Improved Information Filtering and Feature Dimensionality Reduction using Semantic based Feature Dataset for Text Classification: In Context to Social Network

Himanshu Suyal
Research scholar
Department of computer science
GBPEC, Pauri

RB Patel, Ph.D
Associate Professor
Department of computer science
GBPEC, Pauri

ABSTRACT

In Micro-blogging web services such as Twitter, the user is often bombarded with tons of information and raw data, with user unable to classify it into right category. The solution to overcome this problem can be derived from automatic text classification process. Social networking websites often limit their users to put up a short text message of length 140 characters only. Hence classifying this raw data continuously on these microblogging websites is a tedious task, as one has to deal with short text. Short text messages are difficult to classify as they have lack of semantic information and they have high risk of getting misclassified. In this research paper, a methodology has been developed that incorporates preparation of semantic database and then employ it to extract the necessary classification features from the database. This prepared database is then used for binary feature extraction from the set of user tweeted database hence the process of extracting features from the available database based on the semantic database approach has been presented. The basic of this paper is mainly focused on extracting nine features and then reducing the features to seven features using logical operations. The process of reducing the features not only reduces the complexity of the written code but also saves the database memory required to save the extracted feature for master training database. The features so extracted are easier to use and operation has less complexity of generation than compared to features generated by other available algorithms like Bag-of-Words.

Keyword

Text classification, short text, Twitter, semantic, Bag-of-Words

1. INTRODUCTION

Short text plays very important role in very applications like web application and social networking. With the growth of the social networking site, number of short text is increasing. Some popular social network like twitter [1] restricts the message length up to 140 characters. Compared to the long text, short text contains less semantic information, so applying some traditional exiting classification algorithm causes very poor result [2].

Every classification procedure involves preparation of a manually labeled dataset, involving labelling of unlabeled datasets. These labels are nothing but classification of that particular entity to a certain predefined class by heuristic knowledge. Based on which class from the already present list of classes, the particular entity would lay in, defines the type of classification [3]. If there are only two classes present and the entity has to be classified into either of these two classes then the classification scheme is known as Binary classification. If the entity has to be classified in to one of the classes from the list of classes that contains more than two class as entry, then the classification process is known as multiclass classification problem. To classify any particular entity to a certain class, the features describing that particular entity should closely match to the features that belong to the class in which we want to put the entity [4].

The Feature extraction process for short text classification is quiet complex. It becomes more complex in the world of social blogging as the users often try to use synonyms and internet slangs to write a sentence and describe their emotion or state of emergency. For example “as soon as possible” is often referred as ASAP”, laughing out loud is often referred as LOL [5]. In altogether there are more than 2000 internet slangs and abbreviation presents that are used by the users to express their emotions and views and opinions within 140 characters limit.

In this research work we will present a methodology to extract different features from the entities and then we will use these extracted features to train the machine [15]. The, so trained machine will then be used to classify different sentences on micro blogging websites or on social networking websites to fall into certain predefined classes. It will be taken care that the machine is not so finely tuned that its performance degrades for the unseen sentences [16].

Rest paper is organized as follows: Section II presents a detailed description of the previous work done on text clustering. Section III present the detail description of the adopted methodology. Only feature extraction techniques will be discussed here, the features so extracted from the developed methodology can then be used for classification using SVM if it is a two class problem or it can be extended to be utilized with neural networks if it is a multi-class classification problem. Section IV shows the results of feature extraction and section V have a discussion on conclusion and future work.

2. RELATED WORK

Short texts have lack of contextual information so traditional text clustering algorithms have some limitation to classify the text, so to overcome this problem in existing work ,the original short text is enriched by adding some additional information. One way to achieve this is by employing the search engine and utilizing the search results to expand the contextual information of short text [6] [7] [8] and the other way to use the external repository like Wikipedia and open directory, etc. as a background knowledge [9] [10]. Sankaranarayanan et al [11] introduced a new TweetStand to classify tweets as news or non-news. Topic modeling [12] [13] is also widely used to classify the short text. The main idea of these kind of model is to find the topic form the domain related datasets, and assuming that each text is multinomial distribution over these topic. Liliyang et.al [14] introduced a new methodology of classification in which a topic modeling approach is introduced which uses the expected cross entropy to measure the discriminative capacity of words in short text.

