Post Summarization of Micro-Blogs

By
Mehreen Ali Gillani
2010-NUST-MS-PhD-IT-32

Supervisor
Dr. Muhammad Usman Ilyas
Department of Electrical Engineering

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(May, 2013)

Approval

It is certified that the contents and form of the thesis entitled "**Post Summarization of Micro-Blogs**" submitted by **Mehreen Ali Gillani** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Muhammad Usman Ilyas**
Signature: _____

Date: _____

Committee Member 1: **Dr. Ali Mustafa Qamar**

Signature: _____
Date: _____

Committee Member 2: **Dr. Zawar Hussain Shah**

Signature: _____
Date: _____

Committee Member 3: **Dr. Khawar Khurshid**

Signature: _____
Date: _____

Dedication

I dedicate this thesis to my parents who have given me the opportunity of education from best institutions and supported me throughout my life.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Mehreen Ali Gillani**
Signature: _____

Acknowledgments

This thesis would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this research work.

First and foremost with immense gratitude I acknowledge the support and help of my thesis Advisor Dr. Muhammad Usman Ilyas. You have been my inspiration during hard times of research work.

To the committee members, Dr. Ali Mustafa Qamar, Dr. Zawar Hussain Shah and Dr. Khawar Khurshid, thank you for your feedback, encouragement and support.

To my husband, Syed Atif Raza, thank you for encouraging me during this difficult process. Your words have always strengthened me.

To my beloved parents and family, thank you for your continuous financial and moral support.

To Ali Munir, Umay Kalsoom and Talal Hassan who have always given the best advice I often needed along this journey.

My thanks and appreciations also go to my friends at SEECS, NUST.

I also thank to all who, directly or indirectly, have lent their hand in this endeavor.

Above all, utmost gratitude to Almighty Allah, we owe everything to His limitless bounties.

Abstract

Social networking sites e.g. Facebook, Twitter, LinkedIn are becoming popular among users as they are successful in connecting people and have become a great means of information dissemination, communication and entertainment. Popularity of these services motivates us to study characteristics of online social networks. Twitter, is a micro-blogging service launched in July, 2006 by Jack Dorsey. In 2012, more than 500 million users have subscribed Twitter, generating over 340 million posts and 1.6 billion search queries each day. Due to the enormous number of posts generated by Twitter, it is often difficult to understand what is being said by people on a specific topic. Post summarization is a technique to extract short summaries from the collection of posts on a particular topic. In this research work, I have used simple K-Means clustering, an unsupervised learning technique of machine learning to perform post summarization using Twitter status updates. Three different distance metrics: Euclidean, Cosine Similarity and Manhattan were used in K-means clustering. These clustering results were evaluated against previous best post summarization algorithms, and results showed that clustering using Euclidean distance is performing better than existing post summarization algorithms, in terms of Precision, Recall and F-Measure.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

In the last few years social networking sites e.g. Facebook, Twitter, Linkedin are becoming popular among users as they are successful in connecting people and have become a great means of information dissemination, communication and entertainment. Easy accessibility of the enormous content generated by these social networking sites and the popularity of these services motivates us to study characteristics of online social networks.

A relatively new social content is micro-blog in which blog is a short text message of 140 characters or less. The idea of micro blogging was introduced by Jack Dorsey in March, 2006, when he was brainstorming with his colleagues about new ideas for their company. He proposed the idea to share simple messaging service (SMS) of mobile phone as status update among group fellows. From this idea, famous micro blogging site Twitter.com was launched in July, 2006. Since then, Twitter has gained worldwide popularity. According to the company statistic presented in 2012, more than 500 million users have been registered, generating more than 340 million tweets daily and more than 1.6 billion queries each day. Although 40.1% of these tweets are pointless babble, 37.6% are conversational and only 3.6% tweets are regarding main stream news (Pear Analytics, 2009). Most of the conversational tweets, which have been communicated among friends, contain important information. For example, most of the time people tweets describe real time events e.g. political debate, natural disaster, sporting event or TV show. These conversational tweets contain information about what is happening in the real time. But unluckily, the tools available for micro-blog data extraction are in their infancy. For example, the search tool of posts on specific key word provided by Twitter, results in most recent posts instead

of relevant posts. Consequently, it is very likely to receive spam, irrelevant and posts in other languages.

To make use of this huge micro-blog data (being posted every day) and for sake of deep analysis there is a need of automatic post summary to make use of data mining, machine learning and natural language processing.

Post summarization helps user to get these informational tweets with a shorter delay without having a need of reading all irrelevant and redundant tweets being posted. It helps user to quickly identify why a topic is trending and what people say about it over a period of time.

In this research work we have explored the opportunity and challenges of automatic event summary using tweets on sporting events.

1.2 Problem Formulation

The problem considered in this study is how to present an event summary using Twitter status updates only. The problem is formulated as:

- Given an event E, with P number of Twitter updates on E
- Let K is the number of important moments, happened during main event E
- Cluster tweets having timestamp of important moments; extract a set of tweets P_k from P

$$\text{Such that, } P_i ! = P_j \text{ where, } P_i, P_j \subseteq P_k$$

- P_k are the best tweets describing k moments in E

1.3 Prior Work

For the first time, Sharifi [3] in his masters thesis has addressed the need of post summarization. He proposed two algorithms Hybrid Term Frequency Inverse Document Frequency (TF-IDF) and Phrase Reinforcement (PR) to preform single post summary of Twitter trending topic. Inouye [8-9] has applied hybrid TF-IDF to perform multi-post summary of Twitter trending topic.

Nichols et al. [10] have applied Phrase Reinforcement to produce a journalistic summary of sports events using Twitter status updates.

1.4 Proposed Solution

In this work K-Means clustering is applied to perform post summarization of sports event using Twitter status updates. We proposed the multi-phase approach to produce the match summary using Twitter statuses.

In the first, **preprocessing phase**, Twitter posts were collected using hash tags of cricket and football matches. Then we identified peaks in the volume of tweet/minute graph. After this we performed data cleaning. We divided these tweets into two parts: training and test set.

In the second, **feature extraction phase**, most frequently used words were extracted using long tail distribution function [27]. Timestamp of a tweet and top frequently used words were used as features.

