SPORTS MATCH ANALYSIS USING SOCIAL MEDIA MINING

Alfin William
Department of Computer Applications
College of Engineering Trivandrum

***Abstract*—The increase of social media utilisation has allowed streaming the voice of sports fans that have essentially lead to storing fan-generated, large-scale opinions about sports match and team performance. Although research utilizing social media data for the consumer market studies have sharply risen in the recent decade, there is a lack of studies using social media mining approach to improve team performance. In this paper, an opportunity mining approach is proposed to identify opportunities to improve team performance based on text mining and cluster analysis. This mining approach utilises proven data mining techniques like KNN and TF-IDF to effectively extract and categorize fan suggestions and use it to enhance overall team performance.**

***Index Terms*—Text-mining, clustering methods, Team performance, Algorithms, Data mining**

I. Introduction

Apart from having tremendous talent, exceptional team performance, and rigorous training, various sports athletes, coaches, managers, teams, and leagues take advantage of the Web Big Data as it has content to provide true insights regarding the critical factors associated with winning or losing a game. One of the conventional and active methods for team performance analysis is to use real time video data that are captured from on-field stadium cameras and crunched into thousands of data points per second which provides each players' performance metrics, such as player speed, position and possession time. However, this method of using video data is not cost efficient as data processing and analysis are complicated, computationally burdensome, and slow. More recently, IoT devices are incorporated are applied such as using wearable device including GPS in smart watches to measuring and calculating team performance based on running speed, distance, time of a player in realtime. However, this study proposes new approach to analyze team performance, especially focusing on the fans' perspective. It focuses on the value of the fans as a strategic partner in addressing some of challenges that a team has. In the fields of business and man- agement, the expertise, ingenuity, and creativity of individual members of the public are harnessed as an innovative problem solving approach. In this regards, this study suggests finding the factors associated with winning or losing from outside of the field. Before, during and after a sporting event, fans share their opinions and unique analysis of the match. More- over, evaluation and analysis by the experts and sports analysts provide not only useful information and knowl- edge of team performances but also wisdom to win. Thus, the comments, Op-Eds, and even tweets written by the public can be considered as a unique source of data to assess team performance which is essentially equivalent to "Wisdom of the Crowd". In particular, this study uses Outcome-Driven Innovation (ODI) methodology, a strategy and innovation process of making product and marketing decision, to determine potential 'opportunities' with regards to team performance to help teams reach their goals (i.e. providing satisfaction to fans through winning) based on the "Wisdom of the Crowd."

II. LITERATURE SURVEY

This section presents related literature concerning Sports Match Analysis using Social Media Mining.Social Media mining techniques are broadly classified into Data collection phase, Data processing phase, Result representation phase.

A. Data Collection Phase

Web forum comments related to a specific sports match should be collected as the primary step of the proposed approach. The collected material should consist of online comment data generated by sports fans on web forums; our approach identifies major team performance-related topics and criticisms that are currently being discussed by fans and the satisfaction analysis of the team's performance in overall match is performed from a fan-centered perspective therefore comments from fans should be the primary source of data. Collection of online comments may be conducted with techniques used to collect large-scale social media data such as web crawling via interfacing with application programming interfaces (APIs).

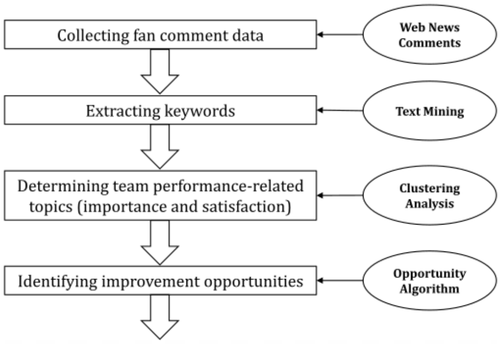
Collecting fan comment data
Web News Comments
Extracting keywords
Text Mining
Determining team performance-related topics (importance and satisfaction)
Clustering Analysis
Identifying improvement opportunities
Opportunity Algorithm

Fig. 1. Multiple Stages of User Comments Retrieval and Processing

1) Social Media API: Social media APIs of various vendors including official Reddit and twitter API is used to collect specific user comments and tweets from a thread using a hashtag like keyword or unique thread id in case of Reddit.

Twitter is an American microblogging and social networking service on which users post and interact with messages known as ”tweets”. Registered users can post, like, and retweet tweets, but unregistered users can only read them. Users access Twitter through its website interface, through Short Message Service (SMS) or its mobile-device application software. [1]

There has been a steady increase in the sheer volume of user tweets proportional to the increase in years. This is largely untouched pool of data that contains insights to many unrecognised patterns and information.

