is tantamount to the model that uses straightforward insights and historical data. Subsequently, this study confirmed that Twitter determined components can in reality provide helpful data to the expectation of football (soccer) results [3].

III. Machine learning

Machine learning explores how computers can learn (or enhance their execution) in view of data. A principal research territory is for computer projects to naturally determine how to perceive complex examples and settle on savvy choices taking data into account. For instance, a run of the mill machine learning issue is to program a computer with the goal that it can consequently perceive written by hand postal codes on mail subsequent to gaining from an arrangement of samples [10].

A. Supervised Learning

Supervised learning is the machine learning assignment of inferring a function from supervised training data set. The supervised learning algorithms used in this research work are:

1) Naïve Bayes Classifier: The Naïve Bayes classifier proffers a straightforward yet effective supervised learning classification strategy. The classifier model accepts all given input variables as being equivalent in significance and autonomous of each other. Naive Bayes classifier depends solely on the established Bayes hypothesis which is related to the probability theory. In basic terms, a naive Bayes classifier accept that the appearance (or non-appearance) of a specific feature of a class is irrelevant to the appearance (or non-appearance) of whatever other feature. Despite the fact that these suppositions are prone to be false, Bayes classifiers are found to be very efficient.
2) Support Vector Machine: A Support Vector Machine (SVM) is a discriminative classifier formally characterized by an isolating hyperplane. In practice, when given marked training information (supervised learning), the algorithm yields an ideal hyperplane which orders new illustrations.
3) Maximum Entropy: The Max Entropy classifier is a probabilistic classifier which has a place with the class of exponential models. Not at all like the Naive Bayes classifier earlier discussed, the Maximum Entropy depends on the Principle of Maximum Entropy and from all the models that fit the training data, of this study chooses the one that has the biggest entropy.

IV. System Design and Structure

The entire process of data mining is the continual repetition of the following three steps as shown in Figure 1:

1. Gathering the data, which are recent tweets from twitter.
2. Performing the mining operation using various data mining techniques.
3. Presenting the results.

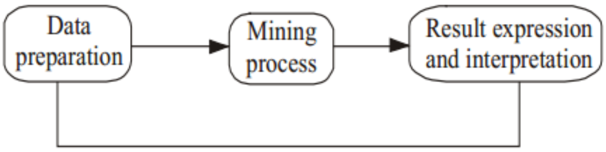

Fig. 1. The Entire Process of Data Mining

A. Classifier Structure

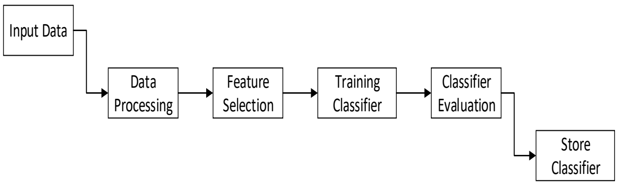

Fig. 2. Classifier Training

Figure 2 shows the workflow of the process of training the classifier.

1) Data Description: The training set of tweets in this study were extracted from the work in [11] on Sentiment140

project work. The Distant Supervision method was used in gathering the tweets over a three month period and it serves as one of the most reliable source of Twitter training Corpus. It is made up of a total of 1.6 million tweets shared evenly among the two polarity types (Negative & Positive).

2) Data processing also known as preprocessing in the Text mining sphere refers to the removal of noise from text. This is performed before the actual classification or learning takes place, this is essentially important especially for text that is unstructured. Some preprocessing techniques applied are:

a) Removal of duplicates: All repeated tweets or retweeted tweets (RT) are removed.
b) Case Folding: All words are converted to lowercase
c) Removal of Punctuation and Special Characters: Web Addresses, RT, punctuations, usernames are removed.
d) Stop word Removal: Common English words known as stop words for example (‘and’, ’the’, ’us’, etc.) are removed from the text were employed.

3) Feature Selection (n-grams): Data during the feature selection, the technique of selecting n-grams was used. An n-gram is a contiguous succession of n things from a given text or speech. The things can be phonemes, syllables, letters, words or base sets as indicated by the application. The n-grams commonly are gathered from a text or speech corpus. At a point when the things are words, n-grams may also be called shingles. An n-gram of size 1 is alluded to as a "unigram"; size 2 is a called "bigram"; size 3 is a "trigram". Bigger sizes are here and there alluded to by the estimation of n, e.g., "four-gram", "five-gram", and so on. In this system, the use of both ‘Unigrams’ and ‘Bigrams’ were employed.

4) Classification: In this classification step, different algorithms namely Naïve Bayes, Maximum Entropy and Support Vector Machine were modelled with the same data source. A total of 1.6 million Tweets were used to build the data set. The training tweet source was collated by students of Stanford University for the sentiment140 project in [11] and was created using the method of distant supervision.

5) Classifier Evaluation: This is the process of determining the accuracy of the classifier by splitting the initial test set into both test set and training set thereby calculating how accurate the classifier is based on how much of the data was classified correctly.

6) Store Classifier: The model of the classifier is stored in a file such that when it would be used, it does not have to be trained again.

