Sentiment Analyses for Kurdish Social Network Texts using Naive Bayes Classifier

Abdusalam Abdulla Shaltooki
Computer Science. University of Human Development
Sulaimany, Iraq
salam.abdulla@uhd.edu.iq

Mzhda Hiwa Hama
Computer Science. University of Sulaimany
Sulaimany, Iraq
Mzhda.hiwa@gmail.com

***Abstract*— Language is a great tool to communicate and carry information. Moreover, it is used to express feeling and sentiment. These days sentiment analysis is one the most active field of research, to discover people's opinion about specific product, service or topic. The task of sentiment classification is to categories reviews of users as positive or negative from textual information of Social Networks like Facebook, Google+, Twitter and Blogs to determine the feeling of majority about specific topics. Kurdish language suffer from the unique and standard writing rules, grammar syntax and alphabet. Therefore, Kurdish people write their feeling in social networks in different ways. Some of them prefer to use the Arabic script style while others prefer to use Latin letters to express their feeling, further some people use their different accents and syntax and even sometimes they use English letters write their emotion. Therefore, for the purpose of analytics for Kurdish sentiment analyses its proposed to use data mining classification techniques such as Naive Bayes classifier because of its strong independence assumption. In Experimental results, the Social Network comments are classified into positive or negative polarities. The accuracy of sentiment analysis is obtained 66% by using Naive Bayes classifier for unigram feature on Kurdish text dataset.**

Keywords—Sentiment Analysis;Kurdish Sentiment;Naive Bayes Classifier.

I. INTRODUCTION

Language is a great tool to communicate and carry information. Moreover, it can be used to express feeling and sentiment [1]. Sentiment Analysis is a Natural Language Processing and Information Extraction task that aims to take writer's feelings expressed in positive or negative comments, tweets, questions and requests, by analyzing many documents. These days sentiment analysis is one the most active field of research, to discover people's opinion about specific product, service or topic. During the current past years, there is a dramatic increase in the Internet usage and specifically in Social Networks. In fact, exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract underlying public opinion and sentiment is a challenging task [2]. Ordinary millions of comments or opinions are posted in websites that provide the facilities for social networks like the Twitter, Facebook and Google+. The author of the comments share their feeling on different topics, discuss current topics even spot accidents or any flu epidemics. These are the treasured source of opinions and sentiments as huge amount of posts are posted by the users according to their used products and services, or express their different views on different perspectives. Researchers are using these posts to measure the public sentiment and to do sentiment analysis [3]. There are several challenges in Sentiment analysis. People do not express opinions in the same way; they use opinion words as positive or negative comments. In traditional text processing, a small differences between two sentences don not change the meaning very much; however, in Sentiment analysis with informal medium like Twitter or blogs, people combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse [4]. Popular machine learning methods that is used for classifying text include Naive Bayes, K-Nearest Neighbor, Support Vector Machines. In this paper, Comments has been collected from the microblogs Twitter, Facebook and Google+ and using Naïve Bayes for classifying "documents" into positive, negative sentiment.

II. RELATED WORK

Kurdish language suffer from the unique and standard writing rules, grammar syntax and alphabet. Therefore, Kurdish people write their feeling in social networks in different ways. Some of them prefer to use the Arabic script style while others desire to use Latin letters to express their feeling, further some people use their different accents and syntax and even sometimes they use English letters write their emotions and feelings. Therefore, the purpose of Kurdish sentiment analyses we need a specific algorithm that only care about the occurrence of each words in negative or positive sentences,

