Spam Filtering of Bi-Lingual Tweets Using Machine Learning

Hammad Afzal*, Kashif Mehmood*

*National University of Sciences and Technology, Islamabad, Pakistan

hammad.afzal@mcs.edu.pk, kashifmehmood.mscs20@students.mcs.edu.pk

***Abstract*—During recent years, usage of social media has increased enormously. Billions of users use Twitter, Youtube etc which has resulted in the increase in spams as well. Spammers use spam accounts and target users on online social media. Whether a user accesses this social media through smart-phone or web, he/she is prone to the spammers on social media websites. This paper analyses different classification techniques that are currently being used in spam filtering in the context of social media. The contents of tweets are unique in nature, and are different from emails due to their less content so some techniques used in emails might be effective while some might not be effective. Moreover, the conversations on social media often comprises of short-forms/slangs and incorrect spellings. Usage of social media has also become popular in local/regional languages. One such language is Urdu which is common in subcontinent Indo-Pak and is written using English alphabets. We have performed spam classification for Roman Urdu tweets, collected from five major cities of Pakistan. Some of the most commonly used algorithms and techniques for spam classification are discussed and evaluated on English and Roman Urdu tweets from Pakistan in this paper.**

***Keywords:* tweets, Roman Urdu, spam filtering, machine learning techniques**

I. Introduction

During last few years, the traffic on internet has increased enormously. A large portion of this traffic is due to social media such as blogs/twitter. In particular, the microblogs such as Twitter has given platforms to the people throughout the world to express their opinions and comment on others'. Considering the wide usage of these platforms, spammers have also changed their means and have started targeting this media for spamming. According to the Federal Trade Commission, spam is defined as any communication with the user that he does not want or needs [1]. Keeping this definition in mind, one can say that all the unwanted tweets, SMS or emails that a user is seeing and he does not want to see them are spam.

According to pear analytics [2] which carried out a study on Twitter and collected 2000 tweets from US over a period of two weeks in August 2009, 40% of tweets are pointless babble while 3.75% of tweets are actually spam, 5.86% of tweets are for self-promotion like ads that are sent in tweets hidden by someURL shortening services like bit.ly etc, 3.99% are news while 8.70% are pass along values which we also call as retweets and are denoted by 'RT' before every tweet.

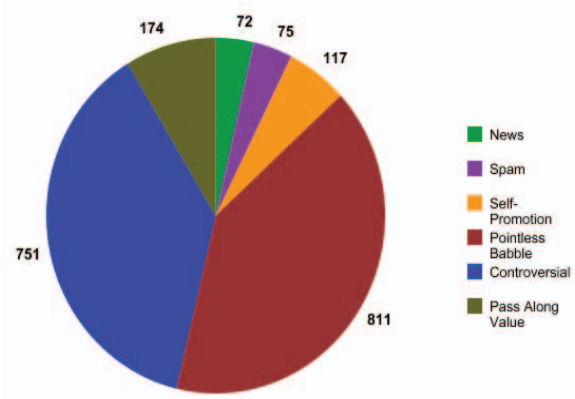

Figure 1. Statistics of Tweet's contents

According to the Twitter website and their statistics, as of March 2015, there are 302 million active users on twitter while 80% of them are active from smart-phones and there are 500M tweets sent per day. 77% of twitter accounts are from outside of United States[1]. There are millions of users that use languages other than English on twitter. Most of the techniques for spam detection are developed/applied for English and other popular languages. The less developed languages such as Urdu, despite being spoken by millions of users around the world (with concentration in subcontinent region) are not much advanced in terms of internet usage and related tools. A large portion of population in subcontinent region uses Roman Urdu (Urdu written using English alphabets) as medium of sharing on Twitter. This population is prone to the spammers who normally send spam messages using local languages and therefore, spam detection tools for English would not be as effective. Therefore, it is of immense importance that tools should also be developed for such languages as well. A number of studies have been carried out for Roman Urdu such as sentiment analysis [3] [4] [5] and on spam detection in Roman Urdu SMS [6]. In this paper, we analysed different machine learning algorithms to evaluate their performance on spam detection in Roman Urdu texts. The paper is structured as section 2 describes literature review,

[1]https://about.twitter.com/company

section 3 defines methodology, and section 4 is about results and discussion while section 6 is conclusion and future work.