In the third, **clustering phase**, we performed K-Means clustering on tweets using three different distance metrics Euclidean, Cosine Similarity and Manhattan. Each tweet was used as vector.

In the **Evaluation phase**, we compared these clustering results of each distance metric with post summary produced by best existing algorithms, Hybrid TF-IDF and Phrase Reinforcement.

1.5 Experimental Results and Findings

In terms of important moments identified in main sport event. Following are the results of all three matches:

1. In Indian Premier League (IPL) final cricket match 2012; Euclidean outperforms Cosine Similarity, Manhattan, Hybrid TF-IDF and PR. Euclidean has 0.896, 0.8 value of Precision and Recall, respectively. Cosine Similarity has second best performance in terms of Precision and Recall.

2. In UEFA Champions League (UCL) final football match 2012; all algorithms have same Recall value as they have identified the equal number of moments. In terms of Precision, Euclidean outperforms all.

3. In UEFA, EURO 11^{th} June, 2012 football match; all algorithms have same Recall value as they have identified the equal number of moments. In terms of Precision, Euclidean and Hybrid TF-IDF both have highest value 0.92. Cosine Similarity and Manhattan are second best having

Precision value 0.86. Phase Reinforcement has lowest Precision value, 0.7.

Results of Automatic Summary evaluation, ROUGE are as follows:

1. In IPL final Cricket match 2012; Euclidean has highest Recall and F-measure value in all three metrics: ROUGE-1, ROUGE-2, and ROUGE-SU. Cosine Similarity has highest Precision and second best value of Recall and F-Measure in all three metrics.

2. In UCL final football match 2012; Euclidean has highest Recall and F-measure value in all three metrics: ROUGE-1, ROUGE-2, and ROUGE-SU. Manhattan has highest Precision but lowest Recall and F-Measure in all three metrics.

3. In UEFA, EURO 11^{th} June, 2012 Football match; Euclidean has highest Precision and F-measure value in ROUGE-1 and ROUGE-2. Hybrid TF-IDF has highest Recall in all three metrics: ROUGE-1, ROUGE-2, and ROUGE-SU. But the Precision and Recall vale is low than of Euclidean. Manhattan has highest Precision in SROUGE-SU but lowest Recall and F-Measure.

In all three matches Manhattan has generated the shortest summary. For IPL and UCL match, Euclidean has generated the longest summary containing 547 and 227 number of words (after removing stop words), respectively. For UEFA EURO 11^{th} June 2012, Euclidean has generated second longest summary containing 136 words and Hybrid TF-IDF has generated the longest summary for this match consisting 167, number of words.

1.6 Key Contribution

Data Collection we have collected Twitter status updates containing hash tags of matches during a match was played. Our data set contain three different matches of Indian Premier League final cricket match, 2012, UEFA Champions League (UCL) 2012 Final football match and UEFA EURO11^{th} June 2012 Football match.

Clustering we have applied K-Means clustering using three different distance metrics: Euclidean, Cosine Similarity and Manhattan, to generate Journalistic summary of sports events using Twitter status updates only. To the best of our knowledge, we are the first who have applied K-Means clustering for post summarization.

Chapter 2

Literature Review

Sharifi [3] trained a naive Bayes classifier to classify posts based upon Google directorys categories (Art and entertainment, Business, Food and drink, Computer game, Health, Politics, Science, Sports, Technology, World). They further proposed two algorithms to perform automatic post summarization. The Phrase Reinforcement algorithm builds graphs from frequently occurring phrases in posts and the path with the highest weight is selected as the post summary [3-5]. The Extended Term Frequency, Inverse Document Frequency (TF-IDF) algorithm was proposed for post summarization which normalizes long posts to get higher weight [3-5]. This method further improves the performance of the Phrase Reinforcement algorithm [3] [7]. Automatic summaries produced by these algorithms are compared with manual summaries [3-10]. Random selections of posts and longest/shortest sentences/posts have served the purpose of baseline results [3] [7]. Recall Oriented Understanding for Gisting Evaluation (ROUGE-1) and content have been used for evaluation metrics [3-5] [7].

Clustering of Micro-blogs is difficult because of non-standard language usage, redundant tweets and limited number of features. Beverungen et al. [6] have observed the impact of post normalization, noise reduction, and different number of clusters used, term expansion and improved feature selection on post clustering performance. They concluded that post normalization slightly degrades the clustering performance while Gap statistic technique performs better to calculate the correct number of clusters for posts. They have also shown that noise reduction improves clustering performance whereas tri-gram for feature selection outperforms uni-gram, bi-gram and term expansion [6]. Inouye [8] has revealed that Bisecting k-means++ algorithm outperforms k-means, bisecting k-means and k-means++ for post summarization.

Twitter tweets on a specific topic mostly contain several subtopics or themes. For multi-post summaries, Hybrid TF-IDF algorithm is used to assign weights to each post and instead of picking single highly weighted post top N posts are chosen, where each post covers a subtopic [8-9]. Then, Cosine similarity measure is used to determine that each post is sufficiently different than other selected posts. In [9] results showed that for post summarizations simple frequency based algorithms i.e. Hybrid TF-IDF and sum-basic summarizer performs better than traditional MEAN, TexRank and LexRank algorithms.

Jeffrey Nichols et al. [10] have proposed an extended Phrase Reinforcement algorithm to produce journalistic summary for sports' special events using Twitter statuses. The algorithm is based on the intuition that sudden increase/spike in number of tweets means that an important event has happened during sports. Then, the algorithm selects highly weighted graph phrase from each spike/sport event. Brendan O Ćonnor et al. [11] have developed an exploratory search application for Twitter topics which cluster tweets into topics based upon terms and keywords frequency.

Chapter 3

Data Set

Twitter is open to data collection. Tweets used in this research work were collected using Python [17] tweet stream 1.1.1 package [18] which provides access to Twitter streaming API. Tweets were recorded using Twitter hash tags for sports event. E.g. for IPL match #ipl for UCL #ucl and for euro #Euro2012 were used.

