Naive Bayes Approach for Feature Selection of Data Stream over Cloud

A. M. Wade[1] Rutuja Shintre[2] Rohan Mandhare[3] Shrikant Dawkhar[4]
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]Smt. Kashibai Navale College of Engineering, India

Abstract— The System uses Feature Selection on live stream data and applies filtering for final results. It does job title classification and job post classification that utilizes machine learning. Machine learning based job type classification techniques for text and related entities have been well researched in academic and also have been successfully applied in many industrial settings. Digital recruitment is a popular online method that has been widely used for attracting individuals who are seeking for job opportunities.
Key words: Feature Selection, Classification, Twitter
General Terms: Big Data, Hadoop, HDFS

I. INTRODUCTION

Feature Selection attempts to identify the best subset of variables (or features) out of the available variables (or features) to be used as input to a classification or prediction method. The main goals of Feature Selection are: to clean the data, to eliminate redundancies, and to identify the most relevant and useful information hidden within the data, thereby reducing the scale or dimensionality of the data. Feature Selection results in an enhanced ability to explore the data, visualize the data, and to make some previousl infeasible analytic models feasible. It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple Feature Selection results in an enhanced ability to explore the data, visualize the data, and to make some previousl infeasible analytic models feasible. It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature

Posterior = prior X likelihood / evidence

$$P(A/B) = P(B/A) \text{ X } P(A) / P(B)$$

The proposed system has a novel approach, a machine learning based semi supervised job title classification system. It leverages a varied collection of classification and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in cascade classification. Architecture of the system, which consists of a two stage Capture with filtration and fine level classification algorithm. The amount of high-dimensional data that exists and is publically available on the internet has greatly increased in the past few years. Therefore, machine learning methods have difficulty in dealing with the large number of input features, which is posing an interesting challenge for researchers. In order to use machine learning methods effectively, pre-processing of the data is essential. Feature selection is one of the most frequent and important techniques in data pre-processing, and has become an indispensable component of the machine learning process. It is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the process of detecting relevant features and removing irrelevant, redundant, or noisy data. This process speeds up data picture, and Table captions should be centered above the table body. The smaller size of the subset that satisfies a certain restriction on evaluation measures In general, the subset with the best commitment among size and evaluation measure. In text mining, the standard way of representing a document is by using the bag-of-words model. The idea is to model each document with the counts of words occurring in that document. Feature vectors are typically formed so that each feature (i.e. each element of the feature vector) represents the count of a specific word, an alternative being to just indicate the presence/absence of a word without specifying the count. The set of words whose occurrences are counted is called a vocabulary. Given a dataset that needs to be represented.

II. BACKGROUND

Feature Selection or attribute selection is a process by which you automatically search for the best subset of attributes in your dataset. The notion of "best" is relative to the problem you are trying to solve, but typically means highest accuracy. A useful way to think about the problem of selecting attributes is a state-space search. The search space is discrete and consists of all possible combinations of attributes you could choose from the dataset. The objective is to navigate through the search space and locate the best or a good enough combination that improves performance over selecting all attributes.
Three benefits of performing feature selection on your data are:

A. Reduces Over fitting

Less redundant data means less opportunity to make decisions based on noise. Improves Accuracy: Less misleading data means modelling accuracy improves. Reduces Training Time: Less data means that algorithms train faster. Privacy Issues-The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviours trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak. Security issues-Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available,

the credit card stolen and identity theft become a big problem. Misuse of information/inaccurate information-Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people. In text mining, the standard way of representing a document is by using the bag-of-words model. The idea is to model each document with the counts of words occurring in that document. Feature vectors are typically formed so that each feature (i.e. each element of the feature vector) represents the count of a specific word, an alternative being to just indicate the presence/absence of a word without specifying the count. The set of words whose occurrences are counted is called a vocabulary. Given a dataset that needs to be represented, one can use all the words from all the documents in the dataset to build the vocabulary and then prune the vocabulary using feature selection. Feature selection is important in fault diagnosis in industrial applications, where numerous redundant sensors monitor the performance of a machine have shown that the accuracy of detecting a fault (i.e. solving a binary classification problem of machine state as faulty vs. normal) can be improved by using feature selection. They proposed to use a global geometric model and a similarity metric for feature selection in fault diagnostics. The idea is to find feature subsets that are geometrically similar to the original feature set. The authors experimented with three different similarity measures: angular similarity, mutual information and structure similarity index. The proposed approach was compared with distance-based and entropy based feature selection, and with SVM and neural network wrappers. The best performance was obtained by combining the proposed geometric similarity approach with a wrapper, so that top 10% of feature subsets were preselected by geometric similarity, following by an exhaustive search-based wrapper approach to find the best subset.