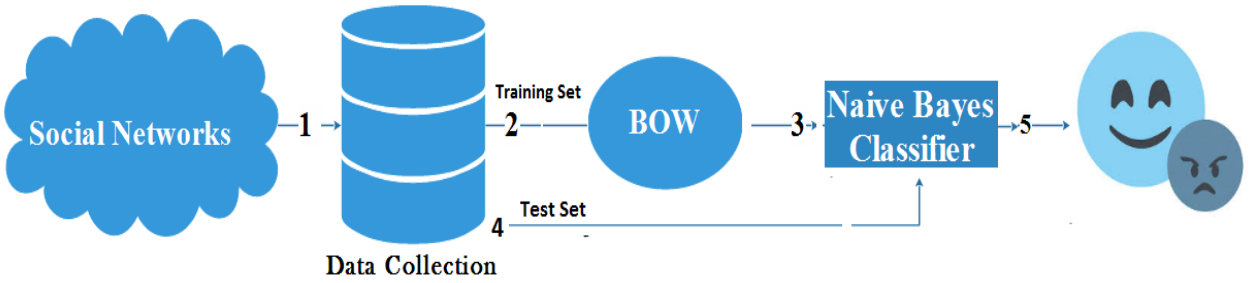

Figure 1: Proposed System Architecture

Regardless of the type of letters script and syntax and other issues. While, many research has recently focused on the analysis of sentiments of social media in order to get a feel for what people think about current topics of interest or specific products or services, in [5] they describes a strategy based on a Naive-Bayes classifier for detecting the polarity of English tweets. The experimental result, shown that the best performance is achieved by using a binary classifier between just two polarity categories: positive and negative. The F-score that is achieves after experiment result was 63%. A system has been designed in [6] for real-time analysis of sentiment toward presidential candidates in the 2012 U.S. Twitter has been used as data source and The statistical classifier have been used for sentiment analysis is a Naïve Bayes model on unigram features. The accuracy that have been achieved was 59% on the four category classification of negative, positive, neutral, or unsure. Sentiment analyzed in [7] on Spanish tweets. Twitter have been used as Twitter can be seen as a large source of short texts (tweets) containing user opinions, emotion analyzing this tweets is a challenge. The approach, Naïve Bayes classifier have been used for detecting the polarity of Spanish tweets. Results shown that accuracy of the system 67% which used for detect six sentiment categories. In [8], the utility of sentiment classification on a novel collection of dataset have been investigated which is Tunisian Facebook users. The originality of this collection leads not only on the nationality of the users, but also on the period of posting their statuses updates which is the Tunisian revolution. The dataset have been preprocessed by removing repeating group and stemming. Machine learning algorithms which are Naïve Bayes and the SVM have been used and result of both algorithm have been compared to each other. Although, Facebook statuses have unique characteristics compared to other corpuses (Reviews, News, etc), machine learning algorithms are shown to classify statuses with similar performance. The overall performance of the proposed methodology is also calculated.

Sentiment analysis in [9] has addressed for reviews expressed in the Arabic language, two dataset have been used the first dataset has been prepared manually by collecting reviewers' opinions from Aljazeera2 website against different published political articles. The dataset includes 322 reviews and the second dataset represents an Arabic opinion corpus that is freely available for research purposes. Naïve Bayes, SVM and K-NN classifier have been used and result of each of them has been represented and discussed.A modern approach towards sentiment classification is to use machine learning techniques which inductively build a classificatio n model of a given set of categories by training several sets of labeled document. Popular data mining methods include Naive Bayes, K-Nearest Neighbour, Support Vector Machines and Neural Network. In proposed system, Naive Bayes supervised method are used for classification.

III. ARCHITECTURE

The proposed system architecture is shown in Figure 1. As has been illustrated in this figure, Kurdish texts is collected from different Social Networks like Twitter, Facebook and Google+. This collected data set is stored in the specific database. Then in the second step, Naïve Bayes Classifier is used to apply on the training set, containing 70% of the data set. In this step, Bag of Words (BOW) for Kurdish Sentimental analyses is built. BOW contains frequency of occurrence of each word in negative and positive documents that is used as a unigram feature for training a classifier. A unigram feature marks the presence or absence of a single word within a text. By applying the classifier on the test set, which is 30% of total data set, and use of the prepared BOW, Kurdish Sentiment Analyzer is built.

IV. NAIVE BAYES CLASSIFIER

The Naive Bayes algorithm is a method that is often used for document classification. The Naïve Bayes method is from a machine learning method based on applying the Naive Bayes theorem with an independent assumption. The Naive Bayes method work performance has two stages in the document classification process: the learning stage and classification stage. In the learning stage, the classification process is done in sample documents that choose words from aspects that express sentiment in a document. Every aspect from an entity represents an opinion in a document by counting the probability of a word surfacing. The Naïve Bayes assumption is every word that expresses an aspect from an entity that is not dependent on the position of where the word is. Next, there is a document classification process by looking for maximum probability values from

every word that arises in a document with considering the Naïve Bayes classifier [10].

The Naïve Bayes model involves a simplifying conditional independence assumption. That is given a class (positive or negative) the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. From numerical based approach group, Naive Bayes has several advantages such as simple, fast and high accuracy [11]. In this algorithm Bayes's rule applied to documents and classes, for each document d and class c. In [12] describes the Bayes rule as :

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

Where c: a specific class
d : Documents wants to classify
P(d) and P(c) : prior probabilities
P(c|d) and P(d|c) : posterior probabilities
The best class the maximum a posteriori class (MAP) that has been looking for to assign this document is:

$$C_{MAP}=argmax\ P(c|d) \quad (2)$$

As in (1), by Bayes rule the equation is:

$$C_{MAP}=argmax\ \frac{P(d|c)P(c)}{P(d)}$$

Since probability of all the document are identical and constant so it is reasonable to eliminate the P(d) .