1. IPL(Indian Premium League) t20 cricket final match between KKR (Kolkata Knight Rider) and CSK (Chennai Super Kings) on 27^{th} May, 2012, at MA Chidambaram Stadium, Chepauk, Chennai, India [30].

2. UCL (UEFA Champion League) football final match between Bayern Munich of Germany and Chelsea of England on 19^{th} May, 2012, at the Allianz Arena in Munich, Germany [13].

3. UEFA European Football Championship known as Euro 2012, football match being held on 11^{th} June 2012, between France and England at Donbass Arena, Donetsk, Ukraine [14].

Table 3.1: Summary of data set

Match	Total Tweets	Max Tweets/Minute	Min Tweets/ Minute
IPL Final, 27^{th} May 2012	46370	1452	85
UEFA Euro, 11^{th} June 2012	233288	3005	1044
UCL final, 19^{th} May 2012	226109	3029	488

In table 3.1, UEFA Euro 11^{th} June 2012 football match has huge number of total tweets, maximum and minimum tweets per minute. Whereas, IPL final cricket match, 2012 has lowest total tweets and minimum tweets/minute. On average IPL has 150-200 tweets/minute. One reason for this is that cricket match has more duration than of football match. The total time of football

match is 90 minute whereas; the duration of IPL final cricket match was 210 minutes.

Chapter 4

Technical Approach

In this chapter the details of our proposed solution is discussed to perform post summarization of sports events using Twitter status updates.

Figure 4.1 is showing the technical approach followed in this study. First of all sports events were identified and their tweets were collected using hashtags .e.g. for IPL #ipl, for ucl #ucl and for UEFA EURO 2012 #euro2012 was used. When a match was ended its commentary was presented on news websites. These news sites were used as reference summary for post summary evaluation.

Then, peaks from tweet/minutes graph were identified. After this data cleaning was performed. The data set was divided into two sets. Training set was almost twice as test set. K, the number of clusters was set on the basis of important moments in test set. Then feature selection and feature extraction was performed. After performing k-means clustering on training set cluster centroids were computed. The distance between these cluster centroids and each point of test set was computed and two tweets of test set having minimum distance with cluster centroid have been picked as cluster result. Similarly k-means clustering has been applied on test set and minimum distance between cluster centroid and training set has been computed. Results of all clusters were combined to form a match summary. This match summary produced by Twitter updates was evaluated against match commentary presented on news websites.

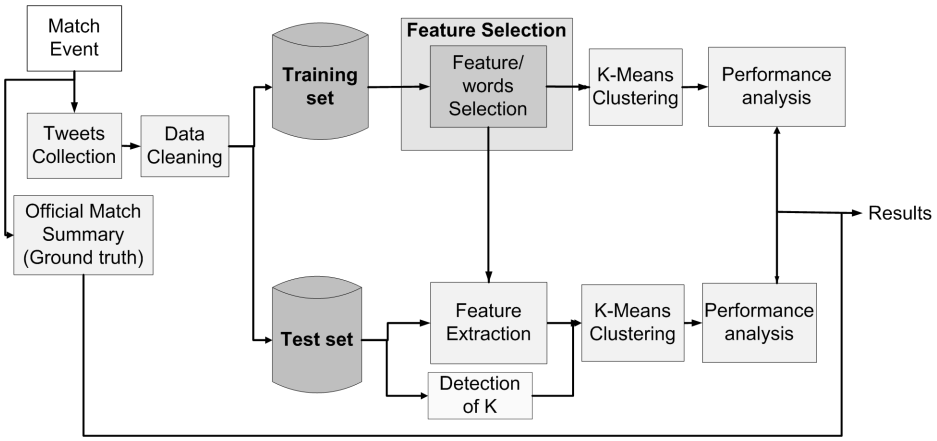

Figure 4.1: Technical approach

4.1 Identifying Peaks from Tweets/Minute Graph

Sudden increase in the volume of tweets indicates that some important moment has been occurred in main sports event e.g. for cricket match the moment can be a wicket, six, end of first inning etc. For football match it can be a goal, half time, match start or end. So, people find a need to comment about this moment.

To identify peaks in tweet volume Local Maxima has been used. It works as follows: At any time i, No. of tweets at time i should be > number of tweets at time *i-1* and at time *i+1*. Figure 4.2 is showing the tweets per minute graph of UEFA EURO 11^{th} June 2012 football match.

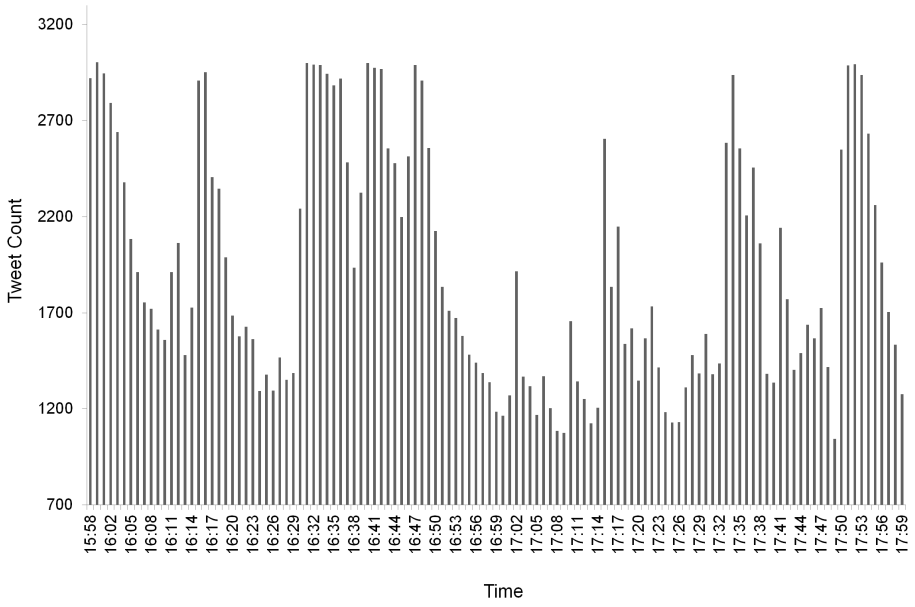

Figure 4.2: Tweet count/minute of UEFA EURO 11^{th} June, 2012, football match

4.2 Methods to Define Minimum Threshold

To set the minimum threshold for peak identification, different methods were tested. Nichols et al. [10] used $Median * 3$ and the other method they have mentioned but didnt work well for their data set is $Median + 2 * std$. Shamma et al. [28] have used $Mean + std$. The results of all three methods are shown in table 4.1. $Median * 3$ has very poor performance as it has not identified any moment in UEFA EURO and UCL final match. Only 1 moment of IPL match was identified. $Median + 2 * std$ has better performance than median *3. Whereas, mean + standard deviation has identified greatest number of important moments. For that reason we have used $Mean + std$ to define minimum threshold.

