Polarity Detection using Effective Machine Learning Classifier

Subbulakshmi R
Student,
Dr.Mahalingam College Of engineering and Technology, Pollachi.

Thirukumar K
Assistant Professor (SG),
Dr.Mahalingam College Of engineering and Technology, Pollachi.

ABSTRACT

Item detection from a tweet is a common task to understand the current movies/topics attracting a large number of common users. However the unique characteristics of tweets (short and noisy content, and a large data volume) make the item detection a challenging task. Existing techniques proposed for item detection uses battery of one class classifier using key word matching techniques and SVM classifier and those techniques provide better accuracy but the features are extracted are found to be noisy, this is a major limitation in SVM classifier.

In this system a SVM classifier with genetic algorithm optimization is proposed. In GA optimization we use 'accuracy 'of SVM as a fitness function; only the best features are selected. And this will improve accuracy for item detection and also the system provides user rating based on the polarity of tweets. This system is expected to improve in terms of classification accuracy when GA is combined with SVM.

General Terms

Support Vector Machine (SVM), Genetic Algorithm (GA), Entropy Weighted GA (EWGA).

Keywords

Item detection, Polarity detection, Twitter.

1. INTRODUCTION

Posting comments about TV programs/Movies using second screen devices[1] (e.g. tablets) is very common. The **Twitter** is the most known and used micro blogs among the preferred social services to post short messages while watching TV. In order to elicit user preferences from a micro blog comment, we need to recognize if the user writing about specific TV shows or Movies .However automatically detecting the subjects of the tweets is a challenging task because of 140 – character limit of tweets (comments) strongly affects the way people write on twitter.

Tweets are unstructured data, Tweet involve

- Minimal contextualization
- Use of slangs
- Abbreviations
- Tiny URLs, etc., so the tweet has more noise and difficult to understand.

An example tweet:

> ***Anyone going to @cstn this year? @theperfectfoil will be presenting BLJ in Imaginarium tent OnThurs evening, July 5th***

Refers to a talk that would be given by Steve Taylor ('The Perfect Foil') about the movie blue like jazz ('BLJ') during Cornerstone music festival in Illinois ('cstn') on Thursday ('Thurs').

In this work a solution to analyze the content of tweets and to identify whether they refer to one (or more) known items is proposed. e.g., movies or TV programs (i.e., Item detection) and item detection prevents the use of traditional classifiers because in that if they want to add to/remove the item means the classifiers required to retrained instead of that here battery of one-class classifiers are used ,each one class classifier is trained for each item in the catalog. When a new item (i.e., class) is added to the catalog, we simply need to train an additional one-class classifier, while no updates are required for the remaining one-class classifiers previously trained.

Implementation of each one-class classifier as a pipeline composed by three stages is decided, An unknown tweet is initially processed by stage 1, that can either assign the tweet to its associated class or it can discard the tweet. Only tweets not classified at stage 1 are processed by stage 2. Similarly, only tweets not classified by stage 2 are processed by the last stage. Tweets not even classified by stage 3 remain not classified by this specific classifier. We used two keyword matching algorithms in stages 1 and 2: these algorithms are based on regular expressions. As for stage 3, we implemented a more advanced approach based on the support vector machine (SVM) and genetic algorithm (GA) algorithm for item detection. After that we can identify the polarity of the tweets of that item using GA based on the positive and negative tweets aggregate counting providing star rating for user reference.

2. EXISTING SYSTEM

An unknown tweet is initially processed by stage 1, that can either assign the tweet to its associated class or it can discard the tweet .Fig 1 shows Only tweets not classified at stage 1 are processed by stage 2. Similarly, only tweets not classified by stage 2 are processed by the last stage. Tweets not even classified by stage 3 remain not classified by this specific classifier. We used two keyword matching algorithms in stages 1 and 2 are based on pattern matching. As for stage 3 implemented a more advanced approach based on the SVM algorithm [2].

