https://doi.org/10.26438/ijsrcse/v6i6.111

International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.6, Issue.6, pp.1-11, December (2018) **E-ISSN:** 2320-7639

An Enlarged and efficient Hash-tagger++ Framework for News Stream in Social Tagging issues

M.Vidhya Lakshmi[1*], P.Radha[2]

[1] Department of Computer Science, Government Arts College, Bharathiyar University, Coimbatore, Tamil Nadu
[2] Department of Computer Science, Government Arts College, Bharathiyar University, Coimbatore, Tamil Nadu

Available online at: www.isroset.org

Received: 07/Dec/2018, Accepted: 22/Dec/2018, Online: 31/Dec/2018

Abstract- In the fashionable notion of Hash-tagger can tag the social media news. The advance mechanism of hash tagger is one kind of metadata tagging community. The social media's one of the micro-blogging of the site – Twitter. The twitter is designed and organized news, stories, group debates and more information's are via tweets. These functionaries have easily connected the twitter crowds and scattered news from that user in media. If the hash-tagger++ can be applied in twitter subsequent to achieve the effectiveness in hash-tag recommendation and the classification. In this amicable part have easily espoused other hash-taggers namely, Multi-Class Hash-tag Classifier _ Support Vector Machine (MCHC_SVM) algorithm and semantic tagger. In this tagging scenario expeditiously classifies the tweets pedestal on its hash-tag. The tagger of semantic means it can detect the similar tweets and recommending the tags also. The proposed system is to can be utilized and abridged the High D of the feature space. The focal goal of this proposed system is easily ordering the twitter crowds, to improving the hash-tag recommendation and achieve high scalability with efficient performance, obviously.

Keywords- Hash Tag Recommendation, Data mining Techniques, Framework for hashtag, MCHC-SVM Algorithm.

I. INTRODUCTION

Twitter is on online news and fast social networking site that relies on micro-blogging for communication. Twitter is to give everyone the capacity to create and share ideas and information instantly, without barriers. It classified among the ten most call on website. The news environment and news consumption practices are changing rapidly. Society is moving from a traditional news cycle dominated by journalism professionals to a more complex information cycle that incorporates ordinary people within the process. Established news media organizations still produce most of the news, including that which circulates through social media and aggregators. Twitter enables its members to post or send short messages called tweets, whereby users can broadcast what they are doing or thinking to the world, to closed list groups or to other individual Twitterers. Many resources note that estimates of the number of tweets and page views over Twitter may be understated, as they may omit users accessing it through third-party clients from their desktop or mobile devices. Twitter gives its users chance to either re-tweet, favorite or reply any initial tweet and the occasional trending topic which will enable an organization know what the consumers are buying at that particular period. It is directly communicate between an individual and any organization. Organizations have the chance to bought ads on twitter, buying twitter ads is very different from newspaper ads. In newspaper are buying a square on a piece of paper which may or may not be read or entertained by thousands of readers but when purchase a twitter ad are actually buying space on an individual's timeline. Twitter made this feature even more unique by allowing advertisers to send ads only to people who mention specific keywords that may relate to their product or organization in their timelines. This feature has authorized organizations to not only select their target individuals but to also have an idea on the amount of prospects available.

The creation of the hash tag this enables other gathering to tweet about the event putting the hash tag sign in the tweet, anybody who views the hash tag will be clever to review others sight with the same hash tag. Hashtag helps journalists find humanity tweeting about topics they are housing. They also help people who are focused in the topics cover find tweets. A hashtag is the symbol, followed promptly, with no space, by a word or phrase. In tweets, the hashtag becomes a highlighted the word can click to go to a search of recent tweets using the hashtag. A news story often consider elements related events, which happen in different time periods and may as well involve different entities.

It assume to have access to a gathering of news articles gloss with social tags extracted in real-time from social media platforms such as Twitter. This approach takes advantage of crowd sourced content as a form of real-time, continuous tagging of news. Twitter has developed into a strong communication and information passing tool used by millions of people around the world to post what is happening now. A hashtag is a keyword, in feature in Twitter to organize tweets and facilitate effective search among a massive volume of data. In this proposed system is an automatic hashtag recommendation system that helps users find new hashtags related to their interests on-demand.

In this system propose a real-time hashtag recommendation approach that is able to efficiently and effectively capture the strong evolution of news and hashtags. Most previous attitude for hashtag recommendation works on static datasets and do not account for the emergence and disappearance of hashtags.

In this proposed paper, Section I contains the introduction of Hash-tagger++ Framework for News Stream in Social Tagging issues Section II contain the related work of Hash-tagger++ Framework for News Stream in Social Tagging issues, Section III explain the Hash-tagger++ Framework for News Stream in Social Tagging issues methodology with flow chart, Section IV describes results and discussion Hash-tagger++ Framework for News Stream in Social Tagging issues, and Section V concludes research work with future directions.

II. RELATED WORK

The previous work discusses the hashtag recommendation for tweets are based on the Multi-Class Classification (MCC) modeling on the static datasets. In this paper **[1]** has presented the concept namely, 'Twitter Hash Tag Recommendation'. The social media is divided into number of micro-blogging services. The twitter is one of the micro-blogging services. Generally the social media applications are used the people in every day, then the social media contents are mostly shared and shares the downloaded files, and more; so every day increased the data in twitter. Twitter is the general term mostly discussed the political awareness and top most news and important news discussed via textual contents. The hashtag is provides the users with tagging mechanism or top ranked news. In this tagging mechanism is too organized, grouped and creates the visibility for their posts. This concept is gives simple and easy methods but it solves the increased infrequent usage problem in twitter. In this system is introducing various methods of the recommending hashtags, which means creates the new posts then to motivate more wide spread adoption and usage. Although, the hashtag recommendation is comes within the website with the following challenges are, (i) processing the huge amount of or volume of streaming data, (ii) the content are hash-tagged with small-l and noisy. So, the proposed system is uses some methods to avoid that problem. The preprocessing methods are reduces the noise in the given data and it determines the effective methods of the hashtag recommendation. It is based on the popular classification algorithms, because the classification logic is splits the twitter news. In this system processes are classified into two main categories are, (i) preprocessing, (ii) Classification. The preprocessing means prevention activities. In this term mostly using the big data related applications. Because it provides the following basic terms like, filtered content, aggregates and actually processing the dataset making with accurate and efficient. To minimizes the noise in the given content and preparing the data for the accurate classification. Secondly, follows the approaches like naïve bayes, K-nearest neighbor, shared the various classifier functions and hybrid classifier.