Table 4.1: Number of important moments identified by different methods to define minimum threshold

Match	**Important Moments**	**Mean + St deviation**	**Median * 3**	**Median +2 * St deviation**
IPL final, 27^{th} May 2012	25	23	1	6
UEFA EURO, 11^{th} June 2012	18	12	Not any	7
UCL final, 19^{th} May 2012	8	7	Not any	5

Figure 4.3 is showing the minimum threshold line drawn by all three methods on tweet/minute graph of UEFA EURO 11^{th} June 2012 football match. The dotted line is drawn by mean + std. It has identified the greatest number of important moments. $Median + 2 * std$ has also identified 5 out of 8 important moments. The line drawn by $Median * 3$ is high enough to identify any important moment in UEFA EURO 11^{th} June 2012 football match.

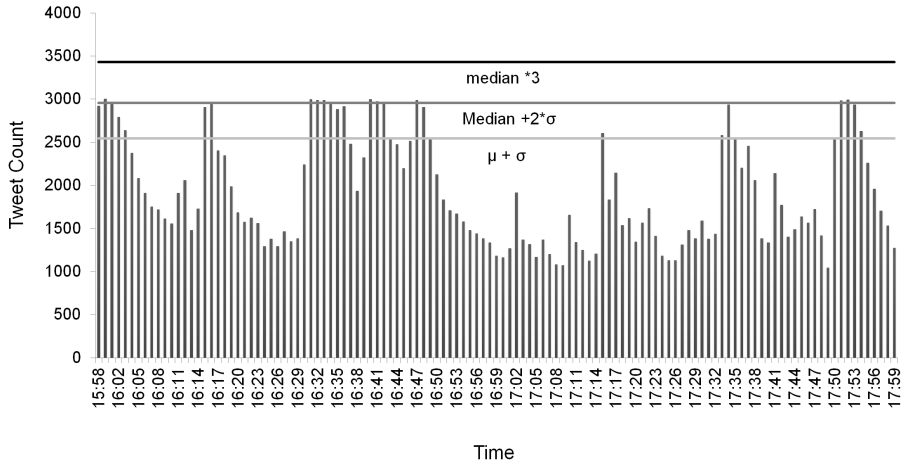

Figure 4.3: Minimum threshold using different methods, of UEFA EURO 11^{th} June, 2012, football match

4.3 Data Cleaning

After identifying peaks from tweets/minute graph, following steps of Data cleaning have been performed:

- *Removal of tweets:*
    1. Language other than English
    2. Containing URLs
    3. Replying other tweets e.g. tweet containing @
    4. Duplicate tweets

Clustering results of without deleting tweets replying to other tweets and after deleting tweets replying to other tweets were compared. The results showed that tweet in reply of other tweets increases noise. For that reason these tweets have been removed. All the tweets with duplicate text are deleted except one.

- *Removal from Tweets:*

  To perform clustering following words have been removed from tweets:

  1. Stop words, hash tags
  2. Special and single character

- *Word normalization:*

  Words have been normalized and unnecessary repetition of letters have been removed e.g. goooooal to goal

- *Lemmatization and Stemming:*

  On tweet text Natural Language Processing Toolkit (nltk) lemmatization and stemming.porter2 was performed.

Lemmatization is a process of grouping different varied forms of a term, which can be analyzed as single term. **Stemming** is a process in which related words are replaced by the stem or root word. Unlike stemming, Lemmatization chooses appropriate lemma by considering the context and part of speech of a term [22].

4.4 Feature Selection

Following features have been selected to perform clustering on tweet text:

1. Time stamp of tweet
2. Each word of a tweet

The intuition behind this is that tweets posted on the same time must contain information about the same important moment occurred on that time.

Tweets must contain repetitive words describing a moment e.g. name of player, wicket, and six etc. for cricket match.

4.5 Feature Extraction

Table 4.2 is showing the total number of unique words after performing data cleaning. These numbers are huge and increase noise during clustering. So the most important words have been extracted using long tail distribution function.

Table 4.2: Number of unique words in each match after performing data cleaning

Matchs	**IPL final, 27^{th} May 2012**	**UEFA Euro, 11^{th} June 2012**	**UCL final, 19^{th} May 2012**
Number of Unique words	3423	3618	2420

Figure 4.4 is showing the word frequency of unique words after performing data cleaning on tweets of UCL final football match, 19^{th} May 2012. Using long tail distribution on average words having minimum frequency, 150 have been selected.

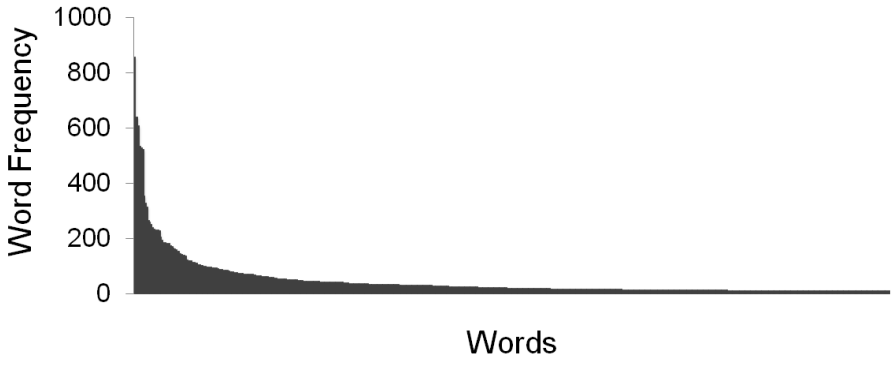

Figure 4.4: Word Frequency of UCL final football match, 19^{th} May 2012

Chapter 5

Clustering

In this chapter the working of K-Means clustering is explained. In this research work K-Means clustering, an unsupervised learning method has been used for micro-blog summarization. For further study on clustering interested candidate may read [15-16, 19-21].

