American Journal of Artificial Intelligence
2017; 1(1): 1-4
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20170101.11

A Model for Clustering Social Media Data for Electronic Learning

Erick Odhiambo Omuya

Department of Computing and IT, Zetech University, Nairobi, Kenya

Email address:
Omuya2005@gmail.com

To cite this article:
Erick Odhiambo Omuya. A Model for Clustering Social Media Data for Electronic Learning. *American Journal of Artificial Intelligence*. Vol. 1, No. 1, 2017, pp. 1-4. doi: 10.11648/j.ajai.20170101.11

Received: April 21, 2017; **Accepted**: May 11, 2017; **Published**: July 3, 2017

Abstract: Through Social media, people are able to write short messages on their walls to express their sentiments using various social media like Twitter and Facebook. Through these messages also called status updates, they share and discuss things like news, jokes, business issues and what they go through on a daily basis. Tweets and other updates have become so important in the world of information and communication because they have a great potential of passing information very fast. They enable interaction among vast groups of people including students, businesses and their clients. These numerous amounts of information can be extracted, processed and properly utilized in areas like marketing and electronic learning. This paper reports on the successful development of a way of searching, filtering, organizing and storing the information from social media so that it can be put to some good use in an electronic learning environment. This helps in solving the problem of losing vital information that is generated from the social media. It addresses this limitation by using the data from twitter to cluster students and by so doing support group electronic learning.

Keywords: Social Media, Twitter Application Programming Interface, Natural Language Processing, Twitter, Corpus

1. Introduction

Clustering is a descriptive task of data mining. A cluster is a collection of data objects that are either similar to one another in the same group or dissimilar to objects in other groups. Clustering uses unsupervised learning technique in finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Its objective is to get groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups [1]. It can be applied in various fields for instance taxonomy of living things, information retrieval from a document, identification of areas of similar land use in an earth observation database, discovering distinct groups by marketers in their customer bases for development of targeted marketing programs and identifying groups of houses according to their house type, value, and geographical location [2]. A number of techniques can be used to do clustering. Some of them include summarization, compression and k-nearest neighbor which localizes search to one or a small number of clusters. Good clustering methods produce high quality clusters with either a high intra-class similarity within clusters or a low inter-class similarity between clusters. The quality of clustering also depends on both the similarity measure used by the method and its implementation [10]. The quality of a clustering method is also measured by its ability to discover more or all the hidden patterns. Clustering is the concept that was used in this research to create groups from social media data which can be used for learning on electronic learning platforms.

2. Methodology

This section looks at how the system for creating discussion groups was developed as well as a detailed explanation of the research method that was used to realize the objective of the study. The system design methodology used was incremental prototyping. In incremental prototyping, the whole requirements are broken down into building blocks which are incremented each time a new component is integrated based on an overall design solution. Typically development starts with the external features and user interface, and then adds features as prototypes are developed. Requirements and Architectural Design can be

done up front and then each prototype developed as the project progresses. The solution is complete when all the components are in place.

Several activities were performed to come up with the system. The first task was to retrieve details of each of the students from their twitter accounts using an extension script which is part of the twitter Application Programming Interface. The second task involved identifying the right kind of data to use for training the expected prototype as well as testing it. This generally dealt with preliminary processing of the data collected from the users to do away with any inconsistencies and outliers [11]. These unwanted features are not very good because they can easily cause the system to perform irregularly. The third step involved using the data already preprocessed above to train the prototype. The machine learning method used was unsupervised learning in which the system was given the data so that it automatically analyzes and creates clusters from the data [3]. The relationship between the data items can be established using the k-nearest neighbor technique. From this, we identified the groups that students' fall that were turned into discussion groups.

The fourth step was testing the system. The end result of the learning process was the model which was able to do classification with very minimal margins of error. The prototype was then subjected to testing using the test data. This is a collection of data whose class labels are already known. They are part of the data that was used to train the system but its results are already known. They were used to confirm that the system indeed accurately did the classification given some data items. Finally, the model was used to classify a new user into a group. This involved picking the details of a new student from twitter and trying to predict the class hence group that he should join.

The illustration of the proposed prototype is given below.