II. LITERATURE REVIEW

In this section, a brief review of work related to spam detection in social media

F. Benevenutoet al. [7] collected 54,981,152 (54.9 Million) user profiles that were connected to each other by 1,963,263,821 (1.9 Billion) social links. They collected tweets posted by these users which amount to 1,755,925,520 (1.7 Billion). The authors collected tweets mainly about three trending topics on twitter namely Michael Jackson's Death, Susan Boyel's emergence and the hashtag #musicmonday. They hired two volunteers to label their dataset of users based on the numbers of hashtags and their tweets. Each volunteer's labelling was given a score and then in-case of a tie, a third volunteer was heard. In total, they got 8207 users labelled with 355 spammers and 7852 non-spammers. The results showed that they classified 70% of spammers and 96% of non-spammers correctly by using Support Vector Machine.

Meda, C. et al. [8] used the same dataset as in [7] but they reduced their number of spammers and non-spammers to 355 and 710 respectively. They applied Random Forest RT (a machine learning algorithm). They performed classification by using a standard 10-fold cross validation. 106 users were used a test set while 959 as training set. By applying RT technique they correctly classified 75% of spammers, which was 70% in-case of [7]. They classified 96% of non-spammers correctly.

Stafford, G. and Yu, L.L. [9] gathered English language tweets from Twitter's public APIfor the world's top ten largest trending topics. They ran their program from 1st to 7th Feb 2013 and collected over 9 million tweets from 801 distinct trending topics on twitter. They labelled their (subset of) dataset of 1500 tweets including 1453 non-spam and 42 spam tweets. Using Naïve Bayesian classifier, they classified the tweets and found that 144 messages were incorrectly classified while rest were correctly classified.

Kandasamy, K. and Koroth, P. [10] collected tweets from twitter and performed classification of URL's in the tweets. They extracted URL's from tweets and matched it with the dataset of URL's on a website that contains all the blacklistedURL's[2]. Apart from that they also used some keywords as spam and labelled a tweet as spam or ham based on those words like xxx, viagra etc. Their Training set consists of 100 tweets and the testing set consists of the same set but without labels. Out of the 100 users 98 were classified correctly and only two of them were classified incorrectly. An accuracy of 98% with naïve Bayes and 87.5% with SVM was attained by them.

Soman S.J and D. S. [11] collected tweets from twitter streaming API for 10 days. After labelling their dataset they classified the spam and non-spam accounts and compared results of SVM and TTBPTF (Tweet Trending Bayesian Probabilistic Tensor Factorization) [11]. The authors used 4 different feature of a tweet including *profile of user*, *location*, *activity pattern like when he tweets* and *text of tweet*. They managed to get an accuracy of 89.3% with SVM while TTBPTF gave them an accuracy of 93.8% on text and content of tweet. On location feature they got an accuracy of 85.4% for SVM and 92.5% for TTBPTF. As the TTBPTF is based on Bayesian probability, it gave best results as compared to SVM.

Apart from spam classification, other types of classification have also been performed by researchers on unwanted social media conversations. For instance, in [12] Chu Zi et al. classified a dataset of 50,000 tweets into three classes as Human, Bot or cyborg. A cyborg is either a human-assisted bot or bot-assisted human. The authors performed classification based on the *timing of tweets*, *the account reputation*, *number of retweets*, *number of mentions* and *number of hashtags* used. They found that a twitter account which is run by human tends to have more followers than followings or friends. For example, famous personalities like Tom Cruise have 5,299,321 followers and 57,595 friends and Russell Crowe have 1,771,952 followers and 98 friends. A bot account tends to follow more people so that it can reach more and more people. A bot account can be used for spamming purpose, and to make sure that the spam reaches more and more people, the bot account need to have many followings or friends. They showed some other interesting features used for this classification like the account reputation that is being calculated as in equation 1.

$$reputation = \frac{no\_of\_followers}{no\_of\_followers + friends}$$

Equation 1: Account Reputation

If an account has more followers, it would have more reputation. A human account has more reputation than a bot or cyborg account because a bot and cyborg has more friends than a human account. From equation 1 we can calculate the reputation of Tom Cruise's account which amounts to 0.98, which is closer to 1.

Content based approaches are assumption based and they read the whole text of a message in order to classify the message as Spam or ham (non-Spam). It analyses the text using features (e.g. tokens) and then decide whether this message is legitimate or Spam. There are many content based approaches available in the literature. Content based filtering is divided into statistical-based and rule based approaches. A rule based technique depends on the analysis of the domain information in terms of a set of rules. It is usually high cost because it keeps a set of rules by obtaining and maintaining it. In [13] Zhenyu found two limitations of this method. Firstly when we try to reduce the spam and insert ham like information into the message, they are then vulnerable to attacks; and second is that there are limited number of training examples.