3. PURPOSE METHODOLOGY

The complete process of feature extraction is shown in figure number 1, the tweets are downloaded from different twitter accounts and then are stored in a central repository. Then the tweets are pre-processed to remove the presence of any special character and stop words that can lead to classification in inaccuracy and also resulting in bad features. After the pre-processing has been done the tweets are now converted into tokens. After the tokens are obtained the N-Feature algorithm operates in the token to extract the features present in each tweet.

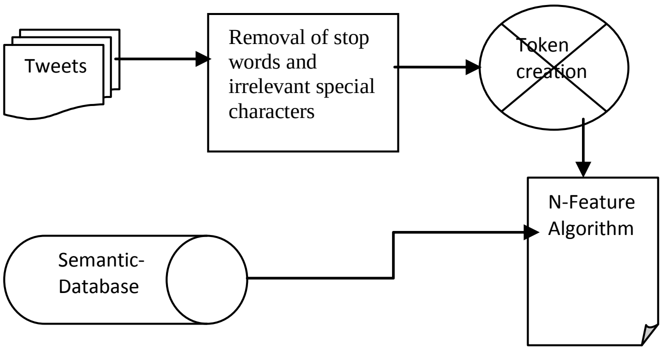


Figure 1: Feature extraction process using N-Feature Algorithm

A. The 'N' Feature Algorithm

The developed methodology adopts a very simple yet effective approach of extraction of features from the entities. By entities author of this article always refer to single sentence picked up from the micro blogging website(s) or from any other social networking website(s), where the data is so overloaded that an efficient classification scheme is needed, so that sentence(s) or micro blog(s) can be clubbed into the category they are likely to belong.

For feature extraction algorithm, the user developed semantic based database. The concept behind the development of a semantic based database was that, it will reduce the classification errors which means a sentence which is more likely to be classified as Class Y will have a very less probability of getting classified as a class X. For this study a 9 feature methodology is selected. The selection of this nine features is based on the detailed study of microblogging website(s) and manually analyzing at least 3000 micro blogs or Tweets. Before selection of these features, classes were defined. They were so defined as to classify the upcoming tweets into 4 categories namely 'Personal Message', 'Deals', 'News', 'Events'. These four classes were only selected because manual analysis of the micro blogs yielded the result that in daily life of a micro blog user he or she restricts oneself to either Personal Message- Exchanging information with his or her friend, depicting joyous emotions etc. Events were selected because they contain information about the most happening events in day to day life. The rest of the two categories were selected because an average person spends almost 45 minutes of his or her total time of day searching for right kind of news or the type of new one wants to read. While the class Deal was selected as it would contain information about daily shopping offers, huge sales, fall or rise of stock market or how one should plan his or her investment policies.

For classifying entities to one of these classes, nine features were selected, these features are binary features with either setting the corresponding entry into the feature matrix with either 1, showing its presence in the sentence or 0 showing its absence in the entity.

The list of features that were adopted are listed below: -

1. Abbreviation present in the sentences: - Authentic Reporting Houses never use abbreviation to convey any information on micro-blogging website(s), such features are often encountered only in personal message or event information.
2. Time-Related Information Event describing words: - The database was so prepared that only describes the events. The words present in the database related to this feature was contextual only to the category event hence almost eliminating the ambiguity of meaning of words present in the micro blogs.
3. Opinion describing words: - A database of words describing opinions or personal emotions was created.
4. Presence of Personal Reference: - A message meant for an individual often contains '@username'. Username is the user id of that particular individually and is always unique and is not to be confused with the on screen name of that blogger.
5. Presence of Currency, deal related information: - A semantic database that refer only to deals and offers was prepared.
6. Presence of Emphasis: - Users often repeat certain character of a word to show there aggression or opinion towards a particular incident or motion. For example "Very Good" often used for appreciation.
7. Presence of date and time information: - An algorithm was developed to detect presence of time and date related information, presence of time related info presence of Am or Pm information etc.
8. Apart from these features two more features were extracted, the features described the presence of time information i.e. before mid-day or after mid-day, date information present in dd/mm/yy or dd-mm-yy format. The features represent event related information in the text.