$$C_{MAP}=argmax\ P(d|c)P(c) \quad (3)$$

Now let us represent the documents by completely set of features: .In that case:

$$C_{MAP}=argmax\ P(x_1, x_2, x_3, x_4, \dots, x_n|c)P(c) \quad (4)$$

To make computations less complex the following simplifying assumptions has been mad:

1) BOW assumption: By assuming that position does not matter.
2) Conditional Independence: To assume that different features probabilities $P(x_i|c)$ are independent given the class c. Therefore, easily it is possible to write:

$$P(x_1, x_2, x_3, x_4, \dots, x_n|c)=P(x_1|c)*P(x_2|c)*P(x_3|c)*\dots*P(x_n|c)$$

Therefore, from (4), easily it can be written:

$$C_{MAP}= P(c_j)\prod_{i=1}^{n} P(x_i|c_j) \quad (5)$$

However, there is a problem with maximum likelihood

TABLE I. EXAMPLE

Set	Document	Words	Class
Training	1	کاتێک خۆشحاڵ نەبم کە گەشەکردن و پێشکەوتنی مرۆڤ نەبینم.	Positive
Training	2	بێزار وە نائومێد بووم لە جەنگ و ماڵوێرانی و ئاوارەیی.	Negative
Training	3	Em barudoxe têperr debêt we aşitî û aramî degerrêtewe.	Positive
Training	4	ئاشتی و پیکەوە ژیانی کۆمەڵایەتی سەقامگیر دەبێت.	Positive
Test	5	خۆشحاڵ نەبم کاتێک جەنگ بۆ ئاشتی بگۆڕێ.	?

because if one of the words in the BOW does not appear in the document, the $P(x_i|c_j)$=0 and it cause C_{MAP} to be equal to zero. To solve this problem classic Laplace or (add-1) smoothing for Naïve Bayes has been used, so in this case by having (6)

$$P(W_i|C_j)= \frac{Count(W_i,c)+1}{\sum_{i=1}^{k}(Count(w,c)+1)} \quad (6)$$

, k is the index of the last word in the BOW. Finally:

$$P(W_i|C_j)=\frac{Count(W_i,c)+1}{\sum_{i=1}^{k}Count(w,c)+V} \quad (7)$$

Where V is the size of the BOW.
An example has been shown in the table (I):

Calculate priori probability of pos and neg in multinomial model by :

P(Pos)=No of positive Docs/Total Number of Docs= 3/4
P(Neg)=No of Negative Docs/Total Number of Docs= 1/4

Calculate maximum likelihood smoothing Naive Bayes estimate by using (7):

p(خۆشحاڵ|pos) = (1+1)/(24+34) = 2/58
p(نەبم|pos) = (1+1)/(24+34) = 2/58
p(کاتێک|pos) = (1+1)/(24+34) = 2/58
p(جەنگ|pos) = (0+1)/(24+34) = 1/58
p(بۆ|pos) = (0+1)/(24+34) = 1/58
p(ئاشتی|pos) = (1+1)/(24+34) = 2/58
p(بگۆڕێ |pos) = (0+1)/(24+34) = 1/58
p(خۆشحاڵ|neg) = (0+1)/(10+34) =1/44
p(نەبم|neg) = (0+1)/(10+34) =1/44

p(کاتێک|neg) = (0+1)/(10+34) =1/44
p(جەنگ|neg) = (1+1)/(10+34) =2/44
p(بۆ|neg) = (0+1)/(10+34) =1/44
p(ئاشتی|neg) = (0+1)/(10+34) =1/44
p(بگۆڕێ |neg) = (0+1)/(10+34) =1/44

Calculate posteriori probability using (4):

$P(pos|d5) = 3/4*(1/58)^3*(2/58)^4$

$P(neg|d5) = 1/4*(1/44)^6*(2/44)^1$

Since P (pos|d5) is greater than P (neg|d5), then the statement in test set has positive polarity.

V. RESULTS

We implemented the classifier in Python and PostgreSQL to store collected documents and the created BOW. All the documents and text are collected from Twitter, Facebook and Google+ one by one. Each of them marked by one as conveying the positive feeling and zero for negatives. It is commonly used dictionary in sentiment analysis, but as mentioned before, a prototype of Kurdish BOW has been built for the purpose of Kurdish Sentimental Analyses. The BOW contains words with their positive and negative rates in different scripts style like Arabic, English and Latin. The BOW has 5,544 total keywords which includes 2,000 negative words and 3,544 positive words. The collected dataset contains 15,000 text files in which 8,000 labeled as positive reviews and the rest 7,000 labeled as negative reviews. The 70% of the dataset has been used in training set for building the BOW. The most frequent 5,544 words has been selected as the BOW. After experiment, the result of sentiment analysis using Naive Bayes classifier is obtained 66% accuracy on test data and F-Score of 0.72.

