2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-7281-5791-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAECT49130.2021.9392398

User Profiling based on Tweeter Data using WordNet and News Paper Archive

Antara Pal[1] and Alok Ranjan Pal[2]

[1]Dept. of Computer Science and Engineering, Pailan College of Management and Technology, Joka, Kolkata-104, West Bengal, India
[2]Dept. of Computer Science and Engineering, College of Engg. and Mgmt., Kolaghat, 721171, West Bengal, India
E-mail: chhaandasik@gmail.com antarapal22@gmail.com

Abstract: **In this paper, a method has been proposed for user profiling based on tweeter data. The sentiments of the tweets are retrieved programmatically with the help of WordNet and News Paper Archive. In this experiment, the English WordNet 2.1 has been used as an online semantic dictionary and machine readable version of the "Times of India" news paper has been used to generate a news paper archive. The algorithm is tested on a data set of 1000 tweets from four different categories which are initially tagged by their innate senses for validation of the derived result.**

First of all, the data set is evaluated with the help of newspaper archive by using lexical overlap and the accuracy in sense retrieval task is 48.7%. The reason behind this scenario is the varieties of representations of a single statement in natural language which creates a mare similarity between the lexical entities of the statements. To overcome this problem, the contexts of the tweets are expanded with the help of WordNet by considering the synonyms of every meaningful word of the tweets and after that the senses of these tweets are evaluated. As the contexts of the statements are expanded in this approach, semantic relatedness between the statements is resolved in an efficient way which leads the system towards a better performance.

Keywords— User Profiling; WordNet; Newspaper Archive; Tweeter Data; Synset Analysis

I. INTRODUCTION

User Profiling is one of the major demands of current era. It is used for several purposes, like- a) personalized recommendation where advertisement of selective products is presented to a user based on his/her activities on internet, b) sentiment analysis of a person whose state of mind may have an impact on his/her surroundings, c) sentiment of a community which could be used for any administrative decision making, etc.

In this experiment, user profiling is performed based on a user's tweets. The overall experiment is carried out in two phases. First, the sentiments of the tweets are derived based on lexical similarity considering the news paper archive as a reference. As the information is stored in the news paper archive in a categorized manner, this knowledge base has been used as a reference for sense resolution. But, the performance of the system was not too much appreciable in this phase. The reason, observed that natural language is creative and every individual expresses his/her views in different ways. So, establishing a relation between a pair of statements based on only lexical similarity is not a wise move. So, in the next step, semantic relation between the statements is calculated beyond the lexical similarity. To do this task, contexts of the tweets are expanded by synset analysis of the meaningful words of the tweets with the help of WordNet. As the contexts of the tweets are expanded, the semantic relation between a pair of statements is identified in a better way which leads the system towards a better accuracy.

In this experiment, English WordNet 2.1 has been used as an online semantic dictionary and the news paper archive is prepared from the online version of the "Times of India" news paper.

II. SURVEY

Zhongqi Lu *et al.* [1] proposed a Collaborative Evolution model, which learns the evolution of user's profiles through the sparse historical data in recommender systems and outputs the prospective user profile for the future. To verify the effectiveness of the proposed model, the authors conduct experiments on a real-world dataset which is obtained from the online shopping website- www.51buy.com and contains more than 1 million users' shopping records in a time span of more than 180 days.

O. Hasan *et al.* [2] proposed a work on user profiling with big data techniques and the associated privacy challenges. The authors also discussed the ongoing EU-funded EEXCESS project as a concrete example of constructing user profiles with big data techniques and the approaches being considered for preserving user privacy.

Grcar Miha *et al.* [3] proposed a work to address the problem of personalized information delivery related to the Web that is based on user profiling. The authors have analyzed different approaches to user profiling, like- content based filtering, collaborative filtering and Web usage mining. They have presented an overview of the approaches including recent research results in the area with especial emphases on user profiling in the perspective of Semantic Web applications.

Schiaffino Silvia and Amandi Analía [4] proposed a work based on the main issues regarding use profiles from the perspectives of different research fields. The authors examined what information constitutes a user profile; how the user profile is represented; how the user profile is acquired and built; and how the profile information is used. They also discussed some challenges and future trends in the intelligent user profiling area.

Farseev Aleksnadr *et al.* [5] have proposed a work based on different user profiling approaches on social networks. The authors have highlighted the challenges, techniques, and future trends. They explained the weakness and strength of these methods and introduce an analytic platform to bridge the gap between social media users, business intelligence and the Big Data.

