Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 3, March -2017

Survey of Classification Techniques for chat message classification

Himanshu Kulkarni[1], Siddharth Bhamare[2], Prasad Khanapure[3], Akshay Agrawal[4], Mrs.V.L.Kolhe[5]

[1,2,3,4] *Department of Computer Engineering, D.Y.Patil College of Engineering, SPPU, Pune.*
[5] *Asst. Professor, Department of Computer Engineering, D.Y.Patil College of Engineering, Pune.*

Abstract — *The Internet has fundamentally changed the way of communicating. Instant messaging is kind of online chat that offers real time transmission over the Internet. Short messages are typically transmitted bi-directionally between users, when each user chooses to complete a message and select "send". With increase in the use of smartphones, IM has changed the trend from traditional communication to smart communication. The large amount of data shared between users leads to the need of analysis and classification of the data. Analysis of IM transcripts can play important role in surveying and knowing the trends, social interests and needs of users. Classification helps in connecting and grouping people from similar social interests. Text classification scenario for such application will be- Multiple class classification like selecting one category among several alternatives. Classification techniques like Naive Bayes,Support Vector Machine, and Decision Tree can be used for text classification.*

Keywords-*IM; Topic Detection; Classifiers; SVM; Natural language processing*

I. INTRODUCTION

The Internet has fundamentally changed the way of communicating. E-mail has been the most rapidly adopted form of communication ever known. People around the globe send out billions of e-mail messages every day. But sometimes even e-mail isn't fast enough. You might not know if a person you want to e-mail is online at that moment. Also, if you're emailing back and forth with someone, you will have to go through a few steps. This is why instant messaging (IM) has become so popular. Instant Messaging is the real-time communication involving exchange of text messages through a software application. Instant messages are basically a chat room for just two people. Users can send and receive messages from others who are online. Instant messaging works on XMPP (Extensible Messaging and Presence Protocol). Jabber is a compelling IM solution that is well-suited to meet todays and tomorrows IM needs. Jabber is freely available set of protocols for building IM systems [6]. It uses XML as its standard data format. In Jabber, the client/server model is heavily weighted to favor the creation of simple clients. Most of the text processing and IM logic is carried out on the server. Further different functionalities can be added to the application on server side like data analysis and processing, security, etc.

Usually there is no restriction to anyone for chatting in groups nor stopping of unwanted texts and notification. Best way is to classify the messages into different classes based on user interests, so that unwanted alerts can be turned off specifically. In this paper different classification techniques are studied for achieving task of topic detection and classification in instant messaging. Text messages will be treated as documents and will be classified in one of the predefined categories like movies, sports, studies, etc. User access will be maintained dynamically. Different classification techniques are compared and reviewed. Contents of the paper are organized as follows: Section 2 presents the related work. Section 3 is comparison between different techniques for feature selection and extraction.

II.RELATED WORK

A.Feature Selection

1) Document Frequency [5]: Document frequency is defined as the frequency of documents in which terms occur (how often the term to be selected as feature occurs in the document). Threshold frequency is predefined according to the size of dataset and feature set required. If DF for term is less than the Threshold then term is rejected from feature set. DF method is scalable up to large datasets with linear computational complexity.
2) Information Gain (IG) [5]: IG is the measure of how much information it contributes in the presence or absence of a term to make the classification decision on any class. It determines the importance of the term for specific class.
3) Best Term [5]: In best term method target class is given and BT will find the documents belonging to that class and features which predict it. In second step, BT will check for absence of the documents in the class and still having those features. Based on first and second step feature set is selected.

1. Algorithm: Class Specific Feature selection [3]
INPUT: Documents for a given training data set with N topics.
PROCEDURE:

1. Form a reference class c0 which consists of all documents;
2. For class I = 1:N do
3. Calculate the score of each feature based on a specific criteria, and rank the feature with the score in a descending order;
4. Choose the first K features zi, the index of which is denoted by Ii;
5. Estimate the parameters * under the reference class c0 and the parameters i under the class ci;
6. End

OUTPUT: Given a document to be classified, Output the class label c*.

B. Classification

1. Naive Bayes [3]: Naive Bayes is conditional probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. Naive Bayesian includes four steps: document preprocessing, feature selection, feature extraction and text classification. For document preprocessing stop word removal, stemming, pruning of word technique, etc. can be used. Feature selection can be defined as a process of selecting the subset from the original feature set on the basis of importance of features. There are three categories of feature selection methods: wrappers, filters and embedded methods. For text classification various classifiers such as decision Tree (DT), K Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB), etc. can be used.
   More recently, researchers apply Bayesian partitioning techniques to estimate the distribution in high dimensional data space. Optional Polya Tree (OPT) to construct a prior distribution, and in derived a closed form of posterior probability using Bayesian sequential partitioning. Class-specific features offer many advantages for multi-class classification. For example, class-specific features carry much more discriminative information from the original raw data, because each class can select the most discriminative features against the other classes. This characteristic makes the PPT different from many other classifications methods which usually need to incorporate a one-vs-all classification scheme to build hierarchical multi classifiers to use class-specific features.
   By itself, the naive Bayes is a relatively accurate classifier if trained using a large data set. However, as in many other linear classifiers, capacity control and generalization remains an issue. In our case, however, the naive Bayes is used as a preprocessor in the front end of the SVM to vectorize text documents before the classification is carried out (by the SVM). This is done to improve the generalization of the overall system while still maintaining a comparatively feasible training and categorization time afforded through the dimensionality reduction imposed by the Bayes formula.
2. Support Vector Machines [2, 11]: Support vector Machine is supervised classification method. It separates two classes in the feature space with the widest possible margin (away from the separating hyperplane in linear version). In multi-class problems, SVM can be used after reformulating the problem as a combination of multiple binary problems. In this case, the correct class is determined based on the output of the binary classifiers, known as MSVM. For example, in one-versus-all winner takes-all approach, one binary SVM classier is constructed for each class to determine whether the message belongs to the class or its complement. The class whose classier produces the highest output function value is chosen as the correct class. SVM algorithm consists of two types of versions: non-linear and linear versions. In nonlinear version, classes are not separated i.e. no straight lines can be found that separate the classes. In linear version, Classes are separated using the hyperplanes.
   Linear classifier is defined as

