ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Post Filtering Mechanism in Social Networking

Pooja B. Bhor[1], Prof. Sheetal Thakare[2]

PG Student, Department of Computer Engg, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India[1]

Department of Computer Engg, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India[2]

ABSTRACT: Online social networks sites have provided platform for online social groups where individuals can collaborate and shuffle their thoughts. In this social space, content will be huge. For specific interest of user, we need mechanism which filter social media posts based on user's interest. This can be done by an automated filtering method; required to categorize members based on their pattern of response. The posts are clustered based on stylistic, thematic, emotional, sentimental and psycholinguistic methods. After members are categorized based on their response to the posts belonging to different clustering methods. The categorization affords recommending the users, posts which might be useful to the interest of the members. The categorization posts show increased performance in case of large number of candidate members by performing clustering based on linguistic features. The contribution work is to implement location-aware personalized posts recommendation using both the users' personal interests and their geographical contexts. The system has been tested on Twitter API group data; where it offers a significant solution to an unaddressed problem associated with social networking groups.

KEYWORDS: Emotion analysis, Psycholinguistics, Sentiment analysis, Stylistics Clustering, Geographical Context

I. INTRODUCTION

The most popular social networking site Twitter has groups with over 100K members in it. Thus, it becomes difficult for the admin to track the members violating group policies. This shows that there exists a need for a measure to categorize the posts made by members in it based on acceptability and group behavior. A community is a fraternity that seeks a platform to discuss subjects that relate to the common cause that motivated the establishment of the group. The members within it enjoy discussion with regard to the purpose of the group. This is especially applicable in the case of academic and interest groups. Thus posting articles irrelevant to these groups can create unnecessary clutter that affects the comfort of the members within the group. Unnecessary advertisements and marketing matters targeting a different audience should be avoided from a group. There is no existing method to deal with this.

A method to control the influx of irrelevant messages within a community is very much essential for the smooth functioning of a social networking group. The existing policies of popular social networking site Twitter enable the administrator of a group to delete and monitor posts made by the members of a community. To screen all the messages posted by the members of a heavily populated group is very difficult. The existing settings provide no option for automated notification for group admin with regard to the members involved in posting articles that do not match the general interest of the group. Our proposed method aims at providing an automated notification to the group moderators regarding suspicious members who frequently post articles that appear to gather negative response from the members within the group. This novel approach has been experimented to account for the hypothesis that, though members may be socially connected; there might be disparity in their likings, thoughts, sentimental orientation and thematic inclination.

The members belonging to the same clusters are those who exhibit resemblances with respect to sentimental, thematic, emotional, writing style and concept-related interests.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

II. POST FILTERING MECHANISM

2.1 Sentiment Clustering

The posts in a network group vary based on the sentimental status associated with the contents. The overall sentiment associated with a post can be positive, negative or neutral. The assessment of the sentiment is made at the keyword level. This enables to classify the posts based on the attitude towards the trends. The response of a member to a post shows his/her attitude towards the entities. This can be illustrated by an **Example :**

- User A says “I love ice cream. I Love to have it every time”
- While User B says “I hate ice cream. I don’t want it at all”

Here the keyword is ice-cream and the sentiment of the keyword with respect to user A is positive while that of user B is negative. Thus, analyze the sentiments associated with prominent keywords associated with a group.

Here, the posts are checked for total sentimental orientation, regarding whether the post is a negative post, positive post or a neutral post. We use K-means clustering to cluster the data. This is mainly due to the fact that K-means clustering exhibits a better computational speed in comparison with
hierarchical clustering. The results guaranteed by K-means give a better closeness in case of our dataset. The experiments have been conducted using hierarchical clustering. The results illustrate a better accuracy in case of K means clustering. The K-value associated with K-means clusteringis obtained by choosing K for which the average distance of points from centroid is minimum.