Rahdari Behnam and Arabghalizi Tahereh [6] proposed different approaches and carried out various experiments to support an explanation concerning the categorization of social media users based on the texts they share about a specific event. The authors took a systematic approach to accomplish this objective by applying topic modeling techniques, using statistical and data mining algorithms, combined with information visualization.

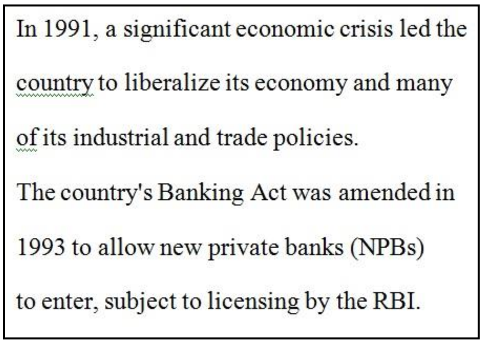
Dickinson Ian *et al.* [7] presented a framework for personal agents that respect the privacy of the individual. The authors presented some motivations and outline a framework for the use of personal agents and user profiling for information systems designed around web services. The key element of their approach in general is to consider the impact of user-profiling and autonomous agents on the user.

Jie Tang *et al.* [8] proposed a work on the problem of user profiling which is aimed at finding, extracting, and fusing the semantic based user profile from the Web. Their work formalizes the profiling problem as several subtasks: profile extraction, profile integration, and user interest discovery. They propose a combination approach to deal with the profiling tasks. Specifically, they employ a classification model to identify relevant documents for a user from the Web and propose a Tree-Structured Conditional Random Fields (TCRF) to extract the profile information from the identified documents. The authors proposed a unified probabilistic model to deal with the name ambiguity problem when integrating the profile information extracted from different sources. Finally, they used a probabilistic topic model to extract user profiles and constructed the user interest model.

III. PREPROCESSING

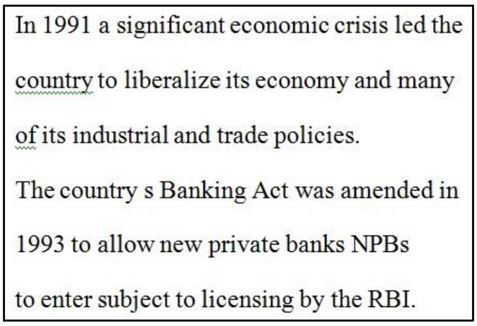
Text Normalization

The tweets, collected from the internet are not adequately normalized. So, a series of text normalization steps have been followed before the work, as- a) detachment of punctuation marks like single quote, tilde, double quote, parenthesis, comma, etc. that are attached to the words; b) removal of angular brackets, uneven spaces, broken lines, slashes, etc. from the sentences; and c) identification of sentence terminal markers (i.e., full stop, note of exclamation, and note of interrogation), etc. Sample non-normalized and corresponding normalized versions of a tweets are given in Figure-1 and Figure-2 respectively.



In 1991, a significant economic crisis led the country to liberalize its economy and many of its industrial and trade policies.
The country's Banking Act was amended in 1993 to allow new private banks (NPBs) to enter, subject to licensing by the RBI.

Fig.1. A sample non-normalized tweet.

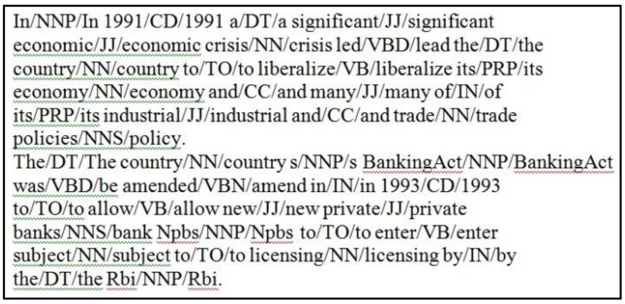


In 1991 a significant economic crisis led the country to liberalize its economy and many of its industrial and trade policies.
The country s Banking Act was amended in 1993 to allow new private banks NPBs to enter subject to licensing by the RBI.

Fig.2. Normalized form of the non-normalized tweet.

Text Lemmatization

Lemmatization is the process of reducing the inflected words into their root forms. To increase the lexical coverage, the words are lemmatized before the execution. In this work, the texts are lemmatized using CST lemmatiser tool. The lemmatized form of the normalized tweet (refer Figure 2) is given in Figure 3.



