A NON-HEURISTIC MACHINE LEARNING APPROACH FOR CLASSIFYING TWITTER CONTENT

By

SRI KANTH KARUPARTHY

Bachelor of Technology in Computer Science

Jawaharlal Nehru Technological University Kakinada

Kakinada, Andhra Pradesh, India

2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2015

A NON-HEURISTIC MACHINE LEARNING

APPROACH FOR CLASSIFYING TWITTER CONTENT

Thesis Approved:

Christopher John Crick

Thesis Chair And Advisor

David Cline

Member

Nohpill Park

Member

Name: SRI KANTH KARUPARTHY

Date of Degree: MAY, 2015

Title of Study: A NON-HEURISTIC MACHINE LEARNING APPROACH FOR CLASSIFYING TWITTER CONTENT

Major Field: COMPUTER SCIENCE

Abstract: In online social networks like Twitter, the users usually get inundated with the continuous stream of short messages or tweets. This problem can be handled using classification. Classification is a supervised data mining technique which involves assigning a label to a set of unlabeled objects. A conventional approach for classifying text or tweets is to extract features from the linguistic content posted by the users. A recurrent problem in classification is feature selection, that is, to decide the best set of features for making a particular classification decision among the infinite possible different sets of features. This process usually involves heuristic approaches that require manual feature selection by experts, which involves guesswork, prior information about the dataset and a great deal of tweaking and experimental validation. To address this problem we propose and employ a non-heuristic machine learning approach which will automatically decide the feature set for a classification task. Our analysis shows that our automated feature selection process for Twitter content classification performs on par with current state-of-the-art approaches which incorporate painstaking, time-consuming human effort to manually and heuristically select a feature set. This approach will improve the timeliness and accessibility of data mining social media data streams.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Emergence of Online Social Networks and increasing popularity of it has been phenomenal. Online social networks provide people with an open platform to share information and opinions on diverse topics. Twitter is a micro-blogging service, which has gained popularity as one of the prominent news source and information dissemination agent over last few years. The content on Twitter can provide rich information, however the users may get inundated by the raw information on Twitter. This problem can be handled using classification. For any classification task on Twitter there exists a conventional approach of manually extracting features from the linguistic content posted by the users. This approach normally includes heuristic methodologies that oblige manual feature selection that is to decide the optimal set of features from the infinite set of features. Experts device several approaches to select the most informative features that involves lot of guess work, prior information of the data.

The aim of this research is to automate the feature selection process for Twitter content classification. We propose and incorporate a non-heuristic machine learning approach which will automatically decide best set of features for a classification task. Our experimental results show that classification made by automating the feature selection process does not hamper the classification accuracy and it is on par with any classification task which involves explicit selection of features.

1.1 Classification

Classification is a supervised data mining technique which involves assigning a category to a set of instances. Based on the number of categories or the classes, classification can be divided into two major types.

- Binary classification
- Multiclass classification

Binary classification is the simplest type of classification problem. In binary classification the target attributes has only two possible values, whereas multiclass problem deals with classification of instances into more than two possible values. Unlike binary classification, multiclass classification is an intricate technique.

Simple applications for classification are.

- Classifying an e-mail as Spam or Not-Spam.
- Classifying a fruit as Mango, Apple, and Banana etc.

1.2 Machine learning

Machine learning deals with the construction and study of systems that can learn from data, rather than following only explicitly programmed instructions [1]. An algorithm which implements classification is known as a classifier. There are two different models through which a classifier can be trained. They are.

- Generative model
- Discriminative model

A generative model learns the joint probability distribution $p(x, y)$ and a discriminative model learns the conditional probability distribution $p(y|x)$. A generative approach models how the data was generated in order to categorize an instance; it makes a prediction based on the knowledge gained. A discriminative model on the other hand, does not attempt to model the generation process. It simply categorizes a given instance based on learning the differences between two categories.

An example application for a generative and discriminative model is in speech recognition problem, where the system attempts to identify the language someone is speaking [3].

- The generative approach is to learn each language and determine as to which language the speech belongs to.
- The discriminative approach is to learn enough linguistic differences between the two languages to identify which is being spoken.

1.3 Text Classification

Text Classification is an area where classification algorithms are applied to text documents, identifying the target class for a particular document based on the content available in the documents. [2] Classifying a corpus of documents into different categories like Science Fiction, Comedy, Fantasy etc. based on the document vocabulary is one such application. Classifying a movie as good or bad based on the reviews posted by several users on online social networks is another application of text classification.

For a classifier to learn how to classify any text, it needs to attain some knowledge and make prediction based on the knowledge gained. To make this possible we provide two sets of inputs given to a classifier, training and testing data. Training data is labelled manually and is given as

an input to the classifier. The ultimate goal is to learn knowledge from the training data and predict the target classes of the test data based on the knowledge gained. The size and choice of the training and testing data plays a major role in building a good classifier.

1.4 Text Representation

The most common and proven technique for text and image classification is a simple "Bag of words" model. The bag of words model is a simplified representation of text as a bag (multiset) of words, disregarding grammar and word order but keeping multiplicity intact. Figure 1 is a simple example which shows the bag of words representation.

$$Documents = \begin{bmatrix} 'Roger\ plays\ tennis\ better\ than\ Novak' \\ 'Nadal\ plays\ tennis\ better\ than\ Roger' \end{bmatrix}$$

```
{
        "Roger"     :     2
        "plays"     :     2
        "tennis"    :     2
        "better"    :     2
        "than"      :     2
        "Novak"     :     1
        "Nadal"     :     1
}
```

Figure 1 Bag of words representation

A bag of words approach is highly dependent on the term frequencies. It is a proven approach if we have a huge amount of content in the documents. However, the documents we consider in this research work are tweets.