Paper **[2]** has presents the 'Suggesting Hashtags on Twitter'. In this concepts solves the problems are how to categorize the posts effectively and posts searching. This problem is reduced by using the simple terms for categorize the user posts. The system facilitates the user may categorizes their posts to help the hashtags and any words or phrases or keywords may be used as the category. If suppose search the postings for facebook, then the user wants to try many different types of hashtags such as, #Facebook, #FB, #Facebook.com, or #Zuckerberg, and etc., it proposes the hashtag implementation and calculates a tool for suggesting related hashtags to a user, given a tweet. This system is initial analyses is suggests the dataset. In the system is to finds out the exact informative distributions of the words or phrases for many different hashtags. That facilitates the naïve bayes model for the hashtag recommendation given a required hashtag or query post. The above systems are used the naïve bayes or Support Vector Machine (SVM) classifier for hashtags. The hashtag is looks a class and then, the tweets are tagged with the hashtag and assumes labeled data for the class. The hashtag recommendation system is used for tweets that can be adapted to recommendation for the news by treating the headline of the post. Then, the MCC approaches are increased by the data scale, sparsity and the noisy tweets.

The Paper **[3]** has presents the concept is 'News-Topic Oriented Hashtag Recommendation in Twitter Based on Characteristic Co-occurrence Word Detection'. In previous system s is to categorize the messages and using the form of conversation for the twitter topics. But it is difficult for the users, because the user hashtags the sharing messages that is form of opinions or interests or comments for their interesting topics. The system introduces the approach for the recommended news topics related hashtags to help the twitter users. The users easy to join the conversion about twitter news topics. Firstly it focus the topic, the topic is a keyword. The topic is a specific information and includes the co-occurred words with a given target word, then it uses the co-occurrence words characteristics it takes from the news articles and to form a vector for news topic representation. The hashtag vector creation is based on the tweets with the same hashtag. Then calculates the similarity between the above two vectors and recommended hashtags of the high similarity scores with the news topics in twitter. Finally the system result gives the recommended hashtags for news which includes highly relevant news topics and helping the tweeter users are sharing their tweets with others in twitter.

In this paper **[4]** has presents the topic is 'Learning-to-Rank for Real-Time High- Precision Hashtag Recommendation for Streaming News'. This system is focuses the high-precision topic are hash-tagged. Its other purpose is for used the large scale topic classification. It classifies set of topics and it includes the huge and highly dynamic topics. But its main theme is based on many applications and provides the information is selects the twitter communities and promoting the original content, index the social news and enable the better retrieval, story tracking and summarization. The learning-to-rank method is modeling the hashtag relevance and presents the methods then, extracts the time-aware features from the highly dynamic content. It uses the data collection then, process the pipeline processing and continuously apply the methodology for gets the low latency, so finally applies the high precision recommendations.

III. PROPOSED SYSTEM

The proposed work enhances the existing approaches and analyzed the advantages and potential of context, L2R approaches and statistical approaches. The proposed system develops a new technique named as hashtagger++. To achieve the effectiveness in hash tag recommendation and classification, the hashtagger++ technique adopts MCHC_SVM [multi class hash Tag classifier _support vector machine] algorithm and semantic tagger. The MCHC_SVM classifies tweets based on its hashtag. The semantic tagger is for detecting similar tweets and recommending tags.

The proposed classification technique extracts categories and hashtag information to generate a compact set of keywords as topic for the tweets. It calculates similarity scores. The similarity score calculation is a measure of tweets affinity towards a category from the training dataset. This assigns the tweets to each category using positive term of keywords and Term Frequencies of tokens. The tweet dataset is labeled to the category for which it obtains highest degree of dependency. Further the algorithm enriches each category hashtag keyword list by choosing semantically related tokens from well classified dataset.

Contributions:

The proposed Tweet dataset Classification and recommender system is named as "Hashtagger++ " because it uses two sources of existing knowledge, and semantic hashtagger determine the semantic content of tweets and map them into a category, which helps to identify the hashtag. Figure 3.1 depicts the overall architecture of the Hashtagger++ based recommender scheme, which is proposed in this chapter. The main contributions of the scheme are highlighted below.

A novel approach becomes necessary in the areas of tweet dataset classification which provides higher efficiency and accuracy. The research study deals primarily with five proposed approaches for tweet dataset classification and management. They are:

- ➢ Effective use of classification algorithms which is created with multi class categorization to find appropriate hashtags from the tweets;
- ➢ Integration of both hashtag and semantic feature of those hashtag pattern facets together rather than using them in two separated stages.
- ➢ So proposed system is hybrid architecture, which is named as Hashtagger++.
- ➢ Hashtagger++ contains the MCHC_SVM pattern mining algorithm for fast and reliable feature extraction
- ➢ Semantic hashtagger with feature priority algorithm for hashtag ranking. This helps to find the effective hashtag in the tweet dataset.
- ➢ Improved hashtag recommendation model: this thesis improves the state-of-the-art hashtag recommendation model which is proposed in the earlier studies by proposing a new approach for data collection and facet computation.

The proposed study introduces a set of approaches for hashtag extraction from twitter and evaluates the end-to-end effect on the recommendation precision, coverage and running time. The proposed work examines three cold-start strategies to enhance the effectiveness and coverage of recommendation for new tweets by its hashtag, by bootstrapping the facet computation using data collected for older stories from the dataset.

The proposed work provides an extensive study of different approaches for hashtag classification and hashtag recommendation, and show that proposed system is more effective than the existing techniques. The results also show that relevance modeling allows us to deliver recommendations to popular as well as less popular tweets, providing high coverage and precision even with small amounts of available data, a setting where most prior models do not perform well. The classification efficiency in terms of time and accuracy is generated.