Clustering is an unsupervised learning approach. It is a type of machine learning in which complex hidden patterns are identified from unlabeled dataset [15]. In our case we have tweets which are not labeled and on the basis of tweet text and tweet timestamp we have clustered these tweets. So the tweets describing the same important moment in main event form a cluster. There are different types of clustering e.g. K-Means clustering, hierarchical clustering, Bi-clustering etc. In this research work, we have used K-Means clustering to cluster tweets.

K-Meaning clustering is a centroid based clustering [15]. In which n observations are clustered into k number of clusters [16]. These clusters are user defined. In our case, we have applied k-means clustering to cluster the tweets into number of important moments occurred during main event. This is how it works:

- Let X_1,...., X_N are data points or vectors
- Each data point will be assigned to only one cluster
- C(i) denotes cluster number for the i^{th} observation
- Given an initial set of k means m_1,....,m_k , it continues by alternating between two steps:

1. **Assignment step**:

   In first step, it assigns cluster number to each data point. It computes distance between data point and all clusters mean. And it assigns the cluster number to one having minimum distance from mean.

$$C(i) = arg\min \|x_i - m_k\|^2_{1 \leq k \leq K}, i = 1, ..., N$$

   $Where, m_k$ is the mean vector of the k^{th} cluster

2. **Update step**:

$$m_k = \frac{\sum\limits_{i:C(i)=k} x_i}{N_k}, k = 1, ..., K$$

   Where, N_k is the number of data points in k^{th} cluster

The algorithm converges until the assignments no longer change.

After performing feature selection, feature extraction and defining k, the number of moments during main sports event. Clustering using different distance metrics has been performed.

5.1 Distance Metrics

Distance measurement in clustering parameter defines how distance between data points is measured. The distance metrics used in this study are as follows:

5.1.1 Euclidean Distance

Euclidean distance finds distance between two points using **Pythagorean metric** [19]. Every tweet is a point, vector in Euclidean n-space. And tweet timestamp and tweet words are its features. If A and B are two data points and n is the number of features or dimensions then it computes distance as follows:

$$d(A, B) = \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + ... + (A_n - B_n)^2} = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

5.1.2 Cosine Similarity

It measures the Cosine angle between two vectors. In information retrieval each document is characterized as a vector and each term of a document is the dimension of vector. The value of dimension is the term frequency in the document. It computes how similar documents are with respect to their topics. It is very efficient to evaluate sparse vectors (which have more zero values) [20].

The Cosine similarity between two vectors a and b is calculated as:

$$a.b = \parallel a \parallel . \parallel b \parallel \cos\theta$$

If two vectors have A, B attributes then similarity is measured as:

$$Similarity = \cos(\theta) = \frac{A.B}{\|A\|.\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

5.1.3 Manhattan Distance

Manhattan or Taxicab metric is a little variation of Euclidean geometry. Manhattan distance simply finds the sum of horizontal and vertical components [21]. The distance between two points is measured as the sum of absolute differences of their coordinates [21]. If A and B are two vectors and n are total number of features this is how it computes distance between them: For text analysis, A and B are the terms frequency of vector documents. The Cosine similarity between two documents ranges from 0 to 1, as term frequency cannot be less than 0 [20].

$$d(A,B) = \parallel A - B \parallel_1 = \sqrt{\sum_{i=1}^{n} |A_i - B_i|}$$

Chapter 6

Micro-Blog Summarization

In this chapter, summary produced by different algorithms has been presented and the comparison of total number of important moments identified by each algorithm is summarized.

After clustering tweets, top 2 tweets having minimum distance from cluster centroids have been used as cluster result. Hybrid TF-IDF and Phrase Reinforcement was implemented. Hybrid TF-IDF and Phrase Reinforcement have been applied on longest sentence of each tweet as mentioned in [3-10], whereas, for clustering complete text of tweet is used. The clustering result of longest sentence and whole tweet text has been compared and the results showed that clustering using full tweet text has high performance. For that reason complete tweet text has been used. As stated in [10] after creating phrase graph, two phrases having highest weight and doesn't contain common token (word) have been selected as moment summary.

Table 6.1-6.3 are showing the two moments summary of IPL cricket final 2012, UCL football final 2012 and UEFA EURO, 11th June 2012, respectively.

Table 6.1: First 2 moments summary of IPL match, 27^{th} May 2012

Moment	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Peak 1 19:32, Match started	final between CSK n KKR starting now! IPL Live: Chennai to bat against Kolkata for the title: Chennai Super Kings won the toss and opted to	final between CSK n KKR starting now!. final time! Think CSK will take it.	final between CSK n KKR starting now! final time! Think CSK will take it.	Such a proud moment every time I hear our national anthem anywhr Some of our Indian players dont know how to stand for our National Anthem.. Shukla, Ashwin, Jadeja and one more in KKR.Junk fellows!	Some of our Indian players dont know how to stand for our National Anthem. Shukla, Ashwin, Jadeja and one more in KKR....Junk fellows! CSK 'Whistle Podu' is going to be what you're going to be hearing in your ears, when Lee Narine knock you over!
Peak 2, 14:59 Lee to M Hussey, SIX. Lee to M Vijay, FOUR	Final 5.5: B Lee to M Hussey, 6 runs, 53/0. Final 5.1: B Lee to M Vijay, 6 runs, 41/0.	Final 5.3: B Lee to M Vijay, 4 runs, 46/0. Final 5.5: B Lee to M Hussey, 6 runs, 53/0	6 again. KKR must win today in	Watching the final. Brett Lee's getting smashed for sixes all over the park! Hussey nd Vijay destroying Brett Lee and bringing the 50 partnership in just 5.5 overs	Hussey nd Vijay destroying Brett Lee and bringing the 50 partnership in just 5.5 overs. Watching final match.but looks like kolkata is passing tough time..GO KOLKATA, GO SAKIB!