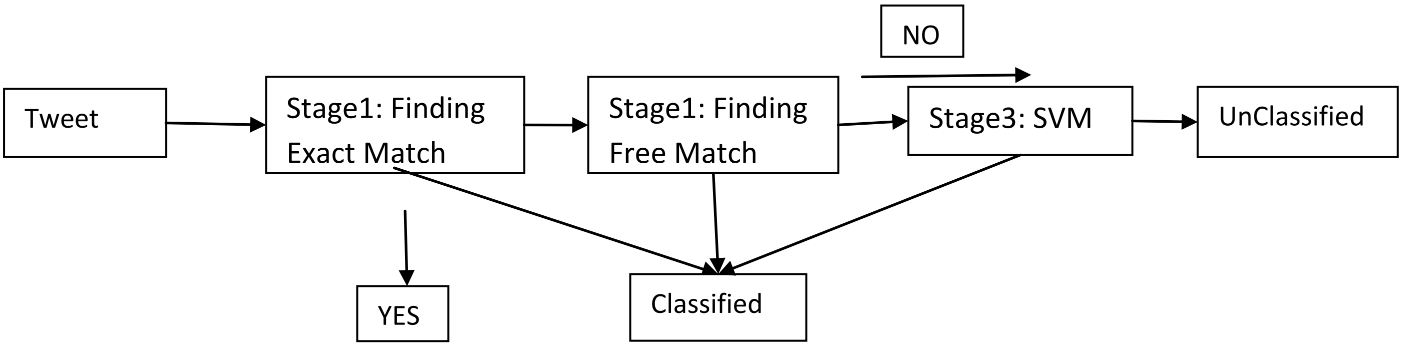

Fig 1: Existing stages

2.1 Stage 1: Exact Match

This stage deploys a simple keyword matching on the basis of the movie title, denoted as exact match. Given the item-i classifier, this algorithm classifies the tweet as belonging to class i if the text contains exactly the title of movie i.

E.g. Tweet: I like Harry Potter
Item name: Harry Potter

2.2 Stage 2: Free Match

Stage 2 of the one-class classifiers is composed of a keyword matching algorithm, denoted as free match. Differently from the keyword matching implemented in stage 1, stage2

(i) Searches for single words of the title within the tweet and
(ii) Searches for words of the title contained in other words

In the tweet. For instance, this stage is able to find a matching between the tweets

E.g. the future is back to the present in the past

And the movie **`Back to the Future'**. Still, this approach allows us to match the movie title also with hash tags and mentions. For instance, the tweet

E.g. atabduction just got home and it was awesome!!!! Is matched by stage 2 because it contains the mention `atabduction', but not by stage 1.

2.3 Stage 3: SVM-Based

2.3.1 Text Preprocessing Tasks

Tokenization, Stop Word Removal, Stemming[3][4] is performed.

2.3.2 Feature Selection

The following four types of features are extracted:

- **Unigrams** correspond to the single terms. Each distinct unigram is represented by a dimension in the feature space. Unigrams formed by stop words are discarded.
- **Bigrams** are pairs of adjacent terms. Each unique bigram corresponds to a dimension in the feature space.
- **Hash tags** represent the tags given by a user to a tweet. However, due to the lack of constraints and common criteria, users can freely use hash tags, either reusing existing tags or defining new ones. Any different hash tag is represented with a dimension in the feature space.
- **Titles** are not linked to specific terms, but it refers to the degree of matching between the tweet content and the item title. Each possible title - related to an item – is represented by a dimension in the feature space.

2.3.3 Feature Normalization

Binary weighting, The TF-IDF schema are used for normalizing the tweets

2.3.4 Classification

Differently from the previous two stages, the SVM algorithm requires also a learning phase to tune its parameter for a specific class. Learning is performed using a 5-fold Cross validation on a set of classified tweets. The output of the SVM algorithm [5][6][7] is a real number, referred to as decision value. Only tweets with a decision value over a fixed threshold are classified.

The performances of the classification system have been measured using a standard hold-out dataset partitioning. For each one-class classifier [8][9], we randomly selected 200 tweets to form the test set. The class (i.e., the item) of tweets in the test set is assumed to be unknown to the classifier and is used to verify if the predicted class corresponds to the actual class. Part of the remaining tweets has been used to form the training set, used by stage 3 as sample tweets for the learning phase.