The improvements achieved in those performance measures have been tested for statistical significance using different functionalities.

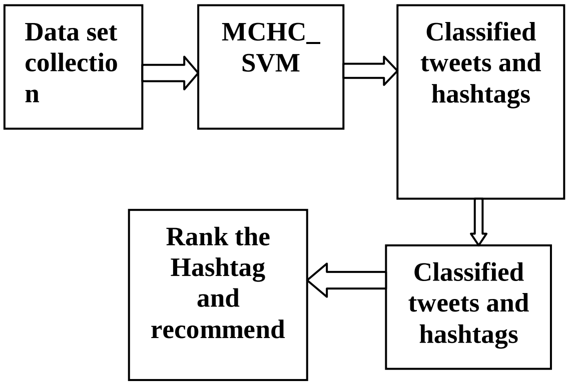
Data set collectio n
MCHC_ SVM
Classified tweets and hashtags
Classified tweets and hashtags
Rank the Hashtag and recommend

Figure 3.1 tweet dataset classification and Recommendation process

The process of classification is populating each category with similar concept hashtag that define the feature space of the category. This chapter shows the well-organized background knowledge in the form of a simple semantic hash tagger, which can improve tweet dataset classification result significantly. Many words have several meanings when used in different contexts. As like, a tweet can be fall under multiple category as well as it contain multiple hashtags. The word which has several meaning is called polysemous words. Semantic process expresses the meaning of each sense of a base word with its Synonym set. For each non-trivial token which is present in a tweet dataset and for each category name, their lexically related hashtag can be found in word all starting with their relevant senses and transitively following different relation types. Lexical cohesion refers to a range of textual cohesion that allows the use of similar meaning words on Synonyms, generalization the concept on Hypernyms, specialized versions of a concept called Hyponyms or constituent parts of an object called Meronyms. Terms that share a common Hypernym are called Coordinate hashtag. The set of lexically related hashtag of a base word are called Lexical Semantics. The concern about generating a suitable list of keywords for a given category can be addressed by extracting all Lexical Semantics of the category from the WordNet.

Positive Term Detection:

The next concern is to extract meaningful words from a tweet dataset. For accomplishing this, the Hashtagger++ uses a facet finding (FF) process, which can find the both positive and negative term and term sets from the set of dataset. The fundamental premise of FF is that if two tokens in a tweet dataset have a common list of Lexical Semantics, then the corresponding facets represent the tweets meaningful intent greatly. Therefore, such tokens can be retained and other tokens can be pruned as their presence is incidental. This reduces the feature space of each tweet dataset.

Topic and Hashtag Detection:

The tokens of the reduced dataset are matched with keywords of each category. Based on the Keyword Strengths and term frequencies of matched tokens, the system finds the topic of the given tweet dataset d. The topic belongs to a tweet dataset to each category is calculated. Finally, the tweet dataset is ascribed to the category with highest similarity value.

Hashtagger++ aided Keyword Enrichment:

With the help of the Reuters dataset, the algorithm enriches category keyword lists by using the tokens of classified dataset. Reuters is a free, open content online repository created through the collaborative effort of a community of users. Reuter's dataset are peppered with n number of topics. It has functions to handle many fundamental tasks in computational linguistics. This hit these powerful facets of UCI to augment more keywords and enhance the feature space of each category. The further chapter gives more detail on the proposal work in subsequent sections.

Framework for Hashtagger++:

This chapter organizes the methodologies and algorithms involved in the proposed system. This thesis focuses on the problem of automatically extracting and classifying twitter dataset based on the weighted positive and negative key terms and based on the sequence, it can detect the exact similarity between the dataset. Finally it finds the hashtags for the tweets and classifies into different categories. The stepwise detailed description of the Hashtagger++ system now follows.

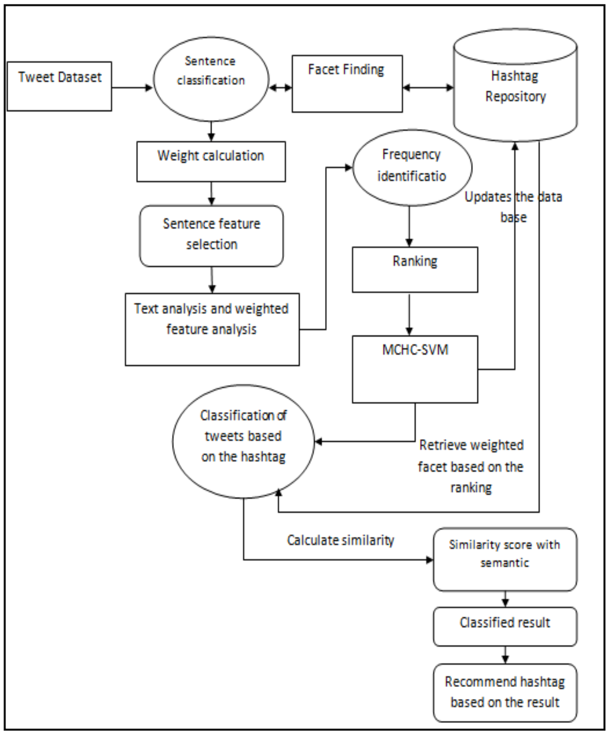
Tweet Dataset
Sentence classification
Facet Finding
Hashtag Repository
Weight calculation
Frequency identificatio
Updates the data base
Sentence feature selection
Ranking
Text analysis and weighted feature analysis
MCHC-SVM
Classification of tweets based on the hashtag
Retrieve weighted facet based on the ranking
Calculate similarity
Similarity score with semantic
Classified result
Recommend hashtag based on the result

Figure 3.2 Hashtagger++ architecture

1) *Training Process:*

This term initiates the algorithm by collecting lexically related hashtags for each category. This input category name to the Hashtagger++ training phase. First synonyms are extracted. Then, similar terms are collected up to the first level only. Next n-grams at the immediate lower level are located. Restricting generalized and specific terms to only single level helps retain the cohesive strength between terms. Sequence terms and term sets are next extracted. At the end of this process, the system is armed with an initial set of keywords **K** that explicitly represent the semantic domain of the category.