[2] www.blacklist.com Blacklisted URL's Database

III. METHODOLOGY

In order to perform experiments on Roman Urdu tweet spam detection, we devised an experimental setup that comprises the following steps.

1. Collection of tweets.
2. Pre-processing of tweets
3. Data Preparation for classification - Labelling of tweets by domain experts as spam/ham
4. Tweet Classification as spam/ham
5. Performance analysis of different classifiers

A. Collection of Tweets

In first step of tweet collection, 2000 tweets are collected using Twitter API from the five major cities of Pakistan including Lahore, Karachi, Islamabad, Quetta and Peshawar. A few sample tweets are given in the Table 1.

TABLE 1: SNAPSHOT OF TWEETS COLLECTED FROM FIVE CITIES

	Language	Text of Tweet
Lahore	Roman Urdu	punjab hakumat ka mukhtalif madaris me zer-e-talim 500 ghair mulki tulba wa talbaat ko de-port krne ka faisla.
Karachi	English	#Karachi Police Chief says killer of KU Prof Shakil Auj, and Dr Sibt-e-Jafar arrested, resident of Liaqatabad affiliated with political party
Islamabad	Roman Urdu	@108kamal @rahehaq ye waqt bht zalim hta h jb us ka pahiya chlta h tu sb kuch barsbr kr dyta h mazlm ki bddua tu arsh hila dytihai
Quetta	English	149,548 people could have seen #PakNeedsJI since its 1st mention until it became a Trending Topic. #trndnl2015
Peshawar	English	Senior citizens r facing problems to get their sims verified by Biometric as their figer prints hav ben vanished.

B. Pre-processing of Tweets

Tweets are processed to remove non alphanumeric characters which do not bear much information in context of spam detection. Tweets of size less than 6 are also removed as they also tend to not carry any information to the classifier. After cleaning of tweets we are left with 1463 tweets.

C. Data Preparation for Classification

Tweets are then labelled as spam or ham (legitimate) by the domain experts. Out of 1463 tweets, 425 (29%) are marked as spam while rest 1038 (71%) are ham. Each data record consists of *user id, hashtags, numbers, URL's* and *labels (spam/ham)*. User id is also text which identifies who created this tweet or retweeted it.

In order to perform a wider variety of classification algorithms, we converted text instances into numeric values (using 'StringToWordVector' function of WEKA which also generates word vectors for classification). Features used in this paper are word grams and bagging. An example of the outcome of this function is shown in Table 2.

TABLE 2 : REPRESENTATION OF FEATURES AS VECTORS

Text	Features	Vector representation
@10pakistan19 raat ko 4 bajya tak online rahti ho	@10pakistan19 Raat Ko 4 Bajya Tak Online Rahti ho	{0 1, 1 1, 3 1, 5 1, 8 1, 9 1, 10 1, 11 1, 13 1}
@asif12252 kahan ja rhy hain	@asif12252 Kahan ja Rahy hain	{2 1, 4 1, 6 1, 7 1, 12 1}

Each word is translated into a numeric representation. First number in each dimension of vector is numeric representation of a number while second is its frequency in the current document. In example shown in Table 2, all words are appearing once in one tweet.

D. Classification

Following algorithms are applied for classification: *Naive Bayes Multinomial, Liblinear, LibSVM, DMNBText* and *J48*. In order to measure performance of different algorithms, accuracy and ROC AUC are used. Accuracy is the number of correctly classified instances as compared to the incorrectly specified.ROC is Receiver Operator Characteristics and AUC is Area Under Curve. ROC curves are used to measure the performance and the more AUC, the good classification performance of the algorithm.

Naïve Bayes Multinomial is a variant of Naïve Bayes with multinomial distribution. Distributions can be of any type like binomial trinomial and so on to multinomial distributions. The multinomial distribution gives the probability of any combinations of number of successes for various categories. While binomial is the probability distributions is the number of successes for one of just two categories.

SVM, proposed in [14], works on vectors which we sometimes call as features. Let's suppose there are two classes in which classification is to be performed, SVM creates a vector and decision boundary in an x-y plane which contains vectors and class labels on the basis of which a certain document is classified into each one of the classes.

DMNB is Discriminative Multinomial Naïve Bayes. It is also a variant of NB and works on the same Bayes rule. It was suggested and experimented by [15].