1.5 Online Social Networks

Online social networks have become extremely popular in the recent past. About two thirds of the global population now uses online social networks. Almost 10% of the total amount of time spent online is devoted to social networks and blogging [4]. Social networks like Facebook and Twitter are already replacing email and search engines as the primary interface to the Internet. This trend is likely to continue, as such networks seek to personalize the web experience [5]. The content posted on social networks is most recent, frequent and updated. This is one obvious reason we chose online social network platform for our research.

1.6 Overview of Twitter

Twitter is a social networking application which allows users to share their perspectives or insights on diverse topics. It helps people to collaborate and socialize with other people all over the world. However, the intuitive messages known as *tweets*, posted by users to express their view, are constrained just to 140 characters. This constraint forces the user to limit his/her views to only few words. In a Twitter network *follower* is a person who subscribes to a user's. *Trends* are the latest topics which are ongoing and being discussed by the majority of users. Usually, trending topics starts with # and are generically termed as *hashtags* (e.g., #ebola). Hashtags in general indicate the main theme of discussions. Despite the fact that, Twitter is considered as a social networking application it is comparable to a daily chronicle in which users post their perspectives on a current trending topic. On Twitter, tweets are presented to the user in a sequential order. Figure 2 gives us an overview of the user space of a Twitter user. This format of presentation is useful to the user since the most recent tweets from the user's followers are rich in recent news which is by and large more interesting than tweets about an event that occurred a quite a while prior.

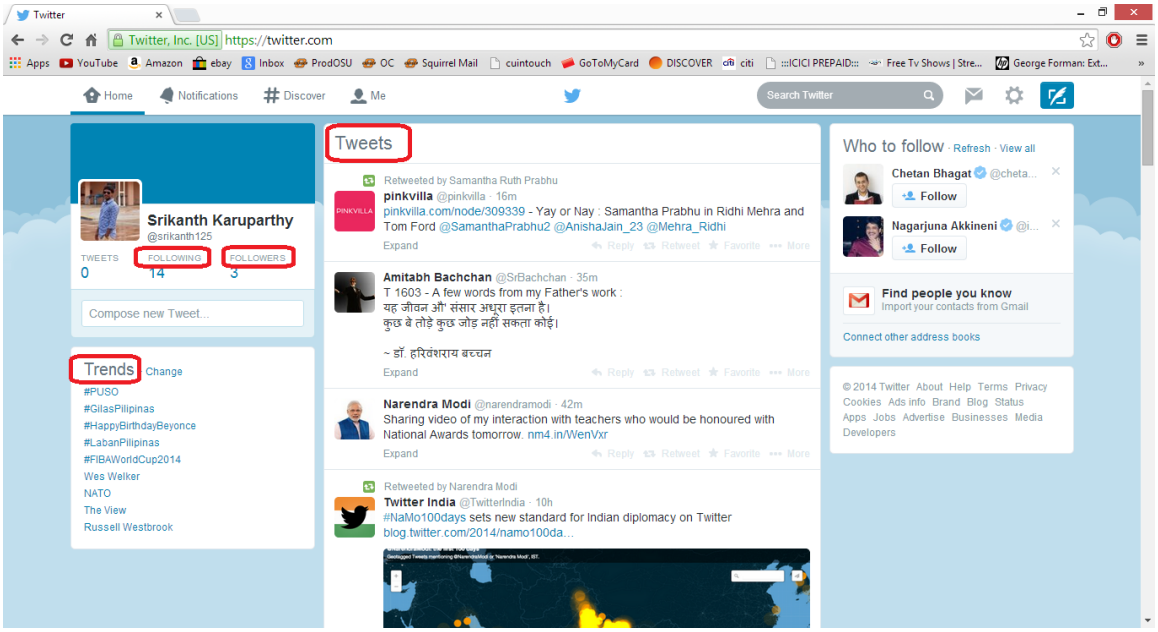


Figure 2 Overview of Twitter

The real downside of this approach is that tweets arrive at a rapid rate. Just displaying tweets in a sequential order may be too overwhelming to the user. Likewise, if the user follows many streams out of whom a couple of people tweet at a rapid rate compared to others, the dominant stream takes a great deal of the user's space. Consequently, tweets from the lesser dominant sources may be lost in the staggering tweet stream. Figure 3 illustrates how the user's get inundated by continuous stream of tweets. The whole user space is occupied by a single user reporting current news. Hence, there is a need to separate the tweets into different categories and present such categories to the user.