2) *Pre-processing tweet dataset:*

The dataset is to be classified need to be preprocessed. Pre-processing includes the following steps. When a new message received in the site, the textual content will be extracted and stored as the dataset, for every data from the dataset, the content separation has been done. For example, if user posts a new tweet, the datasets will be like below.

"More unwanted news posted for Face book, which has had a terrible year for data security.**"**

a) *Replacing all sequences of whitespace characters (tabs, spaces and newline characters) by a single space***:**

This process has been done using simple regular expression concepts. A simple pattern matching functionality can effectively identify these types of characters. To extract the texts and replacing the spaces, newline we use the following, In order to obtain all words that are used in a given input with eliminating tabs and other keywords, this replacing process is required, i.e. a message will split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized keywords representation is then used for further proceedings. The set of all different words have obtained by merging all messages of a dataset.

b) *Stop-word removal:*

After concatenating the words, stop word elimination process will begin. Stop words are a division of natural language. The motivation of the stop-words removal should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words decreases the dimensionality of term space. The most common words in text contents are prepositions, articles and pro-nouns, etc. that does not give the meaning of the dataset. These words are treated as stop words. Stop words are removed from dataset because those words are not measured as keywords in text mining applications. All the stop words, *i.e*. words that appear frequently but do not affect the context are removed from the tweet dataset. Examples of such keywords concludes 'a', 'an', 'and', 'the', 'that', 'it', 'he', 'she' etc. The stop word list contains 1205 words, which do not play any significant role in classification.

c) *Term Weighting:*

For each token *w* in the tweets token set, its number of occurrence is computed. This is called as Term Frequency *TFw*. Term Frequency–Inverse Tweet dataset Frequency (tfidf) is a numerical statistic which reveals that a word is how important to a tweet dataset in a collection. The TF - IDF is often used as an allowance factor in information retrieval and text mining. The value of tf-idf enlarges proportionally to the lot of times a keyword appears in the tweet dataset, but is counteracting by the frequency of the word in the corpus. This can help to control the actuality that some words are generally more common than others. Tf–IDF can be successfully used for stop-words removal this can use filtering in various subject fields including text summarization and classification. Tf–IDF is the product of two statistics which are termed frequency and inverse tweet dataset frequency. To further distinguish them, the number of times each term occurs in each tweet dataset is counted and sums them all together.

3) *Facet finding (FF):*

The pseudocode in **Figure 3.3** explains the process of Facet finding *(FF)*.

```
Algorithm: FF in MCHC-SVM
Input: 1. A set of tokens in a pre-processed article K.
       2. Feature Threshold $T_{FF}$.
Output: A set of semantically related hashtags in new article SK
Steps:
    1. Start the FF process and declare SK=null
        a. SK=0
    2. For each token ($\forall Token\ T$)
        a. $w \in D$
    3. Obtain Lexical Semantic Set Sw from Word dll;
    4. For each token $w \in$ K{
    5. For each other token z¹w, zÎ K
        a. If ($Sw \cap Sz \geq P_{FF}$) then
            i. SK = SK ∪ {w, z}
        b. End if
    6. Delete K
```

Figure 3.3 Facet finding process in MCHC-SVM

By FF, the algorithm extracts strongly related words of the tweet dataset, which are important for classification, which is also known as positive term. It inputs each token w of a tweet dataset to the Word dll to obtain its Lexical Semantics viz., Synonyms and Coordinate hashtag. This is termed as its Lexical-semantics set Sw. Now, starting with first token, the system takes each token w and finds the intersection of Sw with those of all other tokens in the tweet dataset. If the intersection exceeds predefined threshold P_{FD}, the token is retained. Otherwise it is dropped from the tweet dataset. The tokens thus collected from the pre-processed tweet dataset signify the tweets meaning to a greater extent and reduce the dimensionality of tweet dataset representation. Let Nw be the final number of tokens representing the tweet dataset.

Detection of Strong patterns (Semantic Tagger:

This now constructs a Keyword-Strength matrix for each tweet dataset in the training samples. In this process the number of tuples is equal to the number of tokens in the tweet dataset which have been derived by using FF. The number of instances is equal to the number of categories. The elements of this process denote the weighted Keyword (WK) of a keyword for the corresponding topic. The presence or absence of the token is checked in the keyword lists of the topics. If the token is present in any of the topic- Keyword lists, it is a keyword for that category. Let P(w,Tk) be a Boolean function that defines the presence (=1) or absence (=0) of a token w in the topic Tk. The WK of a matched token (equivalently keyword) w to a given category Tk is obtained by the formula below.

$$\text{WK}_{w,k} = \frac{P\,(w,Tk)}{\sum_{j=1}^{N\ topics} P\,(w,Tk)}$$

3.1 equations for weighted keyword detection

The weighted Keyword (WK) are pre-calculated and kept stored in the Keyword Category process.

4) Similarity detection using MCHC_SVM algorithm:

For each tuples in the FF, similarity *detection* metric *Sk*, that reflects the degree of *similarity* of a test tweet dataset to a given topics and its categories *Sk*, is computed using the frequencies of the tokens in the tweet dataset and their corresponding weighted keyword. This metric, given in equation 3.2, is computed for all the columns, *i.e.* for the predefined topics.

$$\text{S}_k = \frac{\sum_{W=1}^{Nw} \text{TF}_w \text{X KS}_{w,k}}{\sum_{W=1}^{Nw} \text{TF}_w}$$

3.2 equations for semantic keyword detection

5) Topic and Hashtag Detection:

The similarity measures obtained for all the columns for a given tweet dataset are compared. The tweet dataset is classified to that category for which this metric has the maximum value.

Topic (Di) = Wk: k=Sim score max (Sj) next Ntopic

Where Topic(Di) be a function that returns the category of a tweet dataset Di and Ntopic is the total number of topics and hashtag categories.