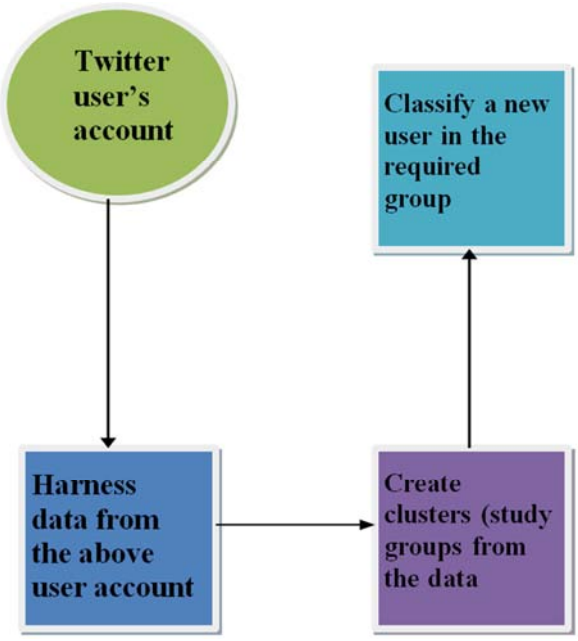

Figure 1. *Proposed Prototype.*

3. Results

The results reported in this paper were obtained from a series of evaluations that were done on the classifier on different parameters including functionality, usability, accuracy, precision and recall.

3.1. Evaluation of the Prototype on Functionality

This is the section that captured the users view on the functioning of the prototype. The first task was to determine if the prototype achieved its overall goal which is grouping students through social media for electronic learning. On this question touching on the overall goal, 90% of the students emphatically agreed that the system actually enabled them to be classified into groups and they were therefore able to know their group members and comfortably interact with them on a given task that they were assigned [4].

They also confirmed that the system simplified the process of group formation and made inclusivity of distant students in the groups possible. A good number of learners indicated that they would continually use the system for the purposes of group formation and discussion. This is summarized in the chart below.

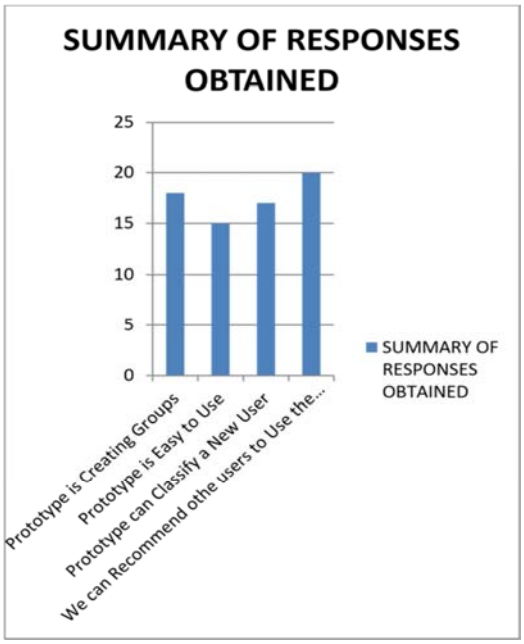

Figure 2. *Summary of Responses from Prototype End Users.*

3.2. Evaluation of the Naïve Bayes Classifier

The Naïve Bayes Classifier was also tested to evaluate its accuracy, precision and recall [9]. In experimenting with the Naïve Bayes Classifier, we relied on the NLTK module which provides functions for calculating these measures for the classifier. A total of 200 tweets were extracted and used for this test which was summarized in a confusion matrix. This matrix consists of the following parameters: TP, TN, FP and FN, which are defined below.

True Positives (TP): number of positive examples, labeled

as such.

False Positives (FP): number of negative examples, labeled as positive.

True Negatives (TN): number of negative examples, labeled as such.

False Negatives (FN): number of positive examples, labeled as negative.

Classifier Accuracy, Precision and Recall

Accuracy: This is the proportion of correct results that a classifier achieved. If, from a data set, a classifier could correctly guess the label of half of the examples, then we say its accuracy was 50%.

$$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN)$$

From these dummy results the accuracy can be calculated as:

$$\text{Accuracy} = (10+100)/(10 + 5 + 15 + 100) = 84.6\%$$

Precision: This measure determines what fraction is correct out of all the examples the classifier labeled as positive.

$$\text{Precision} = TP/(TP + FP)$$

Recall –This measure determines what fraction the classifier picked up out of all the positive examples that were there.

$$\text{Recall} = TP/(TP + FN)$$

The results below illustrate a summary of what was obtained when 200 tweets were used to test the Naïve Bayes Classifier.