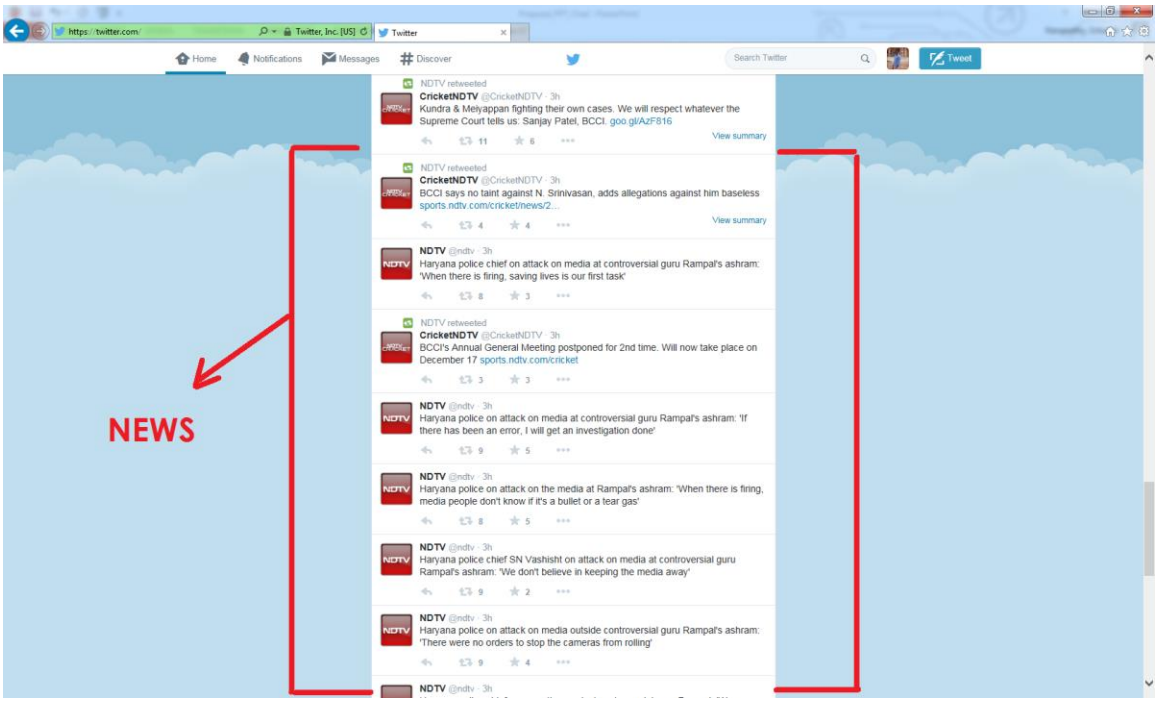


Figure 3 Tweets with same category

CHAPTER 2

REVIEW OF LITERATURE

2.1 Related Work

A number of recent papers have addressed the classification of text posted on social networks. [6] is the first paper to classify users based on their posts' linguistic content. The paper deals with binary classification problems like classifying a user's gender based on the content posted or determining the political orientation using domain specific features. They extracted some features from tweets like SMILEYS (which represent the list of emoticons in a tweet), PUZZLED PUNCT (which represent the combination of characters like '?' and '!') and several other explicitly handpicked features and then classify a user's gender based on their usage of these features in their language. Another heuristic work [7] built upon the earlier work and introduced more categories such as political orientation, ethnicity and whether or not a person is a Starbuck's fan. They built a probabilistic model to extract features based on the user's linguistic style which rely on heuristics such as the use of words like 'dude' or 'lmao' to discriminate between young and old people, usage of expression 'health care' which they use to discriminate between democrats and republicans. [8] is another work in which the authors broadly divide all the tweets into five classes and classify the content based on domain-specific features, such as the presence of word-shortenings and slang, time-event phrases, opinionated words, emphasis words, currency and percentage signs, and the presence of "@username". [9] Defines a set of features associated to Twitter tweets, users, topics and propagation of retweets in order to automatically classify news

events on Twitter as credible or not credible, so as to detect misinformation or false rumors. They show the effectiveness of user and sentiment-based features to this end.

Some of the works like [10] use external knowledge bases like Wikipedia for the classification task. It matches any context posted with the content available in Wikipedia and categorizes it to classes which are defined in Wikipedia. [11] makes use of a similar approach in which they characterize Twitter content based on the co-occurrence of keywords and then compare to similarity measures relying on WordNet synonyms. Yerva et al. [12] classify tweets to identify if they are related to corporate companies or not based on company profiles built using web sources.

A lot of research has also been done on sentiment classification of the Twitter content. Go et al. [13] introduced an approach for automatically classifying the sentiment of tweets with emoticons using distant supervised learning. The authors in [14] also classify the tweets into positive or negative based on POS (Parts of Speech) specific polarity features.

2.2 Existing Approaches

All the existing approaches employ a heuristic approach for classification i.e. they rely on some guess work and great deal of experimental validation for feature selection. They mostly depend on a user's linguistic style for feature selection. Most of the intuitive works handpick features like emoticons, repeated words, presence of punctuation or presence of particular word. However, this is an ad-hoc scheme and is confined only to one particular domain. Selecting features in this fashion does not generalize to additional classification tasks, and requires brainstorming to come up with a feature set.

A few other works rely on external meta-information like Wikipedia, WordNet, web searches etc. The use of meta-information like Wikipedia, WordNet may again create another problem called

the curse of dimensionality [16]. When the feature list becomes exclusively huge, data becomes difficult to visualize and the premise for classification is lost. Hence, there is a need to viably prune features and decrease the feature list size to an ideal value. Additionally, there may be a few enthusiastic and insignificant features that debase the performance of a classifier.

CHAPTER 3

TECHNICAL OVERVIEW

There is a need for an approach which is self-reliant, non-heuristic, non-ad-hoc and does not depend on manual intervention for feature selection. We demonstrate a system that automates the classification of Twitter content, and in particular the feature selection process, while maintaining the consideration of accuracy, speed and minimizing the feature set. For our experiments we arbitrarily created 13 diverse generic classes for tweet categorization. They are *arts, books, business, deals, fashion, food, health, politics, religion, science, technology, sports* and *news*. These classes are just an example; users can create their own categories. Our system can provide suggestions for users based on the current trends globally and locally. To classify this Twitter content into their predefined and user defined classes, we demonstrate a text classification algorithm based on a Bag-of-Words model that automatically selects features based on the TF-IDF (Term Frequency – Inverse Document Frequency) weights for the corresponding bag of words. Figure 4 gives an overview of our model.