The effectiveness of hashtag can be elaborated with the help of the fact that initially after using semantic details, the category list for data mining did not have the names of any of the hashtag2(hallowan) techniques. This has been realized from the set of training dataset.

Architecture of Hashtagger++:

This section discusses the methodologies implemented in the Hashtagger++ technique in detail. The same distributed hash tag selection method is used in this proposed work. Initially, hashtag generation using semantic rule is implemented to the database containing a large amount of tweet dataset. This technique generates the hashtag for the given database. With this semantic hashtagger, the next step is the application of Hashtagger++. Hashtagger++ is used for classifying the dataset in the database with the help of semantic hashtagger that is generated by wordnet technique. The combination of semantic hashtagger and MCHC-SVM helps to increase the accuracy of classification.

The MCHC-SVM classifies the tweet dataset according to the previously detected hashtags and its textual facets. A semantic rule formal context shown in Table 3.1 consists of three objects which denote three dataset.

Table 3.1 story matching score

Data set	Hashtag1 (mettoo)	Hashtag2 (hallowan)	Hashtag3 (facebook security)
1	0.75	0.25	0.5
2	1	0.75	0.25
3	1.0	0.25	0.75

The dataset are named as S1, S2 and S3. Moreover, it has three attributes such as metoo, hallowan and facebook security. An association value between 0 and 1 denotes the relationship between tweet dataset and the hashtag categories. To remove the similarities that have low association values, a confidence threshold T is introduced. Table 3.2 represents the semantic rule provided in Table 4.1 with confidence threshold T as 0.5. Normally, the quality of a formal concept can be considered as the description of the concept. Thus, the similarities between the object and the concept must be the separation of the similarities between the objects and the attributes of the concept. An association value in semantic rule formal context denotes all the relationship between the object and an attribute. Then based on semantic rule theory, the intersection of these association values must be the minimum of these association values. The semantic rule concept lattice can afford additional information, such as association values of objects in each semantic rule formal concept and similarities of semantic rule formal concepts that are important for the construction of concept hierarchy.

Table 3.2 Semantic facet with T >0.5

Dataset	Hashtag1 (metoo)	Hashtag2 (hallowan)	Hashtag3 (facebook security)
1	0.75	-	0.5
2	1	0.75	-
3	1.0	-	0.75

Hashtagger++ for hashtag generation while the formal concepts are also generated mathematically, distinct formal concepts is created on the basis of the difference in terms of attribute object and the traditional concept lattice. This produces the effect of concepts as interpreted by humans. Based on this observation, a group formal concept is infused into conceptual groups of semantic rule conceptual classification. This should have the following properties: The hierarchical similarities should establish groups based on concepts and they are obtained from semantic rule sets using semantic rule lattice concept. That is, a concept indicated by the group can be a subset or superset of another concept inferred in the group.

Algorithm: Hashtag Generation

Input: Beginning concept is CS of concept lattice F (K) and a similarity threshold TS

Output: A set of generated conceptual groups SC

Steps:

```
1: SC {}
2: F(K) An empty concept lattice
3: Add CS to (K)
4: for each sub-concept of CS in F(K) do
5: (C ) Conceptual_Group_Generation(C,
(K), TS)
6: if E(CS,C ) = < TS then
7: SC SC{ (C )}
8: else
9: Insert (C ) to (K) with sup (F (K)) as a sub-concept of
CS
10: end if
11: end for
12: SC ← SC { F(K)}
```

A formal concept should belong to one minimum group generated through concepts, but it can also be more than a single concept group. This attribute is obtained from the characteristic of concepts that an object can belong to more than a single concept.

Semantic tagger Generation:

Semantic tagger is created in this step from the semantic rule context concept hierarchy. This is performed on the basis that both semantic and hashtag maintain the formal definitions of concepts.

Conversely, a concept defined in Hashtagger++ consists of both positive and negative data (data that are defined in unstructured and structured format), whereas a concept in hashtag just emphasizes its intentional characteristics (structured data). For generating the semantic hashtag, both intentional and extensional data need to be converted using Hashtagger++ concepts into the equivalent classes and relations of hashtag

The study of the hashtag in tweet dataset classification process and semantic search process allow one to conclude that association between the two dataset could be helpful, and to have both semantic rules and hashtag based approach. For example, the user selects an existing topic from the topic hashtag, if it is a new one, the user can create the topic and place it in the appropriate position in the hashtag and then formulate a search goal. But in the proposed work the placement and detection of sub category is performed automatically.

Given a user test tweet dataset based on a specific domain D, a list of domain hashtag modules is listed from the case base. The given test tweet dataset content is similar to another one existent in the training set, in this case the user chooses it to reformulate the next category. Many steps are started simultaneously such as: · term detection, semantic based topic detection, indexing and classified by topic and category.

The proposed Hashtagger++ has the option for using prior labeled dataset. Rather it finds the labels. Therefore wrongly classified dataset can be used to enrich category keyword lists. However, note that only those tokens of the classified tweet dataset are being added to the current keyword list which carried out an acceptable intersection with current keywords. This double assurance strategy protects against enrichment by irrelevant tokens.

IV. RESULT AND ANALYSIS

EXPERIMENT SET UP

a. Software and Platform:

The Hashtagger++ framework described above was coded using C#.Net version 4.0 for initialization, keyword database pruning and feature vector preparation tasks. The C#.Net word libraries are used for semantic based evolution. The .Net framework and its library provide numerous advantages, the proposed system is a dynamic one, and datasets are not static. User can use own dataset for the evaluation. MSSQL Connecter was established to make database queries to Hashtagger++. The software was run on an i7 quadcore processor 2.4GHz with Windows 7.

b. Data Sets:

The proposed system used real-time and synthetic datasets. Different corpus adopts different rules and models. Some have dataset with specialized vocabulary containing words that are repeated frequently. On the other hand, corpus derived from certain sources exhibit creative writing style with word occurrences seldom repeated in their dataset. The objective was to achieve a corpus specific combination of statistical and context facets that gives the most accurate classification for varying writing styles and the average size of dataset. In order to validate the efficacy of the approach on varying corpora, the proposed system experimented different sources for the experiments.