In/NNP/In 1991/CD/1991 a/DT/a significant/JJ/significant economic/JJ/economic crisis/NN/crisis led/VBD/lead the/DT/the country/NN/country to/TO/to liberalize/VB/liberalize its/PRP/its economy/NN/economy and/CC/and many/JJ/many of/IN/of its/PRP/its industrial/JJ/industrial and/CC/and trade/NN/trade policies/NNS/policy.
The/DT/The country/NN/country s/NNP/s BankingAct/NNP/BankingAct was/VBD/be amended/VBN/amend in/IN/in 1993/CD/1993 to/TO/to allow/VB/allow new/JJ/new private/JJ/private banks/NNS/bank Npbs/NNP/Npbs to/TO/to enter/VB/enter subject/NN/subject to/TO/to licensing/NN/licensing by/IN/by the/DT/the Rbi/NNP/Rbi.

Fig.3. Lemmatized form of a normalized tweet.

IV. PROPOSED APPROACH

After collecting the tweets from the tweeter handles of different persons, those are passed through a series of preprocessing steps. After that the work is carried out in two phases.

First, the innate senses of the tweets are retrieved through the lexical match between the tweets and the news paper archive. As the information is stored in the newspaper archive categorically, this structured data source has been used for sense retrieval. But in this phase, there were insufficient number of lexical matches between the statements which could not drive the system towards the right direction in all the cases. So, it was necessary to come out of the lexical boundaries of the statements.

For example, suppose there are *n* numbers of tweets, say- $T_{1...n}$ and the news paper archive contains the sense domains $D_{1...k}$. Now, the sense (S_i) of a particular tweet T_i is calculated from the following formula:

$S_i := Max \{ContentWords_{Ti} \cap ContentWords_{D1...Dk}\};$ *where 1<=i<=n.*

The sense of the tweet T_i is evaluated based on the maximum number of overlap between the *Content Words* (noun, verb, adjective and adverb) of the both- tweet and the different sense domains of the archive.

In the second phase, contexts of the tweets are expanded through synset analysis of the content words in it. These synsets are retrieved from the WordNet. As the contexts of the tweets are expanded, a semantic relation has been established between the statements beyond its lexical level information. The resulting semantic relatedness drives the system towards the right direction for sense retrieval.

For example, if the content words of a tweet T_i is $ContentWords_{Ti}$ and the total synsets of those content words is $Synsets_{Ti},$ then the total Meaningful Words (*MW*) are $ContentWords_{Ti}$ *U* $Synsets_{Ti}$ and in the same way the total Meaningful Words (*MW*) of a domain D_j in the archive is $ContentWords_{Dj}$ *U* $Synsets_{Dj},$ *where, 1<=j<=k.* Now, the sense (S_i) of a particular tweet T_i is calculated from the following formula:

$S_i := Max \{MW_{Ti} \cap MW_{D1...Dk}\};$ *where 1<=i<=n.*

The sense of the tweet T_i is evaluated based on the maximum number of overlap between the Meaningful Words of the both- tweet and the different sense domains of the archive.

Algorithm: Sense_Retrieval_from_Regular_Tweet_Record

Input: Tweets and the News Paper Archive.
Output: Senses of the tweets.

Step 1: Tweets are collected from the tweeter portal of different users.
Step 2: Texts are normalized and lemmatized.
Step 3: News paper archive is normalized and lemmatized.
Step 4: Lexical overlap is calculated between a tweet and the different domains of the news paper archive.
Step 5: Maximum overlap with a domain of the archive represents the sense of the tweet.
Step 6: Stop.

Algorithm: Sense_Retrieval_from_Context_Expanded_Tweet_Record

Input: Tweets and the News Paper Archive.
Output: Senses of the tweets.

Step 1: Tweets are collected from the tweeter portal.
Step 2: Texts are normalized and lemmatized.
Step 3: Synonyms of the Nouns, Verbs, Adjectives and Adverbs of the tweets are retrieved from the WordNet.
Step 4: News paper archive is normalized and lemmatized.
Step 5: Lexical overlap is calculated between a context expanded tweet and the different domains of the news paper archive.
Step 6: Maximum overlap with a subject domain of the archive represents the sense of a tweet.
Step 7: Stop.