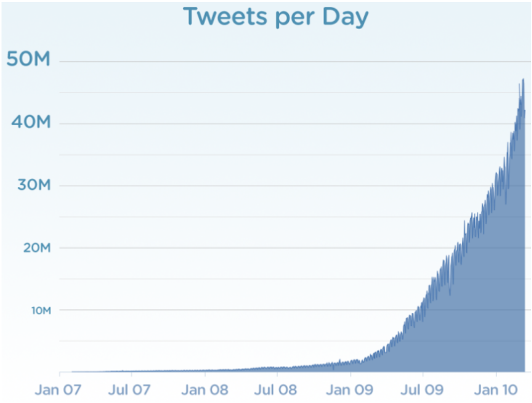

Fig. 2. Increase in tweets with respect to years

Reddit (stylized in its logo as reddit) is an American social news aggregation, web content rating, and discussion website. Registered members submit content to the site such as links, text posts, and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called ”subreddits”, which cover a variety of topics like news, science, movies, video games, music, books, fitness, food, and image-sharing. Submissions with more up-votes appear towards the top of their subreddit and, if they receive enough up-votes, ultimately on the site's front page. [2]

API clients must authenticate with OAUTH verification of vendors to ensure added security and also needs to authenticate unique user id keys with vendors to establish a successful connection to API.

2) Tokenization of data: Tokenization is a very common task in NLP, it is basically a task of chopping a character into pieces, called as token, and throwing away the certain characters at the same time, like punctuation. Tokenization includes word and sentence tokenization . Tokenization a data corpus allows easy computation and manipulation of its relevant data and ignoring the less needed semantic and language jargons.Tokenization of a word corpus is seen as an initial step in performing text classification, intelligent chatbot, sentimental analysis, language translation, etc. It is vital to understand the pattern in the text to achieve the above-stated purpose. These tokens are very useful for finding such patterns as well as is considered as a base step for processes like stemming and lemmatization.

3) Removing Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. This collection of words are grammatically necessary in a sentence but is less useful during processes like data mining. The initial stage of preprocessing of data include removing this stop words from the data corpus so as to simplify the dataset and perform efficient computational tasks in it. Removal of stop words are generally performed using python's Natural Language Toolkit Library.

4) TF - IDF Algorithm: After the extraction/collection of a set of fan comments of the target sports match, the next step of our approach is to extract keywords (or key phrases) from the fan comments to structure each of the comments. This can be performed using Term Frequency - Inverse Document Frequency Algorithm. Term frequency – Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.[1] It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today; 83 percent of text-based recommender systems in digital libraries use tf–idf.[3]

Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. tf–idf can be successfully used for stop-words filtering in various subject fields, including text summarization and classification.

One of the simplest ranking functions is computed by summing the tf–idf for each query term; many more sophisticated ranking functions are variants of this simple model.

Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, ”the brown cow”. A simple way to start out is by eliminating documents that do not contain all three words ”the”, ”brown”, and ”cow”, but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. However, in the case where the length of documents varies greatly, adjustments are often made (see definition below). The first form of term weighting is due to Hans Peter Luhn (1957) which may be summarized as:[4]

The weight of a term that occurs in a document is simply proportional to the term frequency.

Because the term ”the” is so common, term frequency will tend to incorrectly emphasize documents which happen to use the word ”the” more frequently, without giving enough weight to the more meaningful terms ”brown” and ”cow”. The term ”the” is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less-common words ”brown” and ”cow”. Hence an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Karen Spärck Jones (1972) conceived a statistical interpretation of term specificity called Inverse Document Frequency (idf), which became a cornerstone of term weighting:[5]

The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.

The idea behind tf–idf also applies to entities other than terms. In 1998, the concept of idf was applied to citations.[6] The authors argued that ”if a very uncommon citation is shared by two documents, this should be weighted more highly than a citation made by a large number of documents”. In addition, tf–idf was applied to ”visual words” with the purpose of conducting object matching in videos,[7] and entire sentences.[8] However, the concept of tf–idf did not prove to be more effective in all cases than a plain tf scheme (without idf). When tf–idf was applied to citations, researchers could find no improvement over a simple citation-count weight that had no idf component.[8]

Although a list of keywords can be obtained by calculating the TF-IDF, some of the keywords may be irrelevant, grammatically incorrect, or too generic for textual analysis. Identification of keywords can be facilitated by applying TF-IDF while excluding ono- matopoeic words (e.g., ‘haha’, ‘OMG’), emoticons and irrelevant words (e.g., other matches rather than the target one) from the final keyword list. The final step in our proposed approach is to structure the fan comments as an array of keywords and their frequency that appear in their corresponding comments.