1) *Twitter dataset:* Currently the most widely used test collection for proposed hashtag detection and recommendation is twitter articles related to the current scenario. The data was originally collected and labeled by twitter. In the course of implementing the CONSTRUE text categorization system.
2) *Synthetic Domain related Datasets:* In the proposed system, the synthetic dataset are also used. This handcrafted dataset which is known as Domain corpus containing selected news articles from the twitter. This dataset has several categories: hashtag1(metoo), Hashtag2(hallowan), Hashtag3(facebook security) etc. Before applying the Hashtagger++ method to a large data set, there is a need to assess its performance on a prototype corpus with smaller number of dataset. So, therefore generated two data sets for each of the above details by extracting dataset randomly and assigning them to the following datasets:
   a) *Dataset1:* For each of the afore-mentioned sources, the prepared a Dataset1 comprising 12 dataset with 2 dataset in each of the four respective hashtag categories. Among the 2 dataset in each category, 2 of them were selected for training the classifier and the remaining 1 tweet dataset were used for testing the classifier.
   b) *Dataset2* For each of the two corpuses, the Large Datasets comprised 120 dataset with 40 dataset in each of its four categories. Among the 40 dataset in each category, 70 dataset were selected for training the classifier and the remaining 50 were set aside to test the classifier. In this chapter, the Hashtagger++ implementation and performed an analysis of each dataset to make a comparative assessment of their degree of repetitiveness and contextual content.

Table 4.1 summarizes the datasets. Column 1 shows the names of the datasets. Column 2 indicates the average size of the tokenized dataset. Column 3 gives the total number of dataset in the experiment. Column 4 gives an overall measure of the statistical content in the corpus in terms of average TF_IDF per tweet dataset denoted as t. Column 5 gives a measure of the overall contextual content of the corpus in terms of average context score per tweet dataset in the corpus, denoted as k. this will use these values to assess how appropriately the Hashtagger++ assigned the semantic value in each case.

Table 4.1 dataset summary

Datasets	**Average tweet dataset size (No of tokens)**	**Number of dataset**	**Average TF_IDF per tweet dataset**	**Average Context score per tweet dataset κ**
Data set1	123	12	0.07	0.1161
Data set2	808	120	0.09	0.1622

Exploration on different Datasets

The next process depicts results of applying the Hashtagger++ on each of the chosen Datasets. In each case, this plots the classification accuracy, along consecutive generations of the effective feature driven evolution process promulgated by the Hashtagger++ system. For ease of reference and for comparison, this has collects the optimal weights that were experimentally obtained for all statistical, Lexical Semantic and facets at the end of the exploration process for sample dataset is given in **Table 4.2.**

Table 4.2 results obtained at every stage for sample dataset

Dataset	tweet dataset classification is the process of grouping of text mining is an iterative process, which gives high information and effective data in the text mining domain using pattern analysis and pattern classification
Total sentences	2
Total Terms	32
Total Characters	211
TF-IDF	an [1] analysis [1] and [2] classification [2] data [1] tweet dataset [1] domain [1] effective [1] gives [1] grouping [1] high [1] in [1] information [1] is [2] iterative [1] mining [2] of [2] pattern [2] process [1] process, [1] text [1] the [2] using [1] which [1]
Stop word elimination	tweet dataset classification process grouping mining iterative process which gives high information effective data text mining domain using pattern analysis pattern classification
TF-IDF after stop word elimination	analysis [1] classification [2] data [1] tweet dataset [1] domain [1] effective [1] gives [1] grouping [1] high [1] information [1] iterative [1] mining [2] pattern [2] process [2] text [1] using [1] which [1]
Sequence Detection	tweet dataset Start: 0 End: 8 classification Start: 9 End: 19
Weighted Keywords	Classification, Mining, Patterns
Semantic Keywords	Category, group, extract

The performance of the proposed algorithm is measured based on the detection time, accuracy, similarity and unlabeled data handling. The followings are the initial performance analysis of the above sample dataset results.

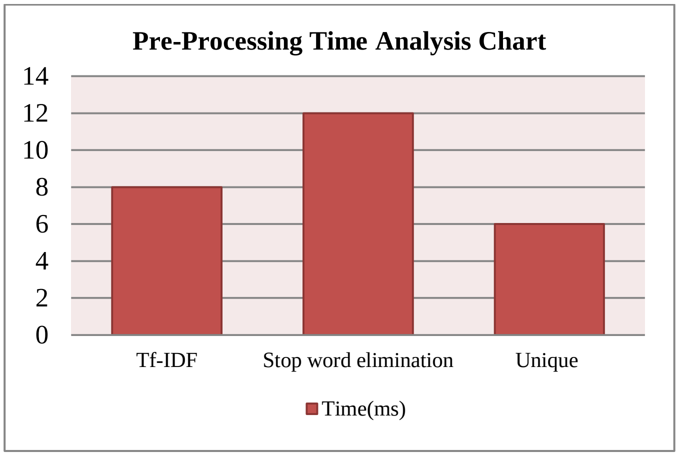

Figure 4.1 Pre-Processing time analysis chart

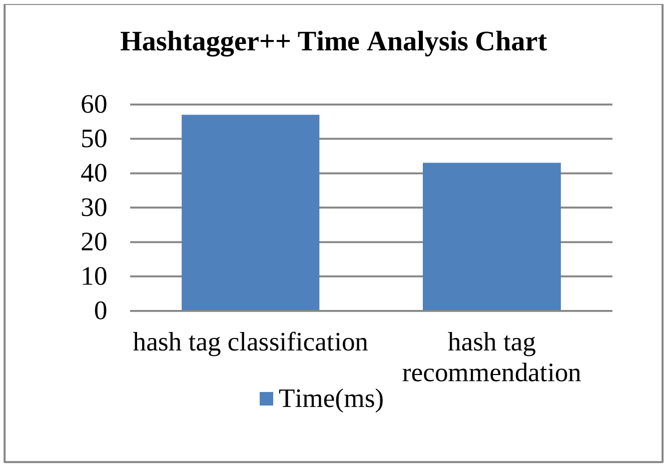

Figure 4.2 Hashtagger++ time analysis chart

For the above given sample dataset, the time is evaluated. The **Figure 4.1** shows the preprocessing time in the Hashtagger++ process. That includes the term detection, frequency and stop word elimination processes. The **Figure 4.2** shows the time taken analysis chart for the Hashtagger++ process. The proposed classifier MCHC-SVM and semantic tagger and recommendation take less time for detection and slightly high for new dataset classification. The recommendation process includes the weighted semantic feature matching, so the result is more accurate and fast.