Table 6.2: Moment summary of UCL final football match, 3^{rd} and 4^{th} peaks 19^{th} May, 2012

Moment	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Peak 3 20:25, Muller Scores, Bayern 1-0 Chelsea	It's about time Bayern scored They've been all over Chelsea the whole game Muller scores the first goal of the match Bayern 1-0 Chelsea	It's about time Bayern scored They've been all over Chelsea the whole game Muller scores the first goal of the match Bayern 1-0 Chelsea	It's about time Bayern scored They've been all over Chelsea the whole game Muller scores the first goal of the match Bayern 1-0 Chelsea	finally Muller scores with a beautiful downward header - just the way they teach it; deserve the lead; 1-0 over Muller heads into the ground and how did it go over Cech into the goal,Bayern 1-0 Chelsea.7mins left.	Muller heads into the ground and how did it go over Cech into the goal,Bayern 1-0 Chelsea.7mins left. Torres to score 2 and win it for Chelsea!
Peak 4, 20:31 Drogba scores, Bayern 1-1 Chelsea	Goal Muller heads the ball into the back of the net Bayern 1 Chelsea 0 82 mins played GOAL!!!! Bayern Munich 1-1 Chelsea Good corner in and a great header by Drogba at the front post Poor keeping by Neuer	Goal Muller heads the ball into the back of the net Bayern 1 Chelsea 0 82 mins played final is insane right now No goals for 83 minutes Bayern scores in 83rd and Drogba comes for Chelsea in 88th	Drogba, i love him but not chelsea, what a goal what a game 1-1 Drogba is the Man	Oh thank you, this game was just resurrected to life with a Chelsea equaliser by Didier Drogba... finals Goal, what a smashing header from Drogba, Chelsea are level.. Game on	Goal, what a smashing header from Drogba, Chelsea are level.. Game on Bayern Munich defending a corner in the final minutes of a Champions League final.

Table 6.3: Last two moments summary of UEFA, EURO 11^{th} June 2012, football match

Moment	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Peak 6 16:47, Half Time	good half of football that, England v France...best all around play I've seen in the tournament so far...1-1 at the break First half of the game England vs France comes to an end Score tied 1-1	Nasri and Lescott both found the net for their country respectively in the first half of the game France 1 England 1 First half of the game England vs France comes to an end Score tied 1-1	France 1 England 1 it's definitely not ending like this and its Half Time France VS England 1-1	Whistle Blown - Half Time - Scores are Equal in England vs France match good half of football that, England v France...best all around play I've seen in the tournament so far...1-1 at the break	good half of football that, England v France...best all around play I've seen in the tournament so far...1-1 at the break How many more times is cabaye going to hack one of our players down before getting booked
Peak 7, 17:52 Match Ended	That's it! France 1-1 England! Could have been better as always! 1-1 good performance from England!! Think England and France would have took a point today	First half was definitely better than the second; France 1, England 1 I said the score would be 1-1 between France and England! I'll go for 2-0 Sweden in the second game	That's it! France 1-1 England! Could have been better as always! 1:1 I'll take that Well done England	FT England 1-1 France ....not the best game that fans would have wanted..but the teams happy wit the result .. England 1-1 France - FT - a good start from England to both teams probably happy with the draw,	A point in opening group game good result against a strong French team onwards, upwards Come on England! Happy with a draw in the end, Considering every1 i spoke to said a France win!!

Table 6.4-6.6 are showing the moments identified by different methods

of IPL cricket final 2012, UCL football final 2012 and UEFA EURO, 11^{th} June 2012, respectively. Where, true means moment is identified and empty means moment is not identified by the method. The grey highlighted boxes are those peaks which are not identified by local maxima. In table 6.4 15:41, 15:42 and in table 6.5 21:27, 21:28 and 21:29 have not been identified by local maxima because these moments are occurring on consecutive minutes and local maxima consider adjacent minutes tweet count. In 6.5 half time in extra time and in table 6.6 the yellow card moments have not been identified because the number of tweets for these moments was low.

For IPL cricket final 2012, Clustering using Euclidean Distance has identified maximum number of moments, 20 out of 25. Coming at the second place, clustering using Cosine Similarity has identified 19 out of 25. Phrase Reinforcement identified 15. Whereas, Hybrid TF-IDF and clustering using Manhattan distance both have identified 14 out of 25 total moments.

For UCL football final 2012 and UEFA EURO 11^{th} June 2012 all methods have identified equal number of moments 10 and 7, respectively.

Table 6.4: Important moments of IPL final cricket match 2012, identified by different methods

Time	Important Moments during IPL Final Cricket Match	Euclidean	Cosine Similarity	Manhattan	Hybrid TF-IDF	Phrase Reinforcement
14:32	Match started	True	True	True		
14:59	Lee to M Hussey, SIX.	True	True	True		
14:59	Lee to M Vijay, FOUR	True	True			
15:25	Bhatia to Vijay, OUT, slower one and what a catch	True	True		True	True
15:33	Kallis to Raina, SIX	True	True	True		
15:41	Pathan to Raina, SIX					
15:42	Pathan to Raina, SIX					
15:43	1. Pathan to Raina, FOUR 2. MEK hussey Fifty	True	True	True	True	True
15:54	1. Narine to Raina, SIX, 2. Fifty Raina	True	True	True	True	True
15:58	Kallis to Hussey, OUT	True	True	True	True	True
16:06	Narine to Raina, SIX	True				
16:11	Shakib Al Hasan to Raina, OUT	True	True	True	True	True
16:12	End of first inning	True	True	True	True	True
16:30	Hilfenhaus to Gambhir, OUT	True	True	True	True	True
16:51	Ashwin to Bisla, SIX	True	True	True		
17:16	Jakati to Bisla, FOUR (Twice)					
17:21	Ashwin to Bisla, SIX,					True
17:36	Morkel to Bisla, OUT	True	True		True	True
17:45	Bravo to Shukla, OUT	True	True	True	True	True
17:49	Bravo to Kallis, SIX	True	True		True	True
17:51	Ashwin to Pathan, OUT	True	True	True	True	True
18:02	Hilfenhaus to Kallis, OUT	True	True	True	True	True
18:06	9 required on 6 Balls	True	True		True	True
18:15	Bravo to Tiwary, FOUR, KKR won the IPL	True	True	True	True	True