B. Searching for the features

After the database and the number of classes are set and well defined, the feature extraction process starts using search algorithm. But before the search algorithm can be started, the entities should be free from the stop words as stop words consume considerable amount of computation time when kept in the entity and often leads to classification errors that are beyond experimental limits. After the stop words are removed from the entities, the special characters like parenthesis, ampersand sign are removed which saves computation time. After the data has been pre-processed and is clean to be used for feature extraction, the process of feature extraction starts. Table 1 shows some of the part of the prepared database for each of the features.

Table 1: Snap view of semantic database prepared

ABD	Already been done
ABT	About
ABT2	Meaning 'About to'
ABTA	Meaning Good-bye (signoff)

(A) ABBREAVTION

bargain
bond
cartel
charter
codicil

(B) Deals

birthday
reveal
reveals
games

(C) EVENTS

C. Reducing the features

There were total of nine features extracted from every text, to store nine features for a set of 3000 and above text messages which is a memory consuming task .It increase the computation time, time to extract feature and accordingly train the classification algorithm to identify the incoming text messages according to the training master dataset. To reduce the features to only seven, a logical operation was performed to reduce three features to a single feature thus reducing the total features used for classification as seven. Since all the features are logical in nature and represented as one and zero, any logical operation can be easily performed. The figure 2 shows the mathematical logical operation developed to reduce the feature. A logical OR operation is performed on the features, that represent time related information and date related information for dd-mm-yy format or dd/mm/yy format (formatting system widely followed in Indian reference).

Reducing the features always reduces the computation time, thus allowing for a faster calculation. The features were reduced to one because they describe event related information that is specific to duration of the day or to a specific day, and is related to time/event only hence can be used to extract a single bit by applying proper logical information.

Output Feature reduced to single bit

Features
Top to down before mid-day, after mid-day and date format

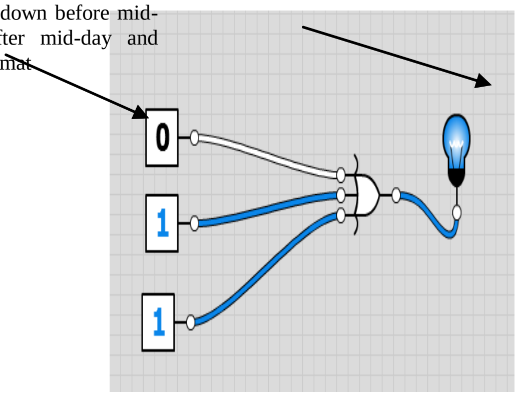
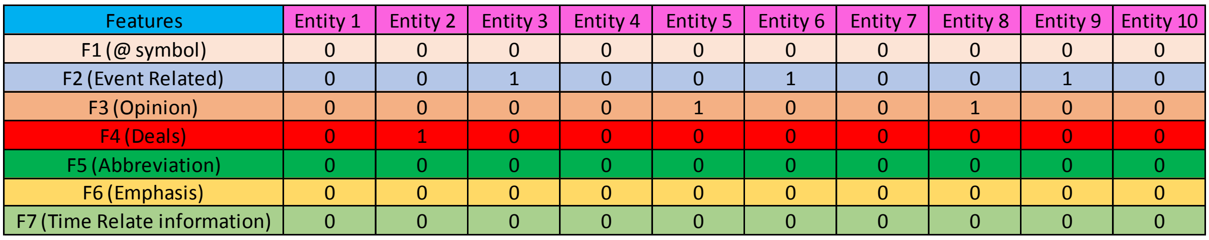


Figure 2: Logical Operation to reduce features in master dataset

4. RESULT

Table 2. Shows the feature vector generated using the so developed algorithm.

Table 2: Extracted feature vector, extraction executed using developed algorithm for News Category