V. Result and Corresponding Evaluation

In the execution phase, first of all, tweets have been collected from the tweeter portal. For the purpose experiment, these tweets have been collected from the tweeter handles of some renowned persons, like Sachin Tendulkar, Raghuram Rajan, Amitabh Bachchan, Rahul Gandhi etc. For validation purpose, the test set was initially tagged with its innate senses. The system generated answers were compared with these tagged answers to evaluate the performance of the system.

Table-1 and Table-2 represent the performance of the system on a regular data and its corresponding context expanded form.

Table I. Performance Of The System On A Regular Data Set

Domain	Number of tweet	Correctly Resolved tweet	% of Accuracy
Cricket	250	107	42.8
Business	250	134	53.6
Film	250	105	42
Political	250	141	56.4
Total	1000	487	48.7

TABLE II. Performance of the system on context expanded data set

Domain	Number of tweet	Correctly Resolved tweet	% of accuracy
Cricket	250	210	84
Business	250	231	92.4
Film	250	215	86
Political	250	225	90
Total	1000	881	88.1

It is clear from Table-1 and Table-2 that due to context expansion strategy, accuracy of the result has been increased.

It is closely observed that performance of the system was comparatively weaker in case of cricket and film related tweets than the other two. The reason behind this scenario is- there are lots of common factors between these two domains. So many common words were found during the evaluation. And, the overall pitfalls of the system were mainly due to the unique and unstructured tweets of different persons. These factors are discussed in section 6.

VI. Conclusion and Future Work

The experiment is carried out to retrieve the innate senses of the tweets posted by the users. Although, the data set was prepared from the structured tweets, in real life scenario- handling the versatility of tweets in computational environment requires a separate experiment. For example, in few cases the tweets are-

Multi-lingual (like: Ab likho kuch acha @PinkvillaTelly itna acha koi celebrity initiative karta for awareness, for country, befaltu ki news likh sakte ho par, yeh likho ab, dalo article. Mohsin Khan doing such great job. He's a gem. @iwmbuzz @PinkvillaTelly @bollywood_life @tellychakkar),

Without any grammatical structure (like: But sid fans trending for sid& it's now on news portal. I am not happy seeing that sana has no single trend today cause it's her first bollywood project also..if we trend #BhulaDungawithSana,will show our ego..but we can definitely trend a positive one by welcoming her in BW),

In transcripted form,

Full of emoticons,

Full of proper nouns, concatenated with first letter in capital, '@', '#' etc. (like: Are you excited to watch #AsimRiaz and his lady love #HimanshiKhurana together in a song with #NehaKakkar today? Comment below and tell us! #AsimRiaz #HimanshiKhurana #AsimManshi #AsiManshiDebut #Himanshiasim #AsimDebut #NehaKakkar),

Phrase-idioms (like: In the land of the blind, the one-eyed man is king.) etc.

A dedicated experiment should be carried out to deal with these situations.

References

[1] Z. Lu, S. Jialin Pan, Y. Li, J. Jiang, Q. Yang, "Collaborative Evolution for User Profiling in Recommender Systems," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), pp.- 3804-3810.

[2] O. Hasan, B. Habegger, L. Brunie, N. Bennani and E. Damiani, "A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case," in Proceedings of IEEE International Congress on Big Data, Santa Clara, CA, 2013, pp. 25-30.

[3] G. Miha, M. Dunja and G. Marko, "User profiling for the web," in Proceedings of Comput. Sci. Inf. Syst., volume 3, 2006, 01, pp. 1-29.

[4] S. Silvia and A. Analía, "Intelligent User Profiling," In: Proceedings of Artificial Intelligence, volume 5640, January, 2009, pp. 193-216.

[5] F. Aleksnadr, A. Mohammad, S. Ivan and C. Tat-Seng, "360° user profiling: past, future, and applications," in Proceedings of ACM SIGWEB Newsletter, 07, 2016, pp. 1-11.

[6] R. Behnam and A. Tahereh, "Event-based User Profiling in Social Media Using Data Mining Approaches," PhD Thesis, 2017.

[7] D. Ian, R. Dave, B. Dave, C. Steve and V. Poorvi, "User Profiling with Privacy: A Framework for Adaptive Information Agents," in Proceedings of Lecture Notes in Artificial Intelligence, 2003, pp. 123-151.

[8] J. Tang, L. Yao, D. Zhang and J. Zhang, "A Combination Approach to Web User Profiling," in Proceedings of ACM Transactions on Knowledge Discovery from Data, March 2010, pp. 1–38.