Table 1. *Actual Classifier Results.*

Feature	Accuracy	POS Precision	POS Recall	NEG Preci-sion	NEG Recall
Uni-gra ms	*0.714*	*0.502*	*0.950*	*0.937*	*0.426*

This classifier was doing the classification using the unigrams. This is where the tweets were being divided into single words which were analyzed before being classified. From this analysis the classifier performed above average with an accuracy of 71.4%. Precision and recall were however average. These measures can be improved if large amounts of data are used to train the classifier before being used to do actual classification.

4. Discussion

Through the study, it can be underscored that inasmuch as the social media has a great potential in education, this has not been exploited to a greater percentage. The techniques that are currently used in group formation and learning are mostly manual and so not efficient. They therefore come with a lot of challenges including time wastage. Through social media a better and more efficient way can be used to enable online learning generally and group formation specifically.

The system that was developed by the researcher demonstrated the learning capability of the social media by coming up with a way of creating study groups from the information shared across the social media. It was able to extract tweets from various social media accounts based on a given hash tag (task) and then pass them to a Naïve Bayes classifier as input. The classifier then grouped the users into different categories based on various tweets that they posted on the task. The classifier was also able to assign other or new users groups also according to their tweets and the learning that the system had undergone.

The system was therefore able to address the limitation of the social media of not being properly utilized as a platform for supporting learning activities like group formation. This paper addresses the limitation of social media of not being properly utilized as a platform for supporting learning activities like group formation. Most of the information that passes through social media was being used majorly for social interaction. The study has proved that it can actually be used constructively in learning in various institutions.

5. Conclusion

Through the study, it was underscored that inasmuch as the social media has a great potential in education, this has not been exploited to a greater percentage. The techniques that are currently used in group formation and learning are mostly manual and so not efficient. They therefore come with a lot of challenges including time wastage. Through social media a better and more efficient way of clustering can be used to enable electronic learning generally and group formation specifically.

References

[1] Dongwoo, K., Yohan, J., Il-Chul, M., and Oh, A. (2010). Analysis of twitter lists as a potential source for discovering latent characteristics of users. Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems.

[2] Yardi, S.; Romero, D.; Schoenebeck, G.; and Boyd, D. (2010). Detecting spam in a twitter network. First Monday 15: 1–4.

[3] Light, V, Nesbitt, E, Light, P & Burns, JR. (2000). Let's You and Me Have a Little Discussion: computer mediated communication in support of campus-based university courses, Studies in Higher Education, vol. 25, no. 1.

[4] Brook, C. and Oliver, R. (2003). Online learning communities: Investigating a design framework. Australian Journal of Educational Technology, 19 (2), 139-160.

[5] Nichols, M. (2003). A theory of eLearning. Educational Technology & Society, 6 (2), 1−10.

[6] Benson, A. (2002). Using online learning to meet workforce demand: A case study of stakeholder influence. Quarterly Review of Distance Education, 3 (4), 443−452.

[7] Hiltz, S. R., &Turoff, M. (2005). Education goes digital: The evolution of online learning and the revolution in higher education. Communications of the ACM, 48 (10), 59−64, doi: 10.1145/1089107.1089139.

[8] Griffiths, T. L., Steyvers, M., &Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114, 211-244. (pdf)(topic modeling toolbox) .

[9] Pak, A., &Paroubek, P. (2010). Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation (pp. 436-439). Association for Computational Linguistics.

[10] Yessenov, K. and Misailovic, S. (2009). Sentiment Analysis of Movie Review Comments. Available: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf. Last accessed 13th July 2011.

[11] Lu H, Sun S, Lu Y (2006). Preprocessing data for effective classification. ACM SIGMOD'96 workshop on research issues on data mining and knowledge discovery, Montreal, QC Nichols, M. (2006). A theory of eLearning. Educational Technology & Society, 6 (2), 1−10.