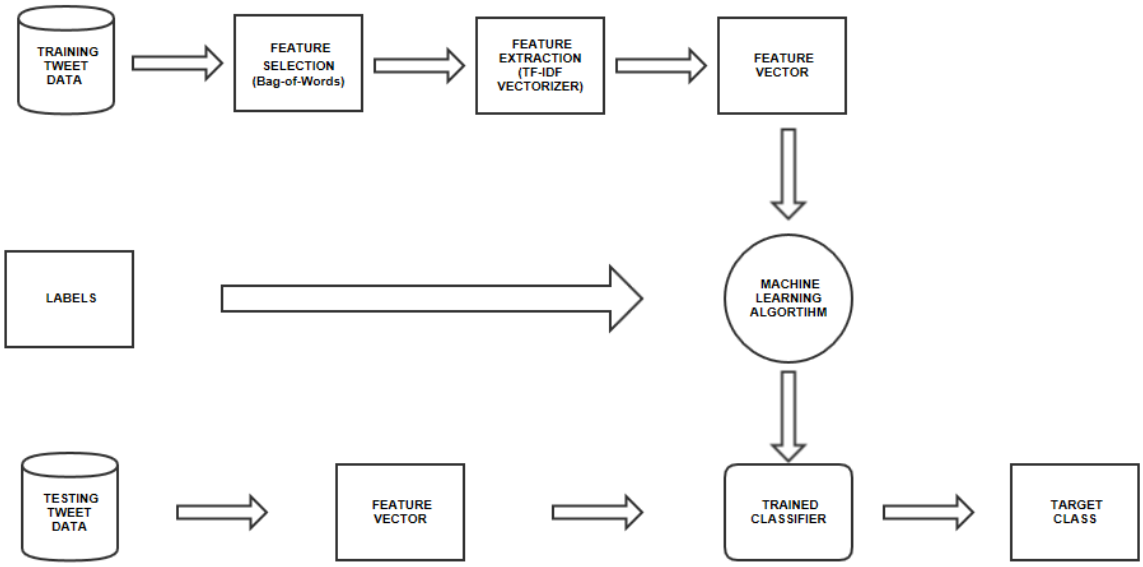


Figure 4 Proposed Model

A training tweet dataset is labelled manually and given as an input to a classifier. The classifier thus learns from the feature vector obtained from the tokenized explicitly labelled tweet data. A set of test data is now given as input to the classifier to identify the target classes for the input tweets. Thus, the classifier tries to predict the target class based on the historical/training data from which it learnt.

CHAPTER 4

METHODOLOGY

4.1 Data Collection

We downloaded all our tweets using a crawler which can pull tweets from the active Twitter stream. We manually labelled all the tweets based on their content into one of the pre-determined classes. We also include a user defined class for any topic of user's choice. For example, if a user is interested in tweets related to "ebola", it is more beneficial to create a separate class called "ebola" rather than have many ebola related tweets in any other categories. By doing so, non-ebola related tweets get more visibility in other categories. Figure 5 (i) and 5 (ii) show example tweets from the training data we use for our experiments. We downloaded 13340 tweets from the Twitter stream for our experiments.

SPORTS	RT @NickJakusz: Highlights from 40-7 @USFSaints football win over LBU...and there were A LOT of highlights!; ; https://t.co/Offq0sEVIp
SPORTS	@kaathleenx @mjr_rjm tbh i havent watched any fsu football yet this season... Damn 2am games
SPORTS	RT @SoDamnTrue: I'm ready for: ; Hoodie weather. ; Football season. ; Cold nights. ; Haunted houses. ; Scary movies. ; Bonfires. ; ðŸŽƒðŸ'»ðŸŒŒðŸ,
SPORTS	@Adri_Hewett It's almost football season! You might want to follow @ProFootballWire & @NFL @ESPN for awesome NFL news tips advice etc
SPORTS	New Story: Fox High School Varsity Soccer beat Jackson High School 1-0 http://t.co/hsZGzcZbKm
SPORTS	Football smores bonfires and beer. last fall of highschool better be great
SPORTS	@DrEsquireZ I know you really like NFL Football too so you should probablly follow @NFL & @ProFootballWire
SPORTS	Anyone who picks Fotheringham to play doesn't deserve to manage my football club @FulhamFC #MagathOut
SPORTS	David Amerson has been #Redskins' best corner http://t.co/M4Zh5aLPoJ Join the legends with your own football card: http://t.co/itF1nUUrxi
SPORTS	#OppoCLT20 Match_(1) ; KKR Win By 7 Wicket!; CSK_vs_KKR; KKR_159/7; Overs_19.0; Doeschate; (52)*; Chawla; (4)*; TarGet 158; #CSKvKKR #CLT20 #Cricket
SPORTS	U.S. Soccer Federation hiring a marketing coordinator - Chicago http://t.co/4lSiNxaG6K
SPORTS	RT @megmuigai: Didier #Drogba is #Chelsea's top scoring player in Champions League football with 34 goals. #CFC #UCL