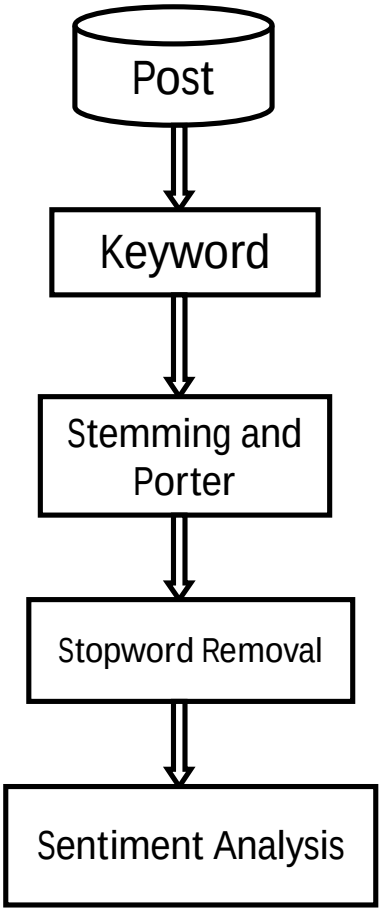

Fig.1 Overview of Sentiment clustering Approach

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

2.2 Theme Clustering

The next type of clustering performed on the posts is the theme based clustering. Here the posts are grouped based on the thematic similarities. After consider concepts, entities and topics for determining the theme. Use the semantically rich common sense knowledge base ConceptNet for extracting contextual and conceptual information of a post. The ConceptNet toolkit provides numerous assertions related to a word [1]. Utilize them to arrive at a general concept of a post. Initially the posts are chunked to obtain keywords and key phrases. These phrases are fed to the ConceptNet to obtain generalized concepts. After exploit the analogy making and topic gisting features of ConceptNet to arrive at a conclusion regarding the important concepts of a post.

Entity recognition is again made with the help of Alchemy API. The entity type is of much importance to us for our study of relatedness between posts. For **example** a post describing about ``iPhone'' very often has an entity type ``technology''. Those people who are technologically interested tend to follow them frequently. Thus, the type of entity is of great importance in the context of the proposed approach.

Once the entities and concepts are obtained, their importance in a post is determined by means of tf(term frequency)-idf(inverse document frequency) value. This gives us a measure of how different types of post differ from each other in terms of concept- entity variation.

$$TF_{concept} = concept_p / concept_{tot} \quad \ldots 1$$

where $concept_p$ refers to the frequency of occurrence of given concept in the post and $concept_{tot}$ refers to the total number of occurrence of concept over.

$$IDF_{concept} = \log_e (P_{tot}/P_{concept}) \quad \ldots 2$$

Where p_{tot} refers to the total number of posts within the group and $p_{concept}$ refers to the number of posts with given concept

$$TDIDF_{concept} = TF_{concept}.IDF_{concept} \ldots 3$$

Using the tf-idf score obtained from the above equation, we formulate a weight vector associated with the concepts and entities prevailing in a group.

2.3 Emotional Clustering

Emotion based aspects can be investigated to gain insight into a person's emotional attitude. The similarity in these aspects can be exploited to predict the liking and sharing probability of members within a community. Thus, the users within a community are grouped together based on the emotional aspects. The clustering based on emotion is performed by taking into account the score of the entire post with respect to the emotional categories namely anger, sadness, fear, disgust and joy. The score corresponding to the categories is again obtained by the service of Alchemy language API. The processis repeated in the same manner as sentiment clustering. Here the prominent keywords obtained using Algorithm 1 are made to undergo emotional analysis. These features are combined together with the overall emotion of a post. This forms the feature vector for emotional clustering.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