2.4 Drawbacks of Existing System

- SVM-Features are selected with noisy data. Features with noisy information increase the complexity of classification[10].
- Rating of the item is not specified

3. PROPOSED SYSTEM

Stage 1 and stage 2 are similar to existing system but in stage 3 for movie detection from tweets feature selection will be done by GA and classification will be done by SVM to improve the item detection accuracy. After detecting the item take all the tweets of that item detect the polarity using sentiment analysis by EWGA. After detecting polarity provide user rating for user reference based on aggregate counting of positive and negative tweets of that item.

3.1 Item Detection by GA+SVM

In stage 3 after text preprocessing task feature extraction will be done by GA is represented in the Fig 2, the same four existing features like unigrams ,bigrams ,hash tags and titles are extracted at last the best feature will be obtained[11][12].

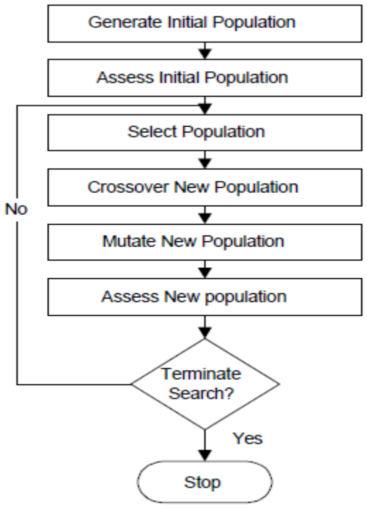

Fig 2:GA process

3.2 Polarity Detection by EWGA

After detecting the item name we will collect all tweets of that item then we will do the EWGA classification for polarity detection. For this purpose we will extract following type of features [13][14][15]

- Parts of speech
- Senti word net
- Frequency
- Stemming
- Chunk label
- Dependency parser
- Positional aspect
- Term distribution
- Thematic word

After applying GA the best feature is selected as a subjective word based on that identifies its polarity whether it is positive or negative.

3.3 User Rating Analysis

After detecting the polarity of user reviews[16] count the number of positive tweets and negative tweets based on the count provide user rating for the movies in the form of stars specified in the below Figure.

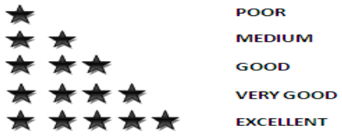

Fig 3: movie rating

3.4 Advantages of Proposed System

- GA+SVM will improve the accuracy.
- Eliminate noise during feature extraction
- EWGA will identify polarity of the tweets.

4. RESULTS

4.1 Data Collection

The input tweets needed for the system is collected from twitter official movie pages. They identified 40 twitter pages concerning of 40 recent movies, 100 tweets for each item, among the most recent, for a total of 4000 tweets. The collected data set is composed of English tweets[17]. Statistical properties of the dataset are

- Items-40
- Tweets-4000
- Average tweets per item-100
- Average terms pertweet-5.2

4.2 Performance Evaluation

4.2.1 Item Detection

By using SVM the item detection accuracy is low due to noise present in the features. An unwanted feature present in the data set creates the noise. While using EWGA+SVM accuracy is increased considerably.

Table 1. Comparison of Accuracy using SVM and GA+SVM

No. Of Tweets	SVM	EWGA+SVM
100	12%	28%
500	38%	38.7%
4000	28.5%	30.2%

Table 1 shows the result of accuracy of SVM and EWGA + SVM while applying for different sets of user movie review. When increasing the number of user movie review the difference between accuracy of two methods is decreasing gradually.

4.2.2 Polarity Detection

In existing system user rating of the movie is not mentioned, but in the current methodology user rating of the movie is calculated which is based on the positive and negative score of the movie review document. From the rating user can know whether the movie is good or bad. Table 2 shows rating of the movies and their document score range.