(ii) Data related to Sports

SCIENCE & TECH	It's a sad day for an iPhone user who doesn't have enough GB to download the new update. This one is actually worth it ðŸ˜-ðŸ˜-
SCIENCE & TECH	RT @ReutersTech: Show's not over yet: #Apple to unveil new iPads operating system in Oct. media report says. http://t.co/gGDRU6ox1u http:â€¦
SCIENCE & TECH	Shots Fired: Samsung Disses The iPhone 6 In New 'Note 4' Advertisement! | New Video http://t.co/ZrQzx7eDy0 #WSHH via @worldstar
SCIENCE & TECH	IG ; JeremyHumilde State of the Art: With Appleâ€™s iPhone 6 Plus Phablets Get Serious: Clunky though the term ... http://t.co/BUMMd0t1Nu
SCIENCE & TECH	RT @ShaunieMcMaster: Is apple taking the piss I have to delete 4 gb of shit na bye
SCIENCE & TECH	iPhone 6 production isn't enough to meet demand: #iPhone6 - Did you pre-order an iPhone 6? If you decided to w... http://t.co/O5yyXlxHpU
SCIENCE & TECH	RT @jamieftgrey: All these people talking about iOS 8 when I don't even have an iPhone http://t.co/zV40c8WoyU
SCIENCE & TECH	@kh79_x unfollowed me. I found out using iPhone app #Tweepr http://t.co/lgdM5aiUZX
SCIENCE & TECH	#HOLYHELL #WTFGOESON RT @tomwarren: iOS 8 requires 5.7GB free on 16GB iPhone or 4.7GB free on 8GB iPhone 5C. Wowzers. http://t.co/kIuOWXNwUS
SCIENCE & TECH	HUGE Daily Deals > http://t.co/FHtgV3bXd8 #6180 Multi-Color iRulu Tablet PC 7' Dual Camera Android 4.2 1.2Ghz B... http://t.co/N6YEJ6kxg5
SCIENCE & TECH	@LiquidGotti I have an iPhone and I don't want to read about the iOS update either. Folks are on the edge of their seats waiting.

(ii) Data related to Science & Tech

Figure 5 Training Data (i) Data related to Sports (ii) Data related to Science & Tech

4.2 Feature Selection

Feature selection or selecting a subset of relevant features for building robust models is our primary research problem. There prevails a heuristic approach of selecting the subset of features manually or explicitly. For our experiments we built the feature set as a Bag-of-Words model taking into account all the text available in the tweets. We did not select any features manually. The set of features in the feature set include the punctuations, stop words, URLs, emoticons etc. which all other works either ignore or eliminate. Our feature set is a list of all the unsanitized data tokenized into a bag of words. This feature set is given as an input to a vectorizer. Figure 6 is a word cloud representing all the data we downloaded for our experiments.

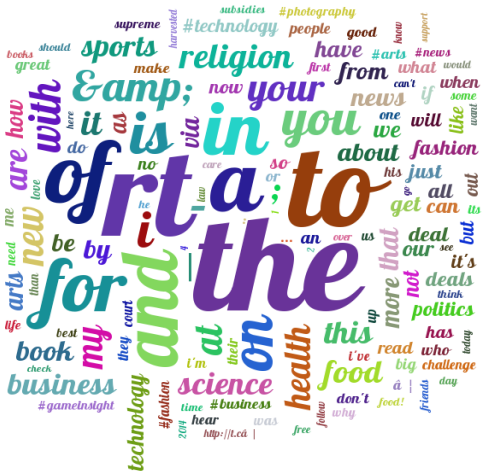


Figure 6 Word Cloud of features

4.3 Feature Extraction

According to [15] feature extraction is a process that extracts a set of new features from the original set of features. In our feature extraction process the Bag-of-Words set is vectorized, i.e. the textual information is converted to a Vector Space Model (VSM). We use a TF-IDF vectorizer for our classification task. The TF-IDF vectorizer is mostly used in information retrieval activity to find the importance of a search term. TF-IDF is composed of two terms TF and IDF. TF stand for term frequency which measures how frequently a term appears in a tweet. IDF stands for inverse document frequency which is used to measure how important a term is. Both are combined i.e. TF.IDF to compute a weight. This weight is a statistical measure used to evaluate how important a word is to a tweet in a collection or corpus. The importance increases proportionally to the number of times a word appears in the tweet but is offset by the frequency of the word in the corpus. We have:

$$IDF(t) = 1 + \log \frac{Total\ number\ of\ tweets}{Number\ of\ tweets\ with\ term\ t\ in\ it}$$

A simple example to explain how TF-IDF works is as follows.

Say we have two tweets in a corpus of tweets/documents.

$$Documents = \begin{bmatrix} 'I\ like\ databases' \\ 'I\ hate\ hate\ databases' \end{bmatrix}$$

The term frequency (TF) for the tweets would be as follows:

$$\begin{bmatrix} ('I', & 1), & ('like', & 1), & ('databases', & 1) \\ ('I', & 1) & ('hate', & 2) & ('databases', & 1) \end{bmatrix}$$

The vocabulary vector is

$$[I, \quad hate, \quad like, \quad databases]$$

The TF vector for first tweet is

$$[1 \quad 0 \quad 1 \quad 1]$$

TF vector for second tweet is

$[1 \quad 2 \quad 0 \quad 1]$

The master tweet term (TF) matrix is

$\left[[1 \quad 0 \quad 1 \quad 1], \quad [1 \quad 2 \quad 0 \quad 1]\right]$

In reality each tweet will be of different size. On a large tweet the frequency of the terms will be much higher than the smaller ones. Hence we need to normalize the tweet based on its size. To normalize a vector is the same as calculating the unit vector and is denoted by $\hat{v}$. The definition of a unit vector $\hat{v}$ is:

$$\hat{v} = \frac{\vec{v}}{\|\vec{v}\|}$$

Where $\hat{v}$ is the unit vector, or the normalized vector and $\|\vec{v}\|$ is the norm (magnitude, length) of the vector $\hat{v}$. $\|\vec{v}\|$ is defined given by the formula:

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + v_3^2 \ldots v_n^2}$$

Our vectors here are:

$\left[[1 \quad 0 \quad 1 \quad 1], \quad [1 \quad 2 \quad 0 \quad 1]\right]$

In order to get its unit vector we plug this vector into definition of unit vector to evaluate it.

$$\hat{v} = \frac{[[1\ 0\ 1\ 1], [1\ 2\ 0\ 1]]}{\sqrt{1^2 + 0^2 + 1^2 + 1^2}\sqrt{1^2 + 2^2 + 0^2 + 1^2}}$$

And we get,

$\left[[0.57735027 \quad 0 \quad 0.57735027 \quad 0.57735027], \quad [0.40824829 \quad 0.81649658 \quad 0 \quad 0.40824829]\right]$

We then compute IDF. When we compute TF all the terms are considered equally important. In fact certain terms that occur too frequently like ‘I’ in our example have little power in determining the relevance. We need a way to weigh down the effects of too frequently occurring terms like ‘The’, ‘I’, symbols, URLs, conjunctions and keywords etc. The major notion being the terms that occur less in the tweet can be more relevant. We need a way to weigh up the effects of

less frequently occurring terms. IDF helps us to solve this problem. By computing the IDF for our vocabulary vector using the formula we get the following vector.