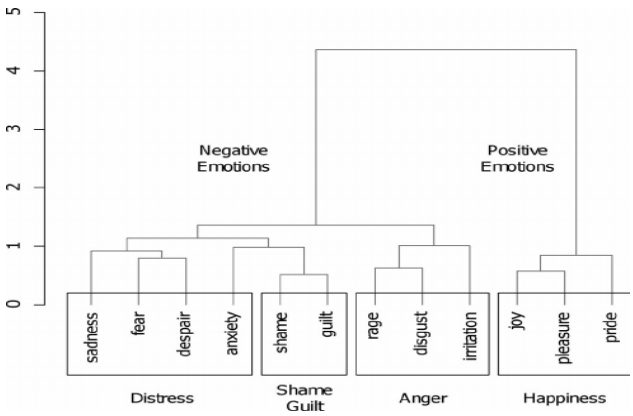

Fig.2 Emotional Clustering

2.4 Stylistic Clustering

Stylistics refers to the writing-style followed in a document. Each person has a unique writing style of his or her own. The writing-style factor has been considered because the writing style followed determines the popularity of an article. There are phrases and usages peculiar to authors that can be of interest to the audience. This aspect has been exploited to find the like-minded audience of a particular style of writing. The stylistic features such as words and character n-grams can capture the writing style effectively. These features are clustered by using K-means clustering. The number of clusters has been determined by cross validation.

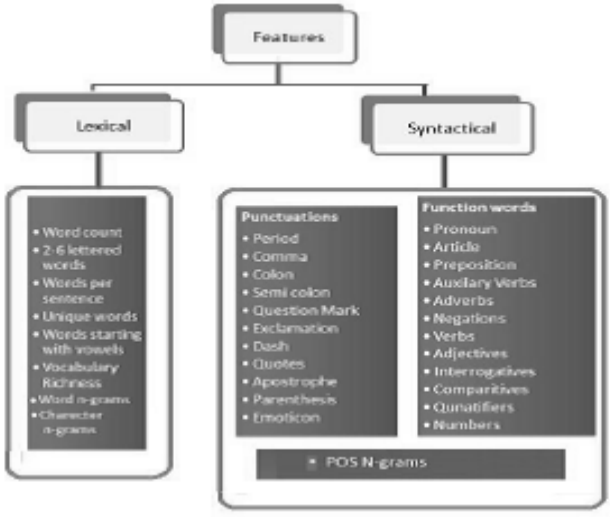

Fig 3. Overview of Stylistic Clustering[1]

2.5 Psycholinguistic Clustering

Psycholinguistics is the study of language processing mechanisms. Psycholinguistics like to study how word meaning, sentence meaning, and discourse meaning are computed and represented in the mind. It is mostly an unconscious process. Example: We think when we're reading words on a page that it's a smooth process, but our eyes actually jerk across the page – a process called saccadic motion. Sometimes they trying to access a word based on meaning, spelling, initial letter, rhyme, etc. Ex: "Oh, it's that color that's really bright green….and it's also a really strong liquor…but it sounds a bit like "loose"….it starts with a "sh" sounds….chartreuse! That's it!" What it shows: how

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

words are organized in the mind = mental lexicon. Access of the mental lexicon must be very quick, since word recognition takes just 1/3 of a second.
The posts have been subjected to the analysis of psycholinguistic orientation of the posts. The psycholinguistic differences contribute to the nature of the posts and hence there will be difference in the audience based on the psycholinguistic perspective of members within a group. Psycholinguistic aspects throw light on various factors that vary based on a personś personality, hobbies, passions, intellects, perception and context of references. Thus it reflects a personś way of responding to a post. As per the theory of homophily [5], which refers to the general proneness of individuals to associate with others with nature similar to them, we can say that people with similar psycholinguistic characteristics show similarity in rating posts. In our approach we make the best use of the psycholinguistic features to categorize people within a social networking community. The psycholinguistic aspect computation has been performed by considering various psycholinguistic features obtained from LIWC(Linguistic Inquiry Word Count) [6] and MRC(Medical Research Council) psycholinguistic database [7].