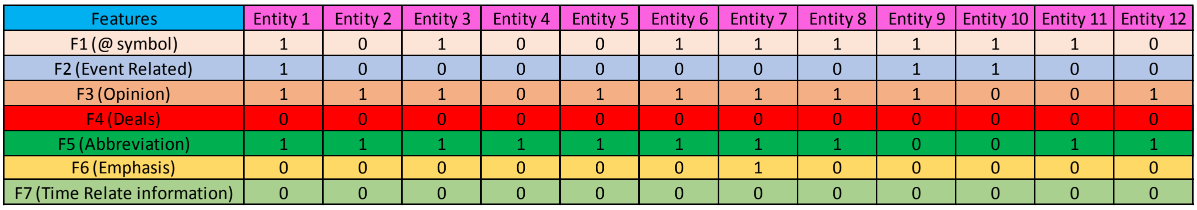


Features	Entity 1	Entity 2	Entity 3	Entity 4	Entity 5	Entity 6	Entity 7	Entity 8	Entity 9	Entity 10
F1 (@ symbol)	0	0	0	0	0	0	0	0	0	0
F2 (Event Related)	0	0	1	0	0	1	0	0	1	0
F3 (Opinion)	0	0	0	0	1	0	0	1	0	0
F4 (Deals)	0	1	0	0	0	0	0	0	0	0
F5 (Abbreviation)	0	0	0	0	0	0	0	0	0	0
F6 (Emphasis)	0	0	0	0	0	0	0	0	0	0
F7 (Time Relate information)	0	0	0	0	0	0	0	0	0	0

The process was executed for 1500 entities all belonging to either of the four class.

Table 3. Shows the feature vector generated using developed algorithm for Personal Message category

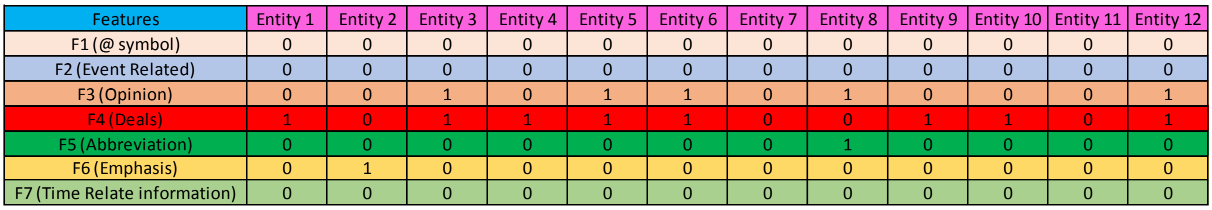
Table 3: Extracted feature vector for Personal Message Category



Features	Entity 1	Entity 2	Entity 3	Entity 4	Entity 5	Entity 6	Entity 7	Entity 8	Entity 9	Entity 10	Entity 11	Entity 12
F1 (@ symbol)	1	0	1	0	0	1	1	1	1	1	1	0
F2 (Event Related)	1	0	0	0	0	0	0	0	1	1	0	0
F3 (Opinion)	1	1	1	0	1	1	1	1	1	0	0	1
F4 (Deals)	0	0	0	0	0	0	0	0	0	0	0	0
F5 (Abbreviation)	1	1	1	1	1	1	1	1	0	0	1	1
F6 (Emphasis)	0	0	0	0	0	0	1	0	0	0	0	0
F7 (Time Relate information)	0	0	0	0	0	0	0	0	0	0	0	0

Table 4 shows the feature vector generated using developed algorithm for Deals category.

Table 4: Extracted feature vector for Deals Category

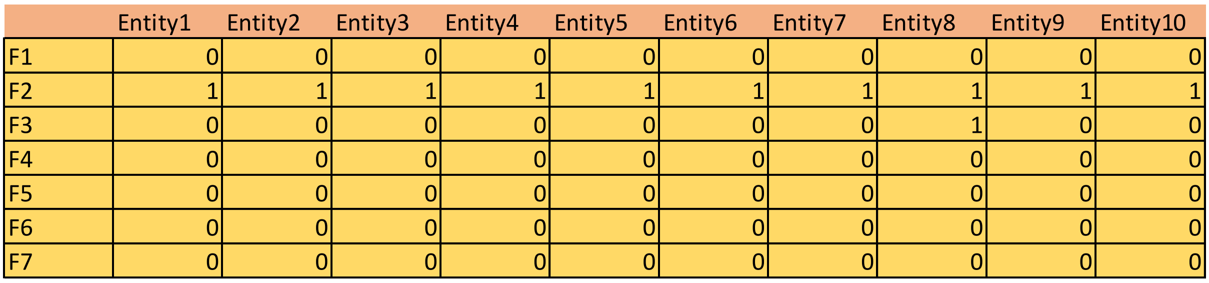


Features	Entity 1	Entity 2	Entity 3	Entity 4	Entity 5	Entity 6	Entity 7	Entity 8	Entity 9	Entity 10	Entity 11	Entity 12
F1 (@ symbol)	0	0	0	0	0	0	0	0	0	0	0	0
F2 (Event Related)	0	0	0	0	0	0	0	0	0	0	0	0
F3 (Opinion)	0	0	1	0	1	1	0	1	0	0	0	1
F4 (Deals)	1	0	1	1	1	1	0	0	1	1	0	1
F5 (Abbreviation)	0	0	0	0	0	0	0	1	0	0	0	0
F6 (Emphasis)	0	1	0	0	0	0	0	0	0	0	0	0
F7 (Time Relate information)	0	0	0	0	0	0	0	0	0	0	0	0

Table 5 shows the feature vector generated using developed algorithm for Event category

Table 5: Extracted feature vector for event category

Color changed for illustration purpose only.