$$W \quad T \quad x \quad + \quad B \quad = \quad O \qquad (1)$$

   Where, W is the direction of the hyperplane and B is the exact position of hyperplane. A machine learning is that expert labeling of large well-known problem in real-world applications of amounts of data for training a classifier is prohibitively expensive. An increasing number of statistical and computational approaches have been developed for document classification, including k-nearest neighbor (k-NN) classification, naive Bayes classification support vector machines (SVMs), maximum entropy, decision tree induction, rule induction, and artificial neural networks. SVMs can be used as a discriminative document classifier and has been shown to be more accurate than most other techniques for classification tasks. The main problem associated with using the SVM for document classification is the effort needed to transform text data to numerical data. We call this essential step, the vectorization step. For the SVM on the other hand, over fitting does not occur since its capacity is equal to the margin of separation between support vectors (SVs) and the optimal hyperplane instead of the dimensionality of the data. What is of concern, however, is the high training time and classification time associated with high dimension data. Using Bayes Naive and SVM, we hope to reduce training and

classification time to a feasible level (reduce dimensions from thousands to 10 or 20) while maintaining acceptable generalization and accuracy.

3. Multinomial Naive Bayes [1]: In multinomial naive Bayes (MNB) classification, it is assumed that the elements of the feature vector have been created by sampling from a multinomial distribution. In MNB classification, the objective is to choose a class that maximizes the posterior probability of the class when the message is given. Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. Whereas simple naive Bayes would model a document as the presence and absence of particular words, multinomial naive bayes explicitly models the word counts and adjusts the underlying calculations to deal with it.
   It describes 3 Naive Bayes models (Multinomial, Binarized and Bernoulli) in the context of Text Classification. Note that Naive Bayes makes the assumption of conditional independence of the features, something that despite it is hardly ever true, it works pretty well. Multinomial Naive Bayes simply assumes multinomial distribution for all the pairs, which seem to be a reasonable assumption in some cases, i.e. for word counts in documents.
   Naive bayes method assumes the independence of feature i.e. one feature should independent from another feature under known priori probability and class conditional probability. Training and testing phase in Naive Bayes is easy for implementation and computation. Multinomial Naive Bayes simply lets us know that each p(fi|c) p(fi|c) is a multinomial distribution, rather than some other distribution. This works well for data which can easily be turned into counts, such as word counts in text. In summary, Naive Bayes classifier is a general term which refers to conditional independence of each of the features in the model, while Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features.

III. COMPARATIVE STUDY OF CLASSIFICATION TECHNIQUES

For chat classification, indicative term based approach is superior to traditional document frequency based approach for feature selection. Topic detection operation can be broadly classified into supervised and unsupervised. Supervised approaches require domain experts for training text documents on predefined conceptual topics, and prediction on topic labels can be then be made on unknown data objects. Unsupervised approaches, on the other hand, clustered text documents into different groups according to the similarity of its contents without involving domain experts for the purpose of retrieving text documents of the same or similar topics.

Table 1: Comparison of classifiers [5]

Sr. No.	Method	Advantage	Disadvantage
1	Decision Tree	Easily pick the best feature from the set of data.	Over-fitting is the problem and so accurate.
2	K-Nearest Neighbor	It is robust to noisy training data.	Classification time is too long and it is difficult to find the optimal value of k
3	Naive Bayes	Easy to implement and it requires less amount of training data.	The assumption of independence of the class results in loss of accuracy.
4	Support Vector Machine	Robust and very accurate.	High algorithmic complexity and extensive memory requirements for large task.

IV.CONCLUSIONS

In this paper different techniques for classification of text messages are reviewed. Supervised approach is used which helps in classifying the messages based on the classes. The Xmpp/Jabber server will be used to provide the connectivity to client IM applications. Classification algorithm execution is on the server side to reduce the number of computations on the client side. Hybrid model involving Naive Bayes as vectorizer and SVM as classifier tends to increase the accuracy by significant amount.

V. REFERENCES

[1] Samir Puuska; Matti J. Kortelainen; ViljamiVenekoski Instant Message Classification In Finnish Cyber Security Themed Free Form Discussion, 2016 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA)
[2] Han Zhang; Chang-Dong Wang; Jian-Huang Lai Topic Detection In Instant Messages, Machine Learning and Applications (ICMLA), 2014 13th International Conference on 2014
[3] Bo Tang; Haibo He; Paul M. Baggenstoss; Steven Kay A Bayesian Classification Approach Using Class-Specific Features for Text Categorization, IEEE Transactions on Knowledge and Data Engineering 2016
[4] Murat Can Ganiz; Cibin George; William M. Pottenger Higher Order Nave Bayes: A Novel Non-IID Approach to Text Classification, IEEE Transactions on Knowledge and Data Engineering 2011
[5] P.Shah; Vibha Patel. A Review on Feature Selection and Feature Extraction for Text Classification.IEEE WiSPNET 2016 conference
[6] Dino isa; Lam hong Lee. Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2008