[1.386294 1.098612 1.098612 1.386294]

It is evident from the vector the weight for the term 'I' which does not infer much information and is repeated in both the tweets has been weighed down using IDF. Thus IDF reduces the impact of common terms that occur with multiple categories.

The corresponding TF-IDF weights would be as follows:

[[0.6185789 0 0.48884832 0.61685789], [0.4707717 0.74615549 0 0.4707717]]

Thus, we see TF-IDF weight for the word 'hate' is highest which means that the word "hate" is the most distinguishing feature between these two tweets. By the end of feature extraction we have a set of features represented in a vector format with corresponding TF-IDF weights. We put a threshold for the TF-IDF weights ignoring the words which have TF-IDF weights lower than the threshold. We experimented using different thresholds and found 0.50 to be optimum threshold for best results.

4.4 Tweet Classification

A classifier is trained with the feature vector built using feature extraction. An algorithm that implement classification, especially in a concrete implementation is known as classifier. There are several state-of-the-art classification algorithms. We perform our experiments using three state-of-the-art machine learning algorithms. Namely:

- Multinomial Naïve Bayes Algorithm [17]
- Linear SVC Algorithm [18]
- Logistic Regression (Maxent) [19]

The trained classifier is then used on a test data set to check the classification accuracy. The results of our experiments, which we show in our next section, were encouraging.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Non-Heuristic Classifier:

The data sets used in these experiments are tweets from Twitter. We only considered those tweets that can be labeled into one our pre-determined classes. All our experiments were run using scikit-learn implementation. We vectorize all the features obtained using TF-IDF vectorizer and use these features to build classifiers. Three classification algorithms namely Multinomial Naive Bayes, Support Vector Machine (SVM) and Logistic Regression are used on the training data. *Precision, recall* and *f1-score* are the basic measures used in evaluating classification algorithms. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records from the data available. It is in general expressed as percentage. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved from the data available. It is usually expressed as a percentage. Precision and recall are defined as:

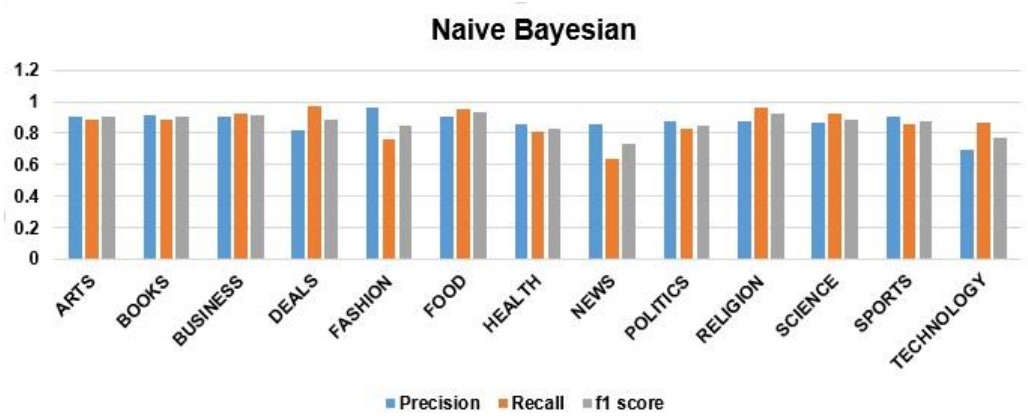
$$Precision = \frac{true\ positives}{true\ positives + false\ postives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

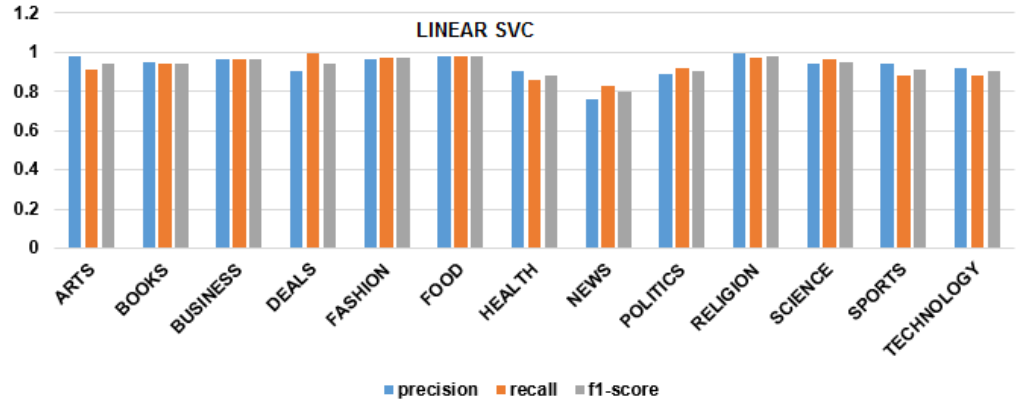
F1-score is the simple harmonic mean of precision and recall. Other related measure used in classification is *Accuracy.* Accuracy is defined as:

$$Accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

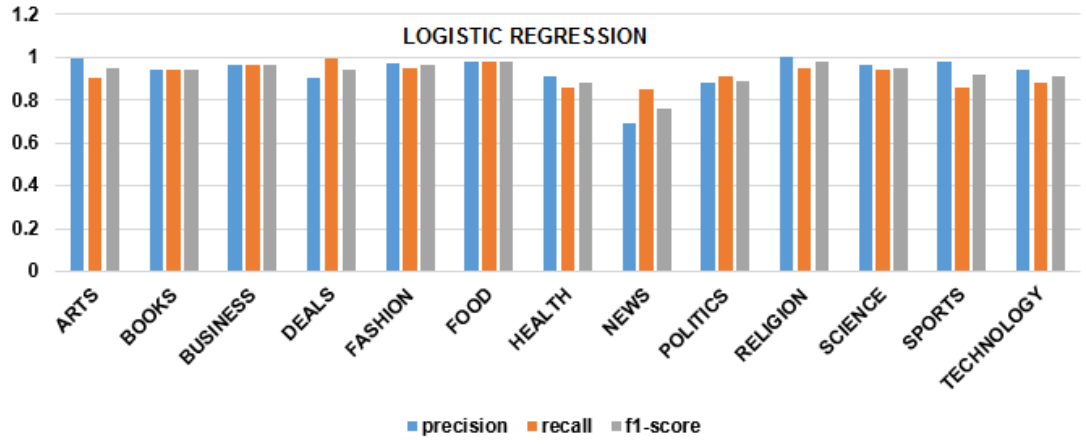
From our experiments we could see very good precision, recall and f1-score by all the three classifiers. The SVM classifier outperforms the other two classifiers performance. Figure 7 shows a detailed classification scores for each classifiers.



(i) Naïve Bayesian Classifier



(ii) Linear SVC Classifier



(iii) Maxent Classifier

Figure 7 Classification Scores for each classifier (i) Naïve Bayesian Classifier (ii) Linear Classifier (ii) Maxent Classifier

Figure 8 shows the accuracy scores for each classifier. The Linear SVC model had an accuracy score of almost 0.9265. The Maxent classifier falls short by a slight margin with an accuracy score 0.9261.

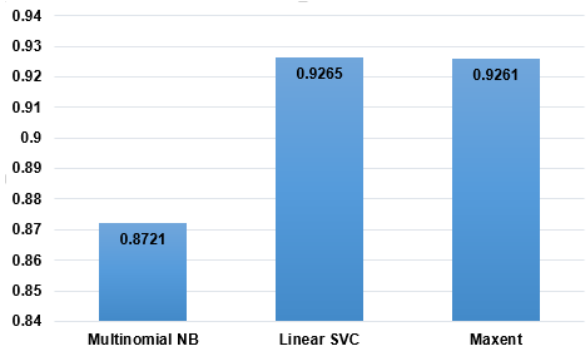
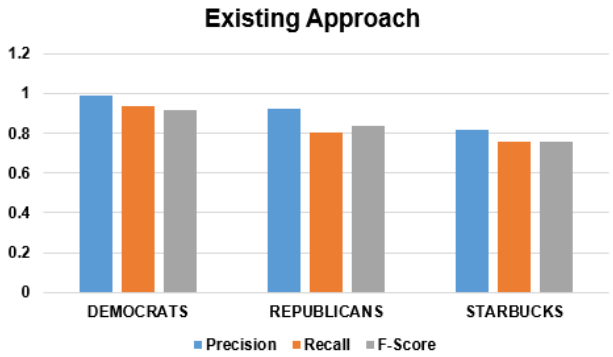


Figure 8 Accuracy Scores for each classifier

5.2 Non-Heuristic Classifier vs Heuristic Classifier:

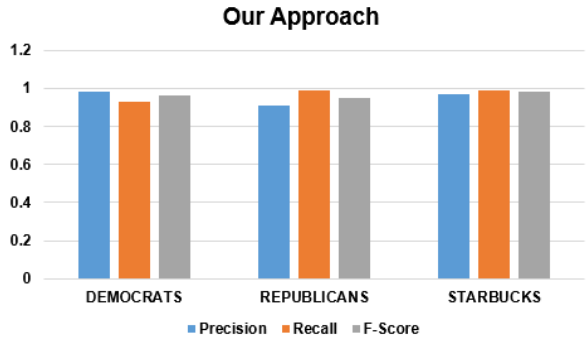
We also compare our results with the existing approaches. Figure 9 shows the classification scores for their experiments and Figure 10 shows the classification score obtained from our

approach. We manually downloaded tweets pertaining to the said categories and ran our experiments using the non-heuristic approach.



	Precision	Recall	F-Score
DEMOCRATS	0.989	0.936	0.915
REPUBLICANS	0.92	0.805	0.84
STARBUCKS	0.817	0.756	0.759

Figure 9 Classification Scores for existing approach [7]



	Precision	Recall	F-Score
DEMOCRATS	0.981	0.931	0.96
REPUBLICANS	0.912	0.984	0.943
STARBUCKS	0.961	0.983	0.974

Figure 10 Classification Scores for our approach

We could perceive from our experiments that the classifications scores for existing and our approach were more or less equivalent. Comparison of the Accuracy score for classification for another existing approach and our approach is shown in Figure 11.