Table 6.5: Important moments of UCL final football match 2012, identified by different methods

Time	**Important Moments during UCL Final Football Match**	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
18:45	Chelsea to kick-off					
19:32	Half time whistle	True	True	True	True	True
20:25	Goal by Thomas Muller	True	True	True	True	True
20:32	Goal by Didier Drogba	True	True	True	True	True
20:37	Match started again In extra time	True	True	True	True	True
20:47	Yellow card to Drogba, Penalty to Bayern, saved by Cech	True	True	True	True	True
21:00	Half-time In extra-time					
21:17	Full-time. Time to Penalty kicks	True	True	True	True	True
21:24	Goal by Lahm (Bayern) Bayern lead 1-0				True	
21:24	Mata (Chelsea) has his Penalty saved by the goalkeeper				True	True
21:24	Goal by Gomez 2-0 Bayern	True	True	True		True
21:25	Goal by David Luiz, Bayern 2-1					
21:26	Goal by Manuel Neuer, Bayern lead 3-1	True	True	True	True	True
21:26	Goal by Lampard 3-2	True	True	True		
21:27	Opportunity missed By Olic Bayern lead 3-2					
21:28	Goal by Ashley Cole 3-3					
21:29	Missed by Schweinsteiger					
21:31	Goal by Didier Drogba, Chelsea won the Champions league	True	True	True	True	True

Table 6.6: Important moments during UEFA, EURO 11^{th} June 2012, identified by different methods

Time	**Euro, 11^{th} June (France Vs. England) Important Moments**	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
16:00	Match started	True	True	True	True	True
16:16	Chance missed by Milner	True	True	True	True	True
16:31	Goal Lescott scores	True	True	True	True	True
16:36	Goal save by Joe Hart	True	True	True	True	True
16:40	Goal Nasri	True	True	True	True	True
16:47	Half time	True	True	True	True	True
17:28	Yellow card to Young					
17:52	Full time	True	True	True	True	True

Chapter 7

Evaluation

In this chapter the effectiveness of our proposed solution has been evaluated. The results of post summaries produced by all methods are presented. Manual and automatic summary evaluation results are discussed.

7.1 Manual Summary Evaluation

True Positive: If a tweet correctly describes a moment it is titled as true positive (TP).

False Positive: If a tweet fails to describe a moment or noisy tweet is titled as false positive (FP).

Precision: In context of information retrieval Precision is the fraction of retrieved instances that are relevant [29]. It is computed as:

$$Precision = \frac{TP}{TP+FP}$$

For a match Precision is: Number of tweets identifying correct moments (TP) / Number of tweets identifying correct moments (TP) + Noisy tweets (FP)

Recall: In context of information retrieval Recall is the fraction of relevant instance that are successfully retrieved For a match Event Recall is:

Number of correctly identified moments / total moments

Table 7.1 presents the Precision and Recall of all three matches using different Methods. Clustering using Euclidean distance has managed the

highest Precision and Recall in all three matches. High Precision means tweets extracted in post summary are mostly describing the important moments in main event. Whereas, Low Precision means more noisy tweets are present in post summary. Clustering using Cosine Similarity has second best value of Precision in IPL final and UEFA EURO 2012 11^{th} June football match. Likewise in terms of Recall it has second best value for IPL Final 2012. All Methods have same value of Recall in UCL final 2012 and UEFA 11th June football match. Because all methods have identified equal number of moments. Hybrid TF-IDF has second best value of Precision in UCL final 2012.

Table 7.1: Precision and Recall in terms of moments identified during main event using different methods

Evaluation Metrics	**Match**	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Precision	IPL final, 2012	0.90	0.88	0.67	0.79	0.48
	UCL final 2012	0.88	0.75	0.75	0.84	0.71
	Euro, 11^{th} June 2012	0.93	0.86	0.86	0.93	0.71
Recall	IPL final, 2012	0.80	0.76	0.56	0.56	0.60
	UCL final 2012	0.83	0.83	0.83	0.83	0.83
	Euro, 11^{th} June 2012	1	1	1	1	1

Table 7.2 shows the Recall of key moments of IPL cricket final match, 2012. Clustering using Euclidean distance has highest Recall value in all moment categories. Hybrid TF-IDF and PR has low Recall in fours and sixes. Clustering using Cosine Similarity is again on second position.

Table 7.2: Recall of key moments of IPL cricket final using different methods

Key Moment	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Match start	1	1	1	1	1
End of first inning	1	1	1	1	1
Four	0.67	0.67	0.67	0.33	0.33
Six	0.86	0.71	0.57	0.29	0.43
Out	1	1	0.75	1	1
Match end	1	1	1	1	1

Table 7.3 is showing the Recall mean of key moments of UEFA EURO and UCL final football 2012 match, identified by different methods. All methods have identified equal number of moments in these two matches. For that reason all have got the same value. 0.5 of game start means all methods

have successfully identified the start of UEFA EURO 2012 football match but couldn't identify the start of UCL final, 2012. Yellow card has low Recall. Because total number of tweets are less about these moments. In both of these matches no red card moment has occurred.

Table 7.3: Mean of UCL final, UEFA EURO 11^{th} June 2012, key moments Identified using different methods

Key Moment	**Euclidean**	**Cosine Similarity**	**Manhattan**	**Hybrid TF-IDF**	**Phrase Reinforcement**
Goal	0.78	0.78	0.78	0.78	0.78
Red Card	No event of type	No event of type	No event of type	No event of type	No event of type
Yellow Card	0.33	0.33	0.33	0.33	0.33
Penalty	1	1	1	1	1
Game Start	0.5	0.5	0.5	0.5	0.5
Half Time	1	1	1	1	1
Disallowed Goal	1	1	1	1	1
Match End	1	1	1	1	1

Figure 7.1-7.3 is showing the Precision and Recall values of all three matches. Clustering using Euclidean has highest Precision and Recall value in all three matches, whereas Phrase Reinforcement has lowest Precision in all three matches. All methods have same Recall value for UCL final and UEFA EURO 11^{th} June 2012 football matches because equal numbers of moments were identified by all methods. In these two football matches Hybrid TF-IDF has second highest Precision, whereas in IPL cricket final Cosine Similarity has second best values of Precision and Recall.