B. Data Processing Phase

After collecting and preprocessing the data, the next step in the approach is to define team performance-related topics and compute the importance value of each topic using clustering algorithm. Clustering algorithms such as hierarchical clustering or k-means are applied to a similarity matrix.

1) Hierarchical Clustering: In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:[9]

Agglomerative: Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

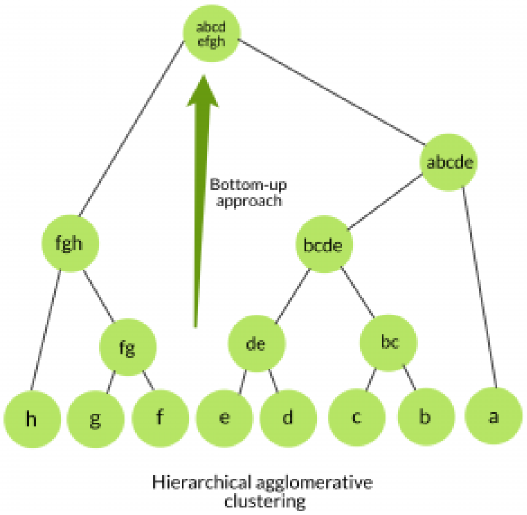

Fig. 3.

Divisive: Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

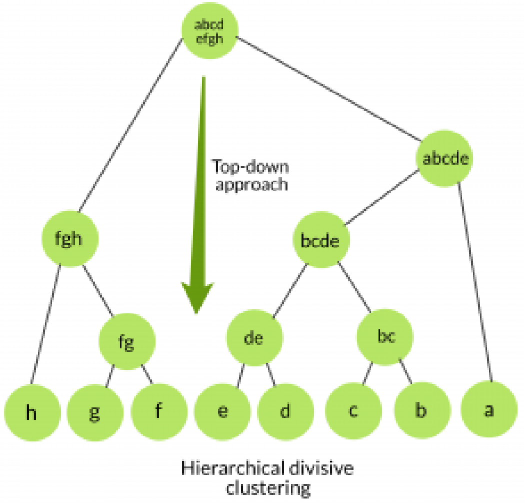

Fig. 4.

2) K-Means Clustering: K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-Means minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can for example be found using k-medians and k-medoids.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

Clustering algorithms such as hierarchical clustering or k-means [16] are applied to a similarity matrix. The results of clustering analysis help ODI experts identify analogous fan requirements, and in turn, fan requirements can be identified by using the identified topics with the assigned importance.

When analyzing clustering results, fan requirements must be inferred from the unambiguous evidence. Good clus- tering assigns one or two customer requirements per cluster. Importance is measured by the frequency of fan comments in each cluster. If many fans mention about a factor, then developing a solution to the factor is an important fan require- ment [10]. In this respect, the number of comments can be used as the surrogate for importance.

For computing the level of satisfaction regarding the topic of interest (i.e. team performance-related topic), the level of satisfaction may be categorized/labeled as satisfied, dissatisfied and neutral. The method for categorizing each comment may be based reaching a consensus (i.e. Delphi method) on the satisfaction level through content analysis or through an algorithmic procedure such as sentiment analysis (as done in Jeong et al. [11]).

Due to the lack of high performing and accurate methods for categorizing/labeling the level of satisfaction regarding each topic, the Delphi method is rec- ommended as of now. Specifically, given a set of topics (i.e. clusters) denoted as T i , for i = 1, . . . , N , where N denotes the total number of topics, the average satisfaction of each topics, denoted TS i as indicated in (2) is calculated as:

$$\mathrm{TS}_i = \frac{\sum_{j=1}^{j_i} CS_{i,j}}{\mathrm{J_i}}$$

Fig. 5. delphi method

3) Sentiment Analysis: Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, ”beyond polarity” sentiment classification looks, for instance, at emotional states such as ”angry”, ”sad”, and ”happy”.

Precursors to sentimental analysis include the General Inquirer, which provided hints toward quantifying patterns in text and, separately, psychological research that examined a person’s psychological state based on analysis of their verbal behavior.[12]

Sentimental Analysis can effectively analyse the fan comments emotions and hence derive their satisfaction and approval regarding individual player performances and about the overall Sports match in general.There are mainly two approaches for performing sentiment analysis.

Lexicon-based: count number of positive and negative words in given text and the larger count will be the sentiment of text.

Machine learning based approach: Develop a classification model, which is trained using the pre-labeled dataset of positive, negative, and neutral.

Multi-Nomial Naive Bayes Classification using scikit-learn is the optimal method to classify fan comments in this specific scenario.