	Entity1	Entity2	Entity3	Entity4	Entity5	Entity6	Entity7	Entity8	Entity9	Entity10
F1	0	0	0	0	0	0	0	0	0	0
F2	1	1	1	1	1	1	1	1	1	1
F3	0	0	0	0	0	0	0	1	0	0
F4	0	0	0	0	0	0	0	0	0	0
F5	0	0	0	0	0	0	0	0	0	0
F6	0	0	0	0	0	0	0	0	0	0
F7	0	0	0	0	0	0	0	0	0	0

Fro the above tables it is evident that the features are so easy to be extracted, since these features are binary in nature they are prefect to be used with any machine learning algorithm, thus reducing computational complexity considerably. These features can now be used with Support Vector Machine or Neural Network Pattern Recognition algorithm.

Since the comparison database is prepared keeping in mind to deal with semantics also, the probability of misclassification using the above obtained features are minimum. The preparation of semantic database deals with the interaction of different words and relation between them.

5. CONCLUSION AND FUTURE WORK

The research will be productive for classification of short text messages that now a days are flooding every micro blogging website. The feature extraction algorithm is so robust and flexible that any new feature can be easily added or omitted to feature vector space as per the need of the user. The same algorithm can then be easily extended to classify the text in to spam or a threat message. As the extracted features are binary in nature they are easily manageable and are more effective when dealing with neural networks using binary functions. In further research work, a machine will be trained using this classification system or testing its accuracy and efficiency. or testing its accuracy and efficiency.

6. REFERNCES

[1] www.twitter.com

[2] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics

from large-scale data collections. In Proceeding of the 17th international conference on World Wide Web, WWW '08, pages 91{100. ACM, 2008.

[3]N.Cohen.Twitteronthebarricades:Sixlesson,learned.http://www.nytimes.com/2009/06/21/weekinreview/21cohenwb.html, Pub. June 20, 2009

[4]http://www.time.com/time/magazine/article/0, 9171, 1044658, 00.html

[5] A. Java X. Song, T. Finin, and B. Tseng, 2007. Why we twitter: understanding microblogging usage and communities. In Process WebKDD/SNA-KDD ‘07 (San Jose, California, August, 2007), 56-65.

[6] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceeding of the 17th international conference on World Wide Web, WWW '08, pages 91-100. ACM, 2008.

[7] Mehran Sahami , Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. Proceedings of the 15th international conference on World Wide Web, 2006.

[8] D Bollegala, Y Matsuo, M Ishizuka. Measuring semantic similarity between words using web search engines. Proceedings of the 16th international conference on World Wide Web, 2007.

[9] Ou Jin, Nathan N. Liu, Kai Zhao , Yong Yu , Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. Proceedings of the 20th ACM international conference on Information and knowledge management, 2011.

[10] Mengen Chen, Xiaoming Jin, Dou Shen. Short text classification improved by learning multi-granularity topics. Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, p.1776-1781, 2011.

[11] Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, and M. D.,Sperling, J. TwitterStand: news in tweets. In Proc. ACM GIS’09(Seattle, Washington, Nov. 2009), 42-51.

[12] Yue Lu, Qiaozhu Mei , Chengxiang Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA, Information Retrieval, v.14 n.2, p.178-203, 2011.

[13] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Volume 1, p. 536C544. 2012.

[14] Yang, Lili, et al. "Combining Lexical and Semantic Features for Short Text Classification." Procedia Computer Science 22 (2013): 78-86.

[15] M.Milian. Twitter sees earth shaking activity during So Caquake. http://latimesblogs.latimes.com/technology/2008-07/twitter-earthqu.html,Pub. July 30, 2008

[16] http://en.wikipedia.orgwiki/Micro-blogging