Performance comparison

Assessment of overall performance

In this subsection, the report gives the results and overall performance of the Hashtagger++ model. So, the first process is comparing its accuracy with that obtained by L2R classification on the same corpora. Then the next ornaments the variety of prior solutions present in the final iteration. Finally this gives the salient performance parameters of the best feature obtained for each dataset and compare them with previously reported results.

Comparison with L2R Model

In order to compare with the Hashtagger++ approach with a L2R, this chapter conducted L2R based tweet dataset grouping process using only *TF_IDF* facets of each tweet dataset in each of the data sets. The existing L2R was trained and tested for each twitter dataset. **Table 4.3** tabulates the accuracy results obtained for the two approaches.

a) The proposed approach performs comparatively better than L2R for both datasets of each hashtag category. The average accuracy for the collaborative method is 95.55% as compared with 76.44% with the L2R method, thus giving an improvement of 25%.

Table 4.3 Performance Comparison between L2R and Hashtagger++ approaches

Datasets	Accuracy using L2R (%)	Accuracy using Hashtagger++ (%)
Dataset0	86	96.5
Dataset1	84	97
Dataset2	83	96

b) In cases where the L2R method gave acceptable results, *i.e.* 86% for the Dataset0 dataset and 84.5 % for the Dataset1, the Hashtagger++ approach enhanced it in both cases to 96.5% and 97% respectively.

c) For the synthetic and large dataset 2, the L2R approach led to rather poor results which were dramatically improved with a collaborative approach. For instance, the classification accuracy of the Dataset2 was only 83% using a L2R approach. This improved to as much as 96% with the Hashtagger++ approach. This is because the HR (Hashtag Recommendation) system was able to utilize the context based featured maximally in the domain corpus.

B) Recall and Precision Analysis

Figure 4.3 exhibits the proposed systems accuracy, which presents in the final iteration produced by the Hashtagger++ for the first dataset, in terms of two conflicting objectives precision and recall. Solution 1 has the highest precision at 87.77% but poorest recall at 78%. Solution 5 on the other hand has the best recall at 87.77% but least precision at 78%. Solutions 2, 3 and 4 are positioned in between. The Hashtagger++ does give the benefit of tradeoff choices between precision and recall. The user has the flexibility to choose a solution that best suits the target application.

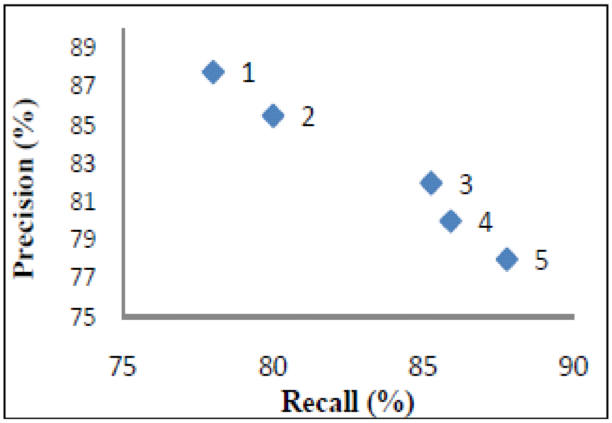

Figure 4.3 precision and recall for the Hashtagger++ process Accuracy, Precision, Recall & F1-Measure

Table 4.4 shows the overall performance of the proposed scheme. It shows the micro-averaged accuracy, precision, recall and F-measure on the above specified 3 different datasets. The L2R ranking process in base paper shows an average precision of 81.56%, an average recall of 82.15% and an average F-measure of 81.56% on Twitter datasets. In comparison, as shown in Table 6, the Hashtagger++ gave an average precision of 95.24%, an average recall of 94.75% and an average F-measure 95.41%.

Table 4.4 Outcome of the proposed Hashtagger++ model and its various parameters

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Dataset0	96	88.56	96	92
Dataset1	100	100	100	100
Dataset2	97.7	100	97.3	98.9

In above table shows the outcome of the proposed Hashtagger++ model and its various parameters are discussed in the table 4.4. The F-measure of the proposed system is approximately 92% (as observed in the graph) on Dataset0 dataset and approximately 98.9% on a small dataset using their large dataset Dataset2. The main process of the Hashtagger++ model it reduces the unwanted facets, In contrast with the existing works that rely only upon multi-word facets within a tweet dataset, Hashtagger++ have utilized the highly structured network of words connected by their lexical relationships as stored in the Hashtag and the well organized categorical information compiled in semantic dictionary. The above results reveal that a Hashtagger++ approach that garners support of hashtag databases such as Word Net and Semantic to perform contextual text analysis and leverages it with L2R have a clear edge over past attempts at combining the two approaches by interpreting contexts in the form of word-groupings present in the same dataset.

Accuracy calculation:

The system finally performs the analysis to show the accuracy of the proposed system. Accuracy refers to the proportion of data classified an accurate type in total data, namely the situation TP and TN, thus the accuracy is

$$\text{Accuracy}= \frac{\text{TP + TN}}{\text{TP+TN+FP+FN}} * 100\ \%$$

4.1 equations for accuracy calculation

Detection and identification of high ranked facets and most negative facets and its frequency can be generalized as the following Table 4.5:

- True positive (TP): the count of dataset detected correctly.
- True negative (TN): the count of dataset detected when it is actually not a specific domain.
- False positive (FP): The count of dataset detected as irrelevant when it is actually relevant one, namely false alarm.
- False negative (FN): The count of dataset detected as relevant one when it is actually irrelevant, namely the dataset which can be detected by Hashtagger++ system.