VI. PERFORMANCE METRICS

Performance metrics are used for the analysis of classifier accuracy. The proposed system is evaluated performance using accuracy and F-Score parameter. The accuracy can be computed using (8):

$$\frac{No\ of\ Correct\ Samples}{Total\ No\ of\ Samples} * 100\% \quad (8)$$

Table II shows the information about total number of documents in test set and correct and incorrect prediction samples of Naïve Bayes classifier. Using (8) and the data inside table (II), it is obvious that the accuracy is equal to 66%.

TABLE II. ACCURACY TABLE

Total No of documents in test set	Correct Sample	Incorrect Sample
4,500	2,991	1,509

Table (III) shows the confusion matrix, which displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data.

TABLE III. CONFUSION MATRIX

Classifier	Actual Class 1	Actual Class 0
Predicted Class 1	True Positive	False Positive
Predicted Class 0	False Negative	True Negative

Table (IV) illustrates the actual and predicted data for both positive and negative documents.

TABLE IV. CONFUSION MATRIX RESULTS FROM TEST SET

Naïve Bayes Algorithm	Actual document 1	Actual document 0
Predicted 1	1,990	999
Predicted 0	511	1,001

In order to compute F-Score, Recall and Precision needs to be computed. Table (V) shows the values for all which have calculated through the (9), (10) and (11) which is 0.72.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (9)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (10)$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (11)$$

TABLE V. RESULTS

Precision	Recall	F Score
0.665	0.79	0.72

VII. CONCLUSION

These days sentiment analysis is one the most active field of research, to discover people's opinion and feeling about a specific product, service or topic from Social Networks. Since Kurdish language has its own characteristics, there are challenges when it is come to Kurdish sentiment analyzing. Naïve Bayes classifier due to its strong conditional independence assumption is a great choice. Moreover, it is easy to implement and extremely fast to train. We also have shown through this paper that reasonable accuracy can be achieved comparing with similar works. The ideas of this paper, can be applied to other domains of Kurdish text classification.

VIII. FUTURE WORK

As future work, many suggestions can be put forward either to improve this work or the other related works. It's possible to make a special pre-processing for accuracy enhancement regarding the rules and syntax of Kurdish language. Use the different available Kurdish dictionaries and our BOW to making a complete BOW with different word scripts and spelling, will give a dramatic enhancement in accuracy of Kurdish sentiment analyses. Using more sophisticated machine learning algorithms like SVM and KNN algorithms for Kurdish sentiment analyses is also would be a great choice.

Acknowledgment

This research was supported by university of Human Development. We thank our colleagues from university of Sulaimani who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

References

[1] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 4 July-August 2014.

[2] Subhabrata Mukherjee,"Sentiment Analysis", Indian Institute of Technology, Bombay Department of Computer Science and Engineering,June 29, 2012.

[3] Md. Ansarul Haque1, Tamjid Rahman"SENTIMENT ANALYSIS BY USING FUZZY LOGIC", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 4,No. 1, February 2014.

[4] Fouzi Harrag ,"Estimating the Sentiment of Arabic Social Media Contents: A Survey", Research Center of College of Computer and Information Sciences, King Saud University under the project code (RC131001).

[5] Pablo Gamallo and Marcos Garcia," Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets∗",Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.

[6] Hao Wang, Dogan, Abe Kazemzadeh, François Bar and Shrikanth Narayanan ,"A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle",2012.

[7] Pablo Gamallo, Marcos Garcia and Santiago, "Tass: A naivebayes strategy for sentiment analysis on spanish tweets," in International Conference on Social Informatics, pp. 215-221, 2012.

[8] Safa Ben Hamouda and Jalel Akaichi," Social Networks' Text Mining for Sentiment Classification: The case of Facebook' statuses updates in the "Arabic Spring" Era", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 5, ISSN 2319 – 4847, May 2013.

[9] Rehab Duwairi , Mahmoud El-Orfali , "A Study of the Effects of Preprocessing Strategies on Sentiment Analysis for Arabic Text", Journal of Information Science, pp. 1-14, 2013.

[10] Gerit John Rupilele, danny manongga, wiranto herry utomo ,"Sentiment Analysis of National Exam Public Policy with Naive Bayes Classifier Method (nbc)", journal of theoretical and applied information technology, vol. 58 no.1, 10th december 2013.

[11] Narayanan, Vivek, Ishan Arora, and Arjun Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model", Intelligent Data Engineering and Automated Learning IDEAL, Lecture Notes in Computer Science Volume 8206, pp 194-201,2013.

[12] K.M.Leung,"Naïve Bayesian classifier," [Online].Available:http://www.sharepdf.com/81fb247fa7c54680a94dc0f3a253fd85/naiveBayesianClassifier.pdf,[Accessed: September 2013].