A new content based method for personalized tweet recommendation, based on conceptual relations between users' topics of interest. The Concept Graph is a way to exploit logical relations between topics of interest in order to provide interesting and efficient tweet recommendations.[1]here the recommender still recommends some tweets that were not retweeted by the user.
A TWIMER framework the use of language models as a basis for analyzing strategies and techniques for tweet advice based on person's interest profiles. TWIMER consists of several components, including tweet retrieval (query formation and relevance model), tweet relevance verification, and final relevance ranking[2].
Collaborative topic Poisson factorization (CTPF) can be used to build recommender systems through gaining knowledge of from reader histories and content material to advocate personalized articles of hobby. CTPF models both reader behavior and article texts with Poisson distributions, connecting to the latent subjects that represent the texts with the latent choices that shows to the readers[5].
Another technique is Co-Factorization Machines (CoFM), which deal with two (multiple) aspects of the dataset where each aspect is a separate FM. This type of model can easily predict user choices even as modeling user interests via content material at the equal time[4]. But the services are only interacting with the user's interest not on user's behaviors.
The algorithms for selecting a characteristic review set, which evaluate via experiments on a wide range of datasets of real reviews from different domains[6]. Here, positive and negative comment can't generalized to arbitrary domain.
Another method is the review set selection problem where given a set of reviews for a specific item, and want to select a comprehensive subset of small size. Provides authentic review using TOPQLTY algorithm sorting technique problem is based on limited review set[7].
Methods to compute quality, diversity and coverage properties using multidimensional content and context data. The proposed metrics so as to evaluate the picture summaries based totally on their illustration of the bigger corpus and the capacity to meet user's information needs[8]. Here Computation is very expensive.
Methods for incorporating social context in the quality prediction: either as features, or as regularization constraints, based on a set of hypotheses. The method proposes quite generalizable and applicable for quality (or attribute) estimation of other types of user-generated content[9]. But a portal may lack an explicit trust network.In multi-document summarization, redundancy is a particularly important issue since textual units from different documents might convey the same information[10]. A high quality (small and meaningful) summary should not only be informative about the remainder but also be compact (non-redundant).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Table I : SUMMARY OF EXISTING APPROACHES

Author	Algorithms/ Techniques	Parameter consider	Advantages	Disadvantages
D. P. Karidi, Y. Stavrakas, Y. Vassiliou, "A Personalized Tweet Recommendation Approach Based on Concept Graphs", In Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016,pp. 253–260	Content based method on conceptual relations between users' topics of interest (ToIs)	Twitter API	1. Provides a social media user with a new timeline that contains messages that strongly match ones interests and that are not necessarily posted by ones followings. 2. This model is effective and efficient to recommend interesting tweets to users.	1. The recommender still recommends some tweets that were not retweeted by the user.
R. Makki, A. J. Soto, S. Brooks, E. E. Milios, "Twitter Message Recommendation Based on User Interest Profiles", In Advances in Social Networks Analysis and Mining (ASONAM),IEEE/ACM International Conference, 2016,pp. 406–410	TWIMER NEI	TREC 2015 Microblog track dataset	1. Automatic query expansion. 2. Higher performance in the language model retrieval entities.	1. Graph-based approaches are not implemented.
P. K. Gopalan, L. Charlin, D. Blei, "Content-based recommendations with Poisson factorization", In Advances in Neural Information Processing Systems, 2014, pp. 3176–3184	Collaborative Topic Poisson Factorization (CTPF) Model	arXiv dataset	1. CTPF performs well in the face of massive, sparse, and long-tailed data. 2. CTPF provides a natural mechanism to solve the "cold start" problem. 3. CTPF scales more easily and provides significantly better recommendations than CTR.	1. This method is not feasible on Facebook group posts.
L. Hong, A. S. Doumith, B. D. Davison, "Co-factorization machines: modeling user interests and predicting individual decisions in twitter" ,In	Co-Factorization Machines (CoFM) Algorithm	Twitter API dataset	1. CoFMcan easily predict user decisions while modeling user interests through content at the same time. 2. Factorization	1. The services are only interacting with the user's interest not on user's behaviors.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 557–566			Machines to text data with constraints can mimic state-of-the art topic models and yet benefit from the efficiency of a simpler form of modeling.	
K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu ,"Collaborative Personalized Tweet Recommendation", In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012, pp. 661–670.	Collaborative Ranking Method For Tweet Recommendation (CTR) model	Twitter Dataset	1. A collaborative ranking method is better than collaborative filtering for different optimization criterion. 2. The proposed CTR method greatly improves the recommendation performance. 3. The CTR method is generic; it is easy to incorporate more information by adding extra features.	1. It only works on user's interests over time not on user's history and tags of the tweet.
T. Lappas, M. Crovella, and E. Terzi, "Selecting a characteristic set of reviews," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp. 832–840.	The Greedy, Integer-Regression and Iterative-Random algorithms.	Review Dataset	1. To accurately emulate the opinion distribution in the underlying corpus. 2. Improvement by previous work.	1. Positive and negative comment can't generalize to arbitrary domain.
P. Tsaparas, A. Ntoulas, and E. Terzi, "Selecting a comprehensive set of reviews," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 168–176.	Greedy Algorithm	Amazon.com and Tripadvisor.com review dataset	1. Performance is statically significant. 2. High quality of the review.	-
P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.	Greedy Algorithm for Summarization	Flickr, Picasa personal photos dataset	1. The greedy algorithm for summarization performs better than the baselines. 2. Summaries help in effective sharing and browsing of the personal photos.	1. Computation is expensive.
Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in Proc. 19th Int. Conf. World Wide	Text-Based Quality Prediction Method	Cellphones, Beauty, and Digital Cameras datasets	1. Improve the accuracy of review quality prediction. 2. The resulting predictor is usable even when social context is	1. A portal may lack an explicit trust network.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 2, February 2019