Liblinear is a variant of SVM with different kernel implementation. It a library for linear classification of large documents. As opposed to its origin SVM, Liblinear is quite

fast and works well on large documents which have been shown in results.

A. Results and Discussion

The results of classification are presented in this section. Some of the algorithms took a long time to build model on the training data and then also on classification. Results have been produced using 10-fold cross validation.

Table 3: Performance comparison of algorithms

	Time taken (seconds)	ROC AUC	Accuracy (%)
NB	0.02	0.973	95.42
DMNBText	0.02	0.984	95.12
LibSVM	0.85	0.5	70.88
J48	11.33	0.929	91.38
Liblinear	0.08	0.9363	94.60

Highest accuracy was reported by Naïve Bayes Multinomial which is 95.42%. DMNBText also showed an accuracy of 95.12%, however, DMNBText returned more false positives than the first one.

LibSVM performed very poorly on this dataset and gave lowest accuracy of 70.88%. The reason behind this is that SVM depends very highly on the attributes being used or generated. SVM showed an improvement in accuracy when feature selection with Information Gain was performed on the dataset. Liblinear showed an accuracy of 94.60% and it took 0.08 seconds.

J48 showed an accuracy of 91.38% while taking a time of 11.33 seconds. This algorithm showed quite good accuracy but it also took much time to build model which might not be feasible in the current smart phone's environment. Results are summarized in Table 3.

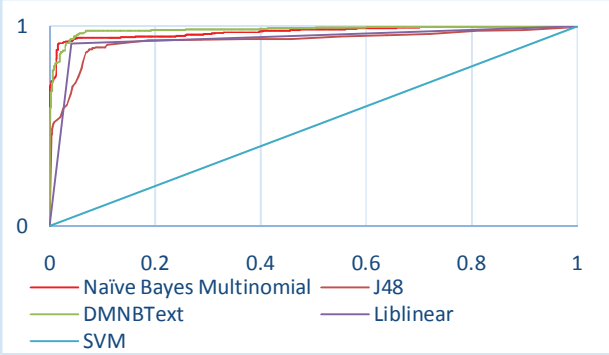

Figure 2: ROC Curves of algorithms

The highest AUC of 0.984 is shown by DMNBText followed By Naïve Bayes Multinomial which showed an area of 0.9736. Liblinear had an area of 0.9363. SVM showed linear curve with an area of 0.5. Linear curve in case of SVM denotes that this algorithm did not perform well and classified the instances randomly not systematically. The reason for this is that there are a large number of attributes in this dataset closer to 2000 while SVM's performance is greatly affected by the number and the quality of attributes that are in the dataset. On the other hand the tree based J48 gave an AUC of 0.929 which is also very good. ROC Curves of algorithms have been shown in the Figure 2.

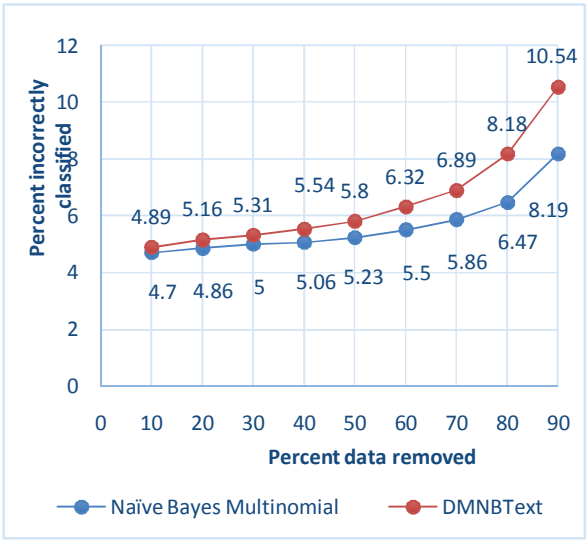

Figure 3: Percent data removed vs percent incorrectly classified instances

In the next experiment we removed percentage of data ranging from 10% - 90% from the dataset and see how the algorithm performs on less data and compare the results in terms of percent incorrectly classified instances and false positives. It can be seen from Figure 3 that when the data reduces the error rate or incorrectly classified instances of DMNBText increases as compared to Naïve Bayes Multinomial. When 10% of data is removed error rate of first one is 4.89 while that of latter is 4.7. When 90% of data is removed and only 10% of data is fed into the algorithm, error rate of first one is very high, 10.54 while that of latter one is 8.19. So we can conclude that Naïve Bayes Multinomial can work better even when less data is present for classification. So from these results it is obvious that Naïve Bayes Multinomial is the best algorithm for classification on SMS and ultimately on smart-phones.