Table 2. Movie Rating

DOCUMENT SCORE RATING	RATING
0.0-0.2	Poor
0.2-0.4	Medium
0.4-0.6	Good
0.6-0.8	Very Good
0.8-1.0	Excellent

5. CONCLUSION

In this work

- Best feature will obtained using GA, so the performance of SVM classifier will increase for item detection.
- User rating of the movie will be obtained from polarity detection using GA

Rating of the movie is detected based on the polarity score of each movie review document so the user can identify whether the movie is good movie or not a good movie based on the star rating provided for each movie. Item detection accuracy will be improved by using the hybrid algorithm (SVM+GA) instead of using SVM. In this work GA is used for best feature selection.GA eliminate the noisy unwanted features during item detection.

The Comparison is done between SVM and GA+SVM for item detection. The existing work accuracy is 28.5% and the current work accuracy is 30.2%. From the result the overall improvement in accuracy is 1.7%. The classification accuracy was improved using GA+SVM.

Future work can be extended including Temporal Information based on **Electronic Programming Guide (EPG)**.

E.g., matching Time Message Posted with the TV program Scheduling

EPG technique provides user of television, radio, and other media application with continuous updated menus displaying broadcast programming based on that finding the item name. It will improve the accuracy of item detection.

Furthermore, the analysis of tweets as conversations (E.g., a sequence of tweets with questions/answers). This additional Information could be used to disambiguate two movies with the same title but released in two different years. It will improve the accuracy of item detection.

6. REFERENCES

[1] M.Lochrie and P.Coulton.(2012), 'Sharing the viewing experience through second screens', In Proceedings of the 10th European Conference On Interactive TV and Video.ACM.pp 199-202.

[2] P.Cremonesi, R.Pagano, S.Pasquali, and R.Turrin, "TV Program Detection in Tweets," Proc. ACM, EuroITV'13, June 24–26, 2013.

[3] M.F.Porter. Readings in information retrieval,1997.

[4] E.Charniak, C.Hendrickson, N.Jacobson and M.Perkkowitz.(1993), 'Equations for part-of-speech tagging', In Proceedings of the 11th National Conference on Artificial Intelligence.pp 784-789.

[5] T. Joachim's.. "Text categorization with support vector machines: Learning with many relevant features" In Proceedings of the 10th European Conference on Machine Learning, ECML '98

[6] J.T.Yau Kwok.(1998), 'Automated text categorization using support vector machine', Proceedings of the International Conference on Neural Information Processing(ICONIP).pp 347-351.

[7] Cortes, C. & Vapnik, V. (1995), ' Support-vector network'**,** Machine Learning20.3, pp 273-297.

[8] M. Manevitz, M.Yousef, "One-Class SVMs for Document Classification", Journal of Machine Learning Research 2 (2001) 139-154 Submitted3/01; Published12/01.

[9] D. Tax, "One-class classification – Concept-learning in the absence of counter-examples" PhD thesis,TU Delft, 2001.

[10] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin.(2010), 'A Practical Guide to Support Vector Classification'.ACM.pp 35-47.

[11] X.Ming Zhao, De-S. Huang, Y.Cheung,H.Wang and X.Huang "A novel hybrid GA/SVM system for protein sequences classification" IDEAL 2004, LNCS 3177, pp. 11–16, 2004.Springer-Verlag Berlin Heidelberg 2004

[12] Liaoyang LIU, Hui FU "A Hybrid Algorithm for Text Classification Problem", PRZEGLĄD ELEKTROTECHNICZNY (Electrical Review), ISSN 0033-2097, R. 88 NR 1b/2012.

[13] A.Das, S.B.Opadhyay "Subjectivity Detection using Genetic Algorithm", the 1st workshop on computational Approach, 2010.

[14] A.Abbasi,H.Chen,andA.Salem "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ,ACM Transactions on Information Systems, Vol. 26, No. 3, Article 12, Publication date: Jun 2008.

[15] B.Pang and L.Lee.(2008), 'Opinion Mining and Sentiment Analysis', Found Trends Inf.Retr.,2(1-2):pp 1-135.

[16] P.Cremonesi,F.Garzotto,R.Turrin(2012), 'User Effort vs. Accuracy in Rating-based Elicitation**'**, Proceedings of the sixth ACM conference on Recommender system.ACM,pp 27-34.

[17] The movie tweets dataset is available for download at,http://homo.dei.polimi.it/cremones/recsys/Microblog_Item_Detection.zip