III. FEATURE SELECTION OF DATA STREAM

The need for parallel processing of the massive volume of data was required, which could efficiently analyse the Big Data. For that reason, the proposed unit is introduced in the real time Big Data processing framework that gathers the massive volume of data from Twitter. Some relational data pre-processing techniques are data integration, data cleaning, and redundancy elimination. Tweepy - An easy-to-use Python library for accessing the Twitter API. It can be used with Python (2.6, 2.7, and 3.x). Tweepy is a Python 2.6, 2.7, and 3.x library for accessing Twitter. It provides access to all Twitter RESTful API methods, including reading and posting of tweets. Twitter is a social networking application which allows people to micro-blog about a broad range of topics. Micro-blogging is defined as "a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web.". Twitter helps users to connect with other Twitter users around the globe. The messages exchanged via Twitter are referred to as micro-blogs because there is a 140 character limit imposed by Twitter for every tweet. This lets the users present any information with only a few words, optionally followed with a link to a more detailed source of information. Therefore, Twitter messages, called as "tweets" are 14 usually focused. In this regard, Twitter is very similar to SMS (Short Message Service) messages exchanged via mobile phones and other hand held devices. In fact, the 140- character limit on message length was initially set for compatibility with SMS messaging, and has brought to the web the kind of shorthand notation and slang commonly used in SMS messages. The 140 character limit has also spurred the usage of URL shortening services such as bit.ly, goo.gl, and tr.im, and content hosting services to accommodate multimedia content and text longer than 140 characters. Several other social networking sites like Facebook [8], Orkut introduced the concept of "Status" messages, some much before Twitter originated. But it was Twitter that went a step ahead and made these "statuses" be sharable between people through mobile phones since its creation. A Twitter user „A" is a person or a system who publish tweets. These tweets are by default public to any user of the system unless the author specifically sets it to be private. All users of a system are identified by a unique user name and user id. When „A' initially registers to Twitter, he has no tweets or no followers or friends (concepts explained later) to begin with. „A" can start posting tweets but these tweets are not read by other users since the user does not have any followers yet. Once 'A' identifies another user (say 'B') to follow, his tweets are visible to 'B'. Consequently, „A" becomes a follower of „B". Thus, „B" becomes a friend of „A". Note that, however, friendship need not be two-way. It is possible that „A" is not a friend of „B" if „B" does not follow the tweets of „A". Thus there exists an asymmetrical relationship between users of Twitter. Twitter also imposes a limit of 2000 friends for a particular user but there are no restrictions on the number of followers since users do not have any control of the number of followers they have. The following information is also optionally stored for each user:

- Language of tweets of the user
- Time zone of the user's location
- Tweet location: the location from which the tweet was tweeted 18
- User's profile picture
- User's location
- User's web page
- Short biography of the user

Tweet A tweet is a Twitter message. It is short message since it is restricted to be within 140 characters by Twitter. This restriction enforces the users to be concise in what they have to say. This is also the reason why users tend to use word shortenings (Eg: "fr"-for, "cud" – could) and abbreviations. Interestingly enough, there is a rich and well understood set of abbreviations which is surprisingly consistent across user groups, and even across other electronic mediums such as SMS and chat rooms [10]. Since users want to convey all they have to say within 140 characters, they could also make spelling mistakes and tweets can be prone to syntactic errors. This makes Twitter a challenging medium to work with. Most of the times, users usually provide links to external resources when they cannot convey the complete information within 140 characters.

These URL links to text, audio or video files are referred to as "Artifacts".

A. Why Mine Twitter?

Twitter serves as a rich source of information. Unlike other information sources, Twitter is up-to-date and reflects the current news and events happening around the world. Conventional news agencies often employ reporters and journalists to gather news. The quality and content is constrained by the number and the type of journalists. Also, once news is published, Web spiders must be updated to crawl for the latest information. On the other hand, Twitter provides information by having millions of users serve as reporters. News or Events here could be global, which means messages that can be understood by a large group of audience or it could be local, which is understood by a small group of people or even one specific individual. It refers to this principle as "push-pull" where on Twitter, information is "pushed" automatically to the users rather than the users "pulling" the necessary information from the web. 22 Another important feature about Twitter is that there is minimal time lag between the time of occurrence of an event and the tweet publication time. Hence, information is conveyed rapidly to users. Tweet/hour relating to Michael Jackson's death. The first tweet regarding the death was reported 20 minutes after the 911 call, which was almost an hour before conventional news media reported the death [10]. Tweets originating from users not only talk about news but a wide range of topics. Java mentions that the user intentions on Twitter include daily chatter, conversations, sharing information/URLs, and reporting news. Since, there is a lot of rich information being transmitted across the globe at a rapid rate, there is a need to mine 23 knowledge from it and provide users with a system that helps them to understand and comprehend the information as easily and quickly as possible.

B. Tweet Classification

On Twitter, tweets are presented to the user in a chronological order. This format of presentation is useful to the user since the latest tweets from the user‟s followers are rich on recent news which is generally more interesting than tweets about an event that occurred long time back. But the major drawback of this approach is that tweets arrive at a furious rate. Merely, presenting tweets in a chronological order may be too overwhelming to the user. Also, if the user has many friends out of whom, few tweet at a rapid rate compared to other friends, the dominant friend takes a lot of the user's space. Hence, tweets from the lesser dominant friends may be lost in the overwhelming tweet 25 stream. Due to these issues, there is a need to separate the tweets into different categories and then present the categories to the user.