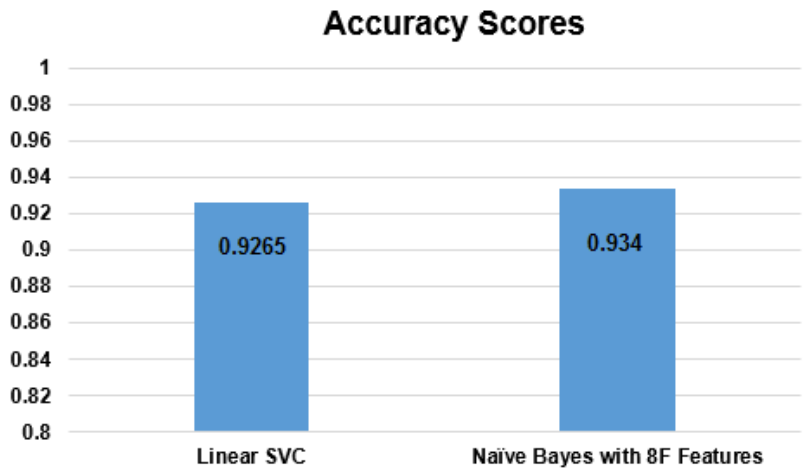


Figure 11 Accuracy Score for our approach (Linear SVC) vs Existing Approach [8]

Thus, our analysis shows that automated feature selection process for Twitter content classification performs on par with current state-of-the-art approaches which incorporate painstaking, time-consuming human effort to manually and heuristically select a feature set.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The work described in this paper is a step towards efficient classification of tweets. Current techniques employ a heuristic and manual approach to select linguistic features for any Twitter classification problems, which requires both manual effort and time. Hence, the traditional approach of handpicking linguistic features is time consuming and requires a lot of guess work.

Some existing works on classification of microblogs integrate text with meta-data from other data sources, for example, Wikipedia and WordNet. Automatic text classification and hidden topic extraction approaches perform well when there is meta-data or when the context of the short text is extended with knowledge extracted utilizing substantial data. Yet these approaches oblige online querying, which is time-consuming and is unfit for real-time applications. When features from external knowledge are used to enhance the feature set, complex algorithms are required to precisely prune the feature set. These approaches eliminate the problem of data sparseness; however they create a new problem of the curse of dimensionality [16]. Hence efficient ways are required to improve the accuracy of classification by using minimal set of features to represent the tweets.

A "Perfect classifier" does not exist. It is always a tradeoff between a few factors that are application dependent. In any case, the fundamental objectives of all classifiers are the same, higher accuracy and better speed. In this research, we have tried to achieve both of these objectives; however there is a scope for great deal of improvement. We intend to utilize our

methodology with a multi-label classifier effectively. Further analysis is required to effectively integrate a multi-label classifier with our system online. Despite the fact that our research concentrates only on Twitter data, there is a need to adapt this approach to function admirably on all other Social Network data. We would like to concoct a non-specific system that can perform reliably well on diverse type of short microblogs. There is a lot of scope to process tweets to capture better information; crawling the tiny URLs is one such approach. We currently do not crawl URLs. We envision building an online classifier which will classify Social Networks content robustly with high speed and precision and with minimal set of features.

REFERENCES

[1] “Machine Learning” http://en.wikipedia.org/wiki/Machine_learning

[2] William Cohen, Text Classification, Tutorial, CMU, CALD Summer Course.

[3] Machine Learning: Generative and Discriminative Models
http://www.cedar.buffalo.edu/~srihari/CSE574/Discriminative-Generative.pdf

[4] Nielsen report 10% time spent on Online Social Networks
http://www.nielsen.com/content/dam/corporate/us/en/newswire/uploads/2009/03/nielsen_globalfaces_mar09.pdf

[5] Jing Jiang, Christo Wilson, Xiao Wang, Wenpeng Sha, Peng Huang, Yafei Dai, and Ben Y. Zhao. 2013. Understanding latent interactions in online social networks. *ACM Trans. Web* 7, 4, Article 18 (November 2013), 39 pages.

[6] Rao, Delip, et al. "Classifying latent user attributes in Twitter." *Proceedings of the 2nd international workshop on Search and mining user-generated contents.* ACM, 2010.

[7] Pennacchiotti, Marco, and Ana-Maria Popescu. "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11 (2011): 281-288.

[8] Sriram, Bharath, et al. "Short text classification in Twitter to improve information filtering." *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010.

[9] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. "Information credibility on Twitter." *Proceedings of the 20th international conference on World wide web*. ACM, 2011.

[10] Genc, Yegin, Yasuaki Sakamoto, and Jeffrey V. Nickerson. "Discovering context: classifying tweets through a semantic transform based on wikipedia." *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*. Springer Berlin Heidelberg, 2011. 484-492.

[11] García-Silva, Andrés, et al. "Characterising emergent semantics in Twitter lists." *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 2012. 530-544.

[12] Yerva, Surender Reddy, Zoltán Miklós, and Karl Aberer. "What have fruits to do with technology?: the case of orange, blackberry and apple." *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*. ACM, 2011.

[13] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* (2009): 1-12.

[14] Agarwal, Apoorv, et al. "Sentiment analysis of Twitter data." *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, 2011.

[15] Motoda, Hiroshi, and Huan Liu. "Feature selection, extraction and construction." *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol 5* (2002): 67-72.

[16] "Curse of Dimesionality" http://en.wikipedia.org/wiki/Curse_of_dimensionality

[17] "Naïve Bayes Classifier" http://en.wikipedia.org/wiki/Naive_Bayes_classifier

[18] "Support Vector Machines" http://en.wikipedia.org/wiki/Support_vector_machine

[19] "Logistic regression" http://en.wikipedia.org/wiki/Multinomial_logistic_regression

VITA

SRI KANTH KARUPARTHY

Candidate for the Degree of

MASTER OF SCIENCE

Thesis: A NON-HEURISTIC MACHINE LEARNING APPROACH FOR CLASSIFYING TWITTER CONTENT

Major Field: COMPUTER SCIENCE

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in May, 2015.