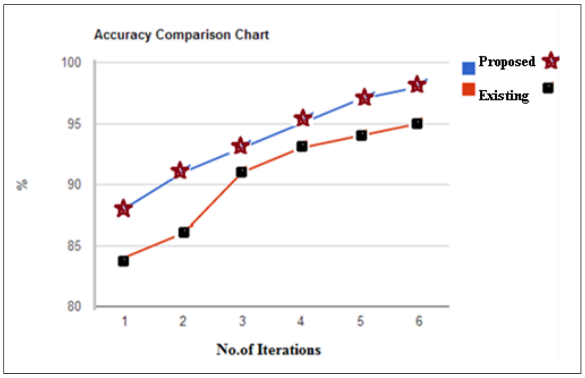

Figure 4.4 Accuracy Comparision Chart

Table: 4.5 Performance comparison table

Metrics	L2R	Hashtagger++
Facet extraction Time(ms)	3.5	2.8
Facet matching Time (ms)	5.6	4.4
Efficiency	Ordinary	Better
Precision (%)	90.7	97.5

Nowadays, tweet dataset classification in system requires high detection rate and low false alarm rate, thus the research compares accuracy, detection rate and false alarm rate, and lists the comparison results of various dataset.

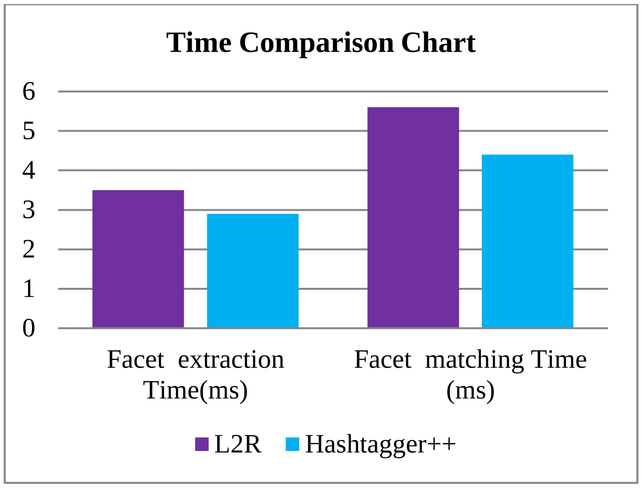

Figure 4.5 Time comparison between existing L2R and proposed Hashtagger++

The above figure represents the comparison between existing and proposed system based on the Training time. This chapter compared the training time of the other classification algorithms with the Hashtagger++.

V.CONCLUSION AND FUTURE SCOPE

The tweet dataset classification and hashtag recommendation is an important research area, where accuracy is only achieved when the hashtags are properly identified. In many of the existing works, hashtag for each of the tweets were manually provided. The development of hashtag with semantic rules provides rich sources of keywords based on rich facet structure. The Hashtagger++ contains MCHC-SVM, which populates the high featured hashtag for training process, and this reduces the test time of the application. Thus, the chapter concludes that Hashtagger++ methods can be utilized to reduce the high dimensionality of the feature space which is typical of statistical facets while still retaining the semantic cohesiveness of dataset. And this hashtag the proposed method enabled user to add more corpus-specific semantic facets from their training dataset, the system generates unique and important facets for robust tweet dataset classification. The proposed work overcomes the cold-start problems, helps to track the stories in twitter domain and many.

Limitations The use of Hashtag and semantic rules Algorithm as a means to perform the weight exploration exhibited its own advantages. This obtained a set of fine iterative solutions with a range of precision-recall tradeoff possibilities. The proposed Hashtagger++ technique does have a high time complexity as it requires an exploration of the weight space with training conducted for each iteration in each generation of evolution.

Future Scope The concept of Belongingness can be utilized for real time online context based posts and mail classification. Different web portals house dataset in differentiated forms such as Tweets, blogs, comments, FAQs, posts, tagged images etc. Each has its unique semantic characteristics that can be subjected to knowledge based contextual analysis for applications such as localized sentiment prediction, articles categorization etc. the next idea is enabling the Hashtagger++ to integrate various contextual facets.

REFERENCES

[1] R.Dovgopol and M. Nohelty, “A hashtag recommendation system for twitter data streams,” Computational Social Networls , 3:3 (2016).

[2] A. Mazzia and J. Juett, “Suggesting hashtags on twitter,” EECS 545m, Machine Learning, Computer Science and Engineering, University of Michigan (2009).

[3] F. Xiao, T. Noro, and T. Tokuda, “News-topic oriented hashtag recommendation in twitter based on characteristic co-occurrence word detection International Conference on Web Engineering, 2012.

[4] B. Shi, G. Ifrim, and N. Hurley, “Learning-to-rank for real-time high-precision hashtag recommendation for streaming news,” Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp. 1191–1202, 2016.

[5] M. Vidhyalakshmi, and P. Radha, “Socaial Hash Tag Techniques Using Data Mining – A Survey”, International Jouirnal of Scientific Research in Computer Science and Engineering, Vol-6, Issue-3, pp.86-92, Jun 2018.

Authors Profile

Ms. M. Vidhyalakshmi pursued Bachelor of Computer Application from Bharathiyar University, India in 2008 and Master of Science from Anna University, India in year 2010. She is currently pursuing Master of Philosophy in Bharathiyar University and currently working as Multimedia Trainer in Rathnavel Subramaniam College of Arts and Science, Coimbatore, Bharathiyar University, Tamil Nadu, India since 2016. Her main research work focuses on Data Mining. She has 2 years of teaching experience.

Ms. P. Radha pursued Master of Philosophy from Alagappa University in year 2002 and Ph.D. from Alagappa University in year 2013. She is currently working as Assistant Professor in Department of Computer Science, Government Arts College, Bharathiyar University, Coimbatore, Tamil Nadu, India since 2014. She has published more than 15 research papers in reputed international journals and conferences including IEEE and it’s also available online. Her main research work focuses on Data Mining. She has 19 years of teaching experience.