C. Result Representation Phase

To effectively represent the derived findings , an optimal and lightweight python GUI framework should be used that can accommodate the entire results of most relevant fan comments and their sentiment analysis and cluster classifications without causing computational and resource overhead.The most optimal GUI framework for achieving this objective is PyQt.

1) PyQt: PyQt is a Python binding of the cross-platform GUI toolkit Qt, implemented as a Python plug-in. PyQt is free software developed by the British firm Riverbank Computing. It is available under similar terms to Qt versions older than 4.5; this means a variety of licenses including GNU General Public License (GPL) and commercial license, but not the GNU Lesser General Public License (LGPL).[3] PyQt supports Microsoft Windows as well as various flavours of UNIX, including Linux and MacOS (or Darwin).[4]

III. CONCLUSION

This study of literature survey realises the utilization of online fan comments for the identification of sports team performance improve- ment opportunities through combining text mining and opportunity algorithm as the core methodology. Regarding the specific steps of the approach, each team performance- related topics from fans’ perspective was defined by clustering of the online fan comment data. Following the clustering, each topic’s measure of importance and satisfaction were computed. Specifically, the concept underlying the computation of importance (of the topic) is contribution (i.e., number of times mentioned) and the concept underling the computation of satisfaction (of the topic) is sentiment (i.e., positive vs. negative). As the final step, the opportunity algorithm is utilized in order to assess the opportunity value and improvement direction of team performance-related topics from the perspective of sports fans. The approach was demonstrated using the 2018 FIFA World Cup final qualification of the Korean National football team matching against Uzbekistan in September 5 th , 2017. The performance improvement directions based on the top 16 topics for the Korean National football team were found through this case study. The topic with the highest opportunity was about the role of the sports governance body; specifically, the management practice and policies of the Korean National football team yielded the largest opportunity value. The proposed approach contributes to the identification and assessment of new strategy opportunities across various sports domains, using fan comments and social media data. It thereby assists team coaches, athletes, and sports governance bodies to identify areas for improving performance by capturing potential opportunities in the perspective of sports fans. Furthermore, given the underlying method provides "opportunities" to be mined as long as there are the "Wisdom of the Crowd", the proposed method is generalizable to other domains. For instance, utilizing google and yelp reviews as the "Wisdom of the Crowd" may allow restaurant owners to improve aspects of their hospitality management without having to spend an arm and a leg on management consultants. As such, if there is an opportunity for improvement in a certain area in addition to the "Wisdom of the Crowd", the proposed method should be able to reveal opportunities based on the concept of importance and satisfaction.

References

[1] Aliza Rosen: 'Tweeting made easier', IEEE J. Ocean. Retrieved November 7, 2017.

[2] Treibitz, T., Schechner, Y.Y.: 'Reddit Competitive Analysis, Marketing Mix and Traffic. Alexa Internet. ', MRetrieved July 12, 2019.

[3] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text" in Proc. 14th ACM Int. Conf. Multimedia, 2006, pp. 221–230

[4] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in Proc. ACM Int. Conf. Intell. User Interfaces, 2012, pp. 189–198.

[5] Manning, C.D.; Raghavan, P.; Schutze, H. (2008). "Scoring, term weighting, and the vector space model" (PDF). Introduction to Information Retrieval. p. 100. doi:10.1017/CBO9780511809071.007. ISBN 978-0-511-80907-1. "TFIDF statistics — SAX-VSM".

[6] Bollacker, Kurt D.; Lawrence, Steve; Giles, C. Lee (1998-01-01). CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. Proceedings of the Second International Conference on Autonomous Agents. AGENTS '98. pp. 116–123. doi:10.1145/280765.280786. ISBN 978-0-89791-983-8.

[7] Sivic, Josef; Zisserman, Andrew (2003-01-01). Video Google: A Text Retrieval Approach to Object Matching in Videos. Proceedings of the Ninth IEEE International Conference on Computer Vision – Volume 2. ICCV '03. pp. 1470–. ISBN 978-0-7695-1950-0.

[8] Beel, Joeran; Breitinger, Corinna (2017). "Evaluating the CC-IDF citation-weighting scheme – How effectively can 'Inverse Document Frequency' (IDF) be applied to references?" (PDF). Proceedings of the 12th IConference.

[9] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

[10] A. W. Ulwick, What Customers Want: Using Outcome-Driven Innovation to Create Breakthrough Products and Services. New York, NY, USA: McGraw-Hill, 2005

[11] Liu, C., Meng, W.: 'Removal of water scattering'. Proc. Int. Conf. Computer Engineering and Technology, Chengdu, China, April 2010

[12] T.-M. Yen, Y.-C. Chung, and C.-H. Tsai, "Business opportunity algorithm for ISO 9001: 2000 customer satisfaction management structure," Jan. 2007.