Web, 2010, pp. 691–700.			unavailable.	
H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.	Modified Greedy Algorithm	DUC04 dataset	1. The best performance is achieved. 2. Submodular summarization achieves better ROUGE-1 scores.	1. The proposed system very expensive to solve.

III.CONCLUSION

The paper discusses post recommendation technique in social networking sites such as Twitter. The method follows a differentapproach by considering linguistic features of data that arereadily available from a social networking platform. The posts are clustered based on stylistic, thematic, emotional, sentimental and psycholinguistic methods. The method shows a considerable degree of accuracy in predicting the response ofa member to a post. After members are categorized based on their response to the posts belonging to differentclustering methods. Then user will get post recommendation based on their interest and their geographical location.

REFERENCES

[1] D. P. Karidi, Y. Stavrakas, Y. Vassiliou, "A Personalized Tweet Recommendation Approach Based on Concept Graphs", In Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016,pp. 253–260
[2] R. Makki, A. J. Soto, S. Brooks, E. E. Milios, "Twitter Message Recommendation Based on User Interest Profiles", In Advances in Social Networks Analysis and Mining (ASONAM),IEEE/ACM International Conference, 2016,pp. 406–410
[3] P. K. Gopalan, L. Charlin, D. Blei, "Content-based recommendations with Poisson factorization", In Advances in Neural Information Processing Systems, 2014, pp. 3176–3184
[4] L. Hong, A. S. Doumith, B. D. Davison, "Co-factorization machines: modeling user interests and predicting individual decisions in twitter" ,In Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 557–566
[5] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, Y. Yu ,"Collaborative Personalized Tweet Recommendation", In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, 2012, pp. 661–670.
[6] T. Lappas, M. Crovella, and E. Terzi, "Selecting a characteristic set of reviews," in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2012, pp. 832–840.
[7] P. Tsaparas, A. Ntoulas, and E. Terzi, "Selecting a comprehensive set of reviews," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2011, pp. 168–176.
[8] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, 2011, p. 4.
[9] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 691–700.
[10] H. Lin and J. Bilmes, "Multi-document summarization via budgeted maximization of submodular functions," in Proc. Human Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 912–920.