B. Incorrect Classification/ Learning by Induction

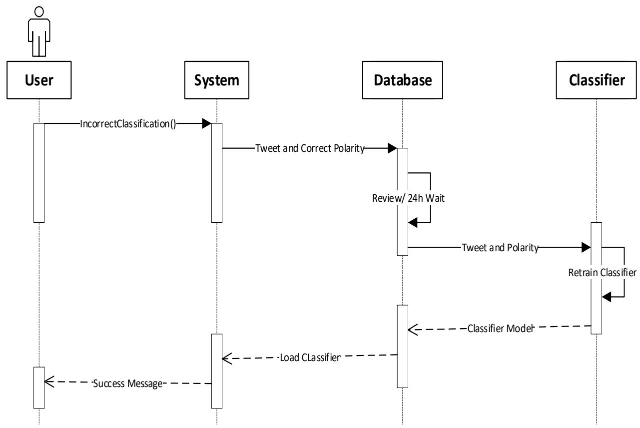

Fig 3: Sequence diagram for incorrect classification

The sequence diagram in figure 3 shows the steps that would be taken if the system classifies a tweet wrongly. Initially, the user signifies to the system that a particular tweet is wrongly classified by providing its actual sentiment (Positive/ Negative). The system then preprocesses the text and stores it. Due to the time taken for training, it is done once every 24 hours and the model for each classifier is stored again, overwriting the previous one.

V. IMPLEMENTATION

The Tools used in the implementation of the system include:

1. Python: Python is a broadly used high-level, universally useful, interpreted, dynamic programming dialect. Its outline rationality underscores code decipherability, and its linguistic structure permits software engineers to express ideas in less lines of code than would be conceivable in languages, for example, C++ or Java. Python was the choice programming language due to its wide array of packages available for machine learning such as Scikit-Learn, Numpy, and Scipy amongst others. Also partially due to the fact that it was easier to develop a web application from it rather than languages such as Java or R.
2. HTML: Hyper-Text Markup Language, regularly alluded to as HTML, is the standard markup dialect used to design website pages. Alongside CSS (Cascade Style Sheet), and JavaScript, HTML is a foundation innovation used to make pages, and additionally to make user interfaces for portable and web applications (Flanagan).
3. SQL: SQL (Structured Query Language) is special programming language use for administering information and data stored in a relational database management system (RDBMS), or for streaming in a relational information stream management system (RDSMS).
4. PostgreSQL: is an object-relational database management system (ORDBMS) with an accentuation on extensibility and measures consistence. It can deal with workloads running from little single-machine applications to substantial Internet-confronting applications with numerous simultaneous users.
5. Python Machine Learning Packages: Packages are namespaces which contain multiple packages and modules themselves, the machine learning, numerical & scientific

packages used in this project research includes:

a) SciKit-Learn: It is an open source machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN (Density Based Spatial Clustering of Application with Noise) – a popular unsupervised learning algorithms.

b) NumPy: It is an augmentation to the Python programming language, including support for vast, multi-dimensional clusters and lattices, alongside an extensive library of high-level numerical capacities to work on these exhibits. A required library for use by SciKit-Learn.

c) SciPy It is an open source Python library used by researchers, examiners, and designers doing scientific computing and technical computing. Also a required by Scikit-Learn.

d) Pandas: Pandas is an open source, (Berkeley Software Distribution) BSD-authorized library giving high-execution, simple to-use information structures and information investigation devices for the Python programming language.

6. Flask: Flask is a miniaturized scale web system written in Python and taking into account the Werkzeug toolbox and Jinja2 template engine. Other options include Django, Bottle and Web.py but flask was favored because of its convenience and straight forward execution.

7. PyCharm: PyCharm is an Integrated Development Environment (IDE) used for programming in Python. It gives code investigation, a graphical debugger, a coordinated unit analyzer, mix with rendition control systems, and backings web improvement with Django and Flask. PyCharm was produced by the Czech organization JetBrains. It is a cross-platform application that can be used in Windows, Mac OS X and Linux.

VI. RESULT

After eight tweets were corrected by the user for wrongful classification, the accuracy of the Support Vector Machine classifier increased from 79.80% to 80.17% as seen in figure 4. The result also showed that Maximum Entropy outperform both Naive Bayes and Support Vector Machine in Tweets classification.

1	Maximum Enthropy	83.52%
2	Naive Bayes	79.89%
3	Support Vector Machine	80.17%

Fig 4: Current accuracy of the classifiers

VII. CONCLUSION

This study has shown that Twitter contains enough information useful for predictive modeling in both Text mining and Sentiment analysis implementation. If the system is left for public use with an administrator in charge of the incorrectly classified tweets, a much more significant improvement on the accuracy of the classifiers can be seen. Nevertheless, conclusion of the study is the fact that Twitter contains information which is useful for predictive modeling. There is still so much to be done in the field of Text mining.

ACKNOWLEDGMENT

We acknowledge the support and sponsorship provided by Covenant University through the Centre for Research, Innovation and Discovery (CUCRID).

REFERENCES

[1] McKinsey Global Institute, "Big data: The next frontier for innovation, competition and productivity," McKinsey Global Institute, 2011.

[2] B. Pavlyshenko, "Forecasting of Events by Tweet Data Mining," 2013.

[3] S. Kampakis and A. Adamides, "Using Twitter to predict footbal outcomes," arXiv, 2014.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Munich, 2010.

[5] J. Bollen, H. Mao and X.-J. Zeng, "Twitter mood predicts the stock market," in Journal of Computational Science, 2011.

[6] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," CoRR, 2010.

[7] M. J. Paul and M. Dredze, "You are what you Tweet: Analyzing Twitter for public health.," ICWSM, pp. 265-272, 2011.

[8] L. Shi, N. Agarwal, A. Agrawal, R. Garg and J. Spoelstra, "Predicting US Primary Elections with Twitter," 2012.

[9] N. A. Vidya, M. I. Fanany and I. Budi, "Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers," in The Third Information Systems International Conference, 2015.

[10] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2012.

[11] A. Go, R. Bhayani and L. Huang, "Twitter Sentiment Classification using Distant Supervision," 2011.