IV. FUNCTIONALITY

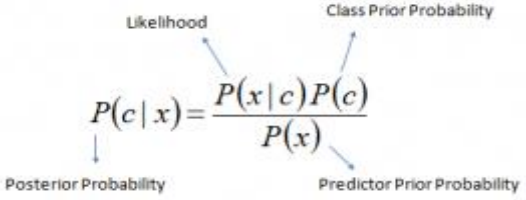
The System uses Feature Selection on live stream data and applies filtering for final results. It does job title classification using naïve Bayes classification and job post classification that utilizes machine learning. Machine learning based job type classification techniques for text and related entities have been used. The proposed system takes live data stream from twitter and applies machine learning based semi supervised job title classification system. The system takes the live input using twippy API. It leverages a varied collection of classification and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in cascade classification. Time will also be a crucial element looking at the amount of data that is being generated on the Web today. Collecting opinions and event wise data on the web will still requires processing that can filter out un-opinionated user-generated content and also to test the trustworthiness of the data and its source. The System requires a response time of 1000 tweets per second. As the data to be used is taken from live feed the database used by the system is Mongo DB. The basic requirement of Mongo DB database is client controller and server controller. It will require a minimum ram of 4GB and a memory of 20GB. The main expectation of user from the System is to acquire a generalised result. The result must be heterogeneous and classified. User also expects the response time of the System to be quick. The System should function without generating any fault and should have a stable bandwidth. The main factor responsible for the reliability of the proposed system is the computing power and its processing ability. The system has a processing power of 100mbps. The proposed system makes use of Machine learning to generate results thus it has a very high availability. As it uses Semi-Supervised learning methodology it does computational task even when the user is unaware. The cloud system known as Linode cloud system is used for all the computational processing. As the system uses Cloud Computing it is highly portable and can be accessed from anywhere.

USER	CHARACTERISTICS
ADMIN	Data Analysis Data Processing Job Searching Code Changes
USER	Data Processing Job Searching

Table 1:

A. Algorithm

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

V. CONCLUSION

The proposed system takes live data stream from twitter and applies machine learning based semi supervised job title classification system. It leverages a varied collection of classification and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories.

ACKNOWLEDGEMENT

The authors are thankful for the guidance and support from Prof. A.M.Wade.

REFERENCES

[1] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993
[2] P.-F. Pai and T.-C. Chen, "Rough set theory with discriminant analysis in analyzing electricity loads," Expert Syst. Appl., vol. 36, pp. 8799–8806, 2009.
[3] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining data streams: A review," ACM SIGMOD Rec., vol. 34, no. 2, pp. 18–26, Jun. 2005.
[4] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," SIGKDD Explorations, vol. 14, no. 2, pp. 1–5, Dec. 2012.
[5] A. Murdopo, "Distributed decision tree learning for mining big data streams," Master's of Science thesis, European Master Distrib. Comput., Jul. 2013
[6] S. Fong, X. S. Yang, and S. Deb, "Swarm search for feature selection in classification," in Proc. 2nd Int. Conf. Big Data Sci. Eng., Dec. 2013, pp. 902–909.
[7] L. Rokach, and O. Maimon, "Top-down induction of decision trees classifiers-a survey," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 35, no. 4, pp. 476–487, Nov. 2005.
[8] C. C. Aggarwal Data Streams: Models and Algorithms, vol. 31. New York, NY, USA: Springer, 2007.
[9] P. Domingos, and G. Hulten "Mining high-speed data streams," in Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, 2000, pp. 71–80.
[10] B. Pfahringer, G. Holmes, and R. Kirkby, "New options for Hoeffding trees," in Proc. Australian Conf. Artificial Intell., 2007, pp. 90–99.
[11] J. G. Cleary and L. E. Trigg, "K_:An instance-based learner using an entropic distance measure," in Proc. 12th Int. Conf. Mach. Learning, 1995, pp. 108—114.
[12] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in Proc. SIAM Int. Conf. Data Mining, 2007,
[13] I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, "Active learning with evolving streaming data," in Proc. Eur. Conf. Mach. Learning Knowl. Discovery Databases, 2011, vol. 3, pp. 597–612.
[14] S. Fong, S. Deb, X.-S. Yang, and J. Li, "Metaheuristic swarm search for feature selection in life science classification," IEEE IT Prof. Mag., vol. 16, no. 4, pp. 24–29, Aug. 2014.
[15] X.-S. Yang, S. Deb, and S. Fong, "Accelerated particle swarm optimization and support vector machine for business optimization and applications," in Proc. 3rd Int. Conf. Netw. Digital Technol., Macau, China, Jul. 11–13, 2011, pp. 53–66.
[16] S. Fong, J. Liang, R. Wong, and M. Ghanavati, "A novel feature selection by clustering coefficients of variations," in Proc. 9th Int. Conf. Digital Inf. Manag., Sep. 29, 2014, pp. 205–213
[17] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Mateo, CA, USA: Morgan Kaufmann, 2005