IV. Conclusion and Future Work

As there are number of algorithms and techniques available for spam filtering, there are somewhat less available for tweets spam filtering. Number of results presented for less developed languages such as Urdu is even lesser. Not all the algorithms can yield best results on smaller text such as tweets. According to the literature review that has been performed in this paper, Naïve Bayes Multinomial is the simplest and easy to implement technique while SVM is slower on larger datasets and also does not performs optimally in case of large number of attributes. Both the DMNBText and Naïve Bayes Multinomial performed very well but the later had an upper hand in case of the false positives which were less as compared to the first algorithm. To distinguish further between these two algorithms we conducted another experiment that has further revealed the strengths and

weaknesses of these two algorithms. Furthermore, the techniques have been applied using numeric representation of words without considering domain/language information. Language dependent approaches such as that [16] considers contextual profiles of significant terms in Roman Urdu text is likely to improve results of classification.

V. BIBLIOGRAPHY

[1] [Federal Trade Commission, "Unsolicited Commercial E-Mail," 3 Nov. 1999.

[2] Pear Analytics. (August 2009) Twitter Study. [Online]. HYPERLINK "http://web.archive.org/web/20110715062407/www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf"

[3] Iqra Javed and Hammad Afzal, "Creation of Bi-lingual Social Network Dataset using Classifiers," in *Machine Learning and Data Mining in Pattern Recognition*. St Petersburg, Russia: Springer International Publishing, 2014, pp. 523-533.

[4] Iqra Javed and Hammad Afzal, "Opinion analysis of Bi-lingual Event Data from Social Networks," in *ESSEM@AI*IA*, Italy, 2013, pp. 164-172.

[5] Iqra Javed, Hammad Afzal, Awais Majeed, and Behram Khan, "Towards Creation of Linguistic Resources for Bilingual Sentiment Analysis of Twitter Data," in *Natural Language Processing and Information Systems:19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014*. Montpellier, France: Springer International Publishing, 2014, pp. 232-236.

[6] Kashif Mehmood, Hammad Afzal, Kashif Majeed, and Hassan Latif, "Contributions to the study of bi-lingual Roman Urdu SMS Spam Filtering," in *National Software Engineering Conference*, Rawalpindi, 2015.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," in *CEAS - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, Redmond, Washington, US., 2010.

[8] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, "A machine learning approach for Twitter spammers detection," in *International Carnahan Conference on Security Technology (ICCST), vol., no., pp.1,6, 13-16*, Oct. 2014.

[9] G. Stafford and L.L. Yu, "An Evaluation of the Effect of Spam on Twitter Trending Topics," in *International Conference on Social Computing (SocialCom), vol., no., pp.373,378, 8-14*, Sept. 2013.

[10] K. Kandasamy and P. Koroth, "An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques," in *IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2014.

[11] S.J. Soman and S. Murugappan, "Bayesian Probabilistic Tensor Factorization for Malicious Tweets in Trending Topics," in *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). vol., no., pp.895,900, 10-11*, July 2014.

[12] Zi Chu, S. Gianvecchio, Haining Wang, and S. Jajodia, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?," in *Dependable and Secure Computing, IEEE Transactions on , vol.9, no.6, pp.811,824,* Nov - Dec 2012.

[13] Kang Li, Zhenyu Zhong, and L. Ramaswamy, "Privacy-Aware Collaborative Spam Filtering," in *IEEE Transactions on Parallel and Distributed Systems, vol.20, no.5, pp.725,739*, May 2009.

[14] V. Vapnik, and D. Wu. H. Drucker, "Support vector machines for spam categorization.," in *IEEE Transactions on Neural Networks, 10(5):1048–1054*, 1999.

[15] Jiang Su, Harry Zhang, Charles X. Ling, and Stan Matwin, "Discriminative Parameter Learning for Bayesian Networks," in *Proceedings of the 25th International Conference on Machine Learning*. Helsinki, Finland: ACM, 2008, pp. 1016--1023.

[16] Hammad Afzal, Robert Stevens, and Goran Nenadic, "Towards semantic annotation of bioinformatics services: building a controlled vocabulary," in *Third International Symposium on Semantic Mining in Biomedicine.*, Turku, 2008, pp. 5-12.