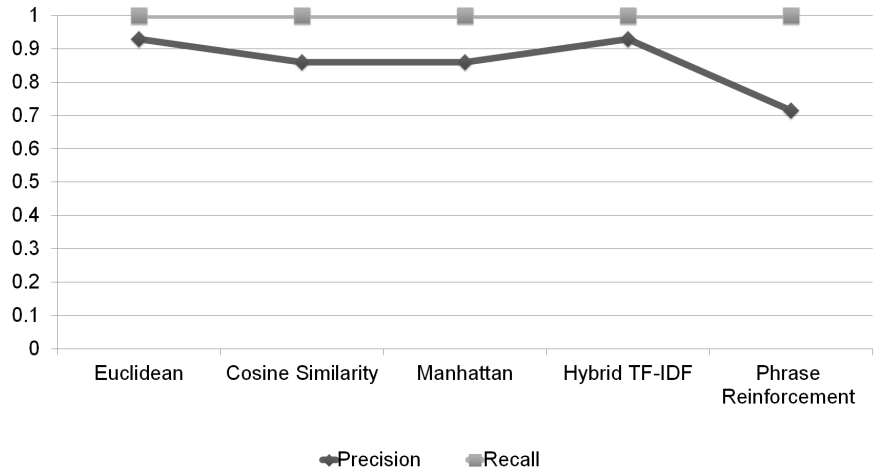

Figure 7.1: Precision and Recall of UEFA EURO football match, 11^{th} June, 2012

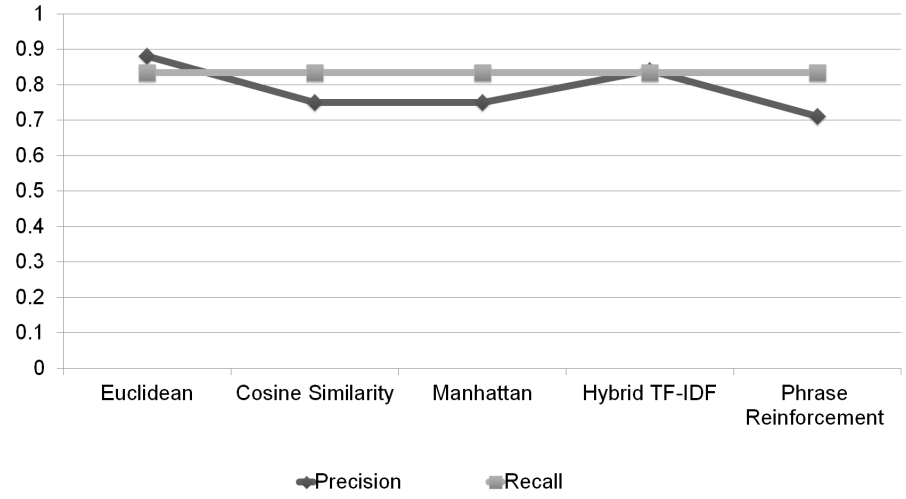

Figure 7.2: Precision and Recall of UCL final football match, 2012

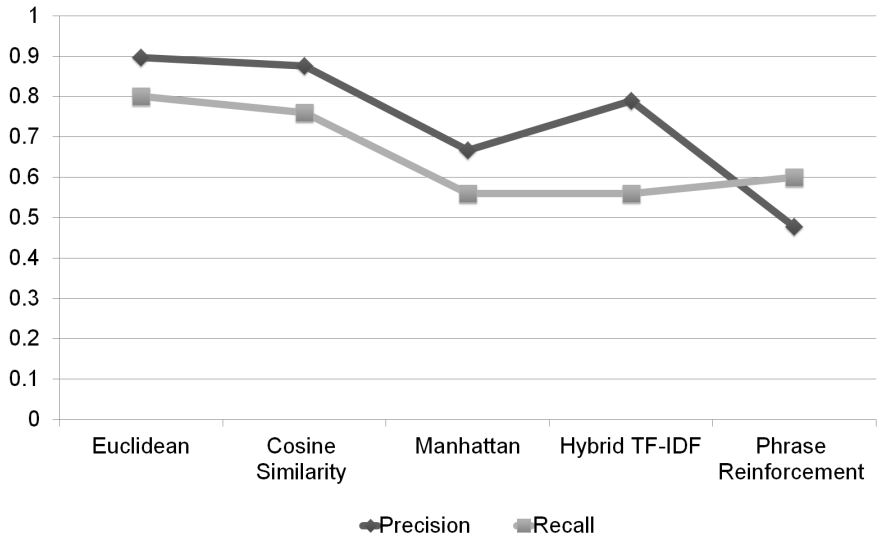

Figure 7.3: Precision and Recall of IPL final cricket match, 2012

7.2 Automatic Summary Evaluation

Till now the evaluation of post summaries have been performed manually. First we read the sentences of summary and then titled them true positive or false positive on the basis of summary presented on mainstreams news. Although these manual evaluations were carefully performed, it might have errors. To evaluate these post summaries generated by different methods we have also performed automatic summary evaluation. In this study automatic summary evaluation has been performed using ROUGE.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package used for automatic summary Evaluation. These Metrics compare an automatically produced summary against a reference or a set of references summary. ROUGE is developed by chin yew lin in university of southern California. It has been used in Document understanding conference 2004 [23]. ROUGE has four basic metrics ROUGE-N, ROUGE-S, ROUGE-W and ROUGE-L which checks overlaps of unigrams, word pairs and word sequences. ROUGE-N works as:

$$ROUGE - N = \frac{\sum_{s \in manualsummaries} \sum_{n-gram \in Match(n-gram)}}{\sum_{s \in manualsummaries} \sum_{n-gram \in count(n-gram)}}$$

Where, n is the length of n-gram, Match (n-gram) is the number of n-grams common in manual summary and automatic summaries. Count (n-gram) is the total number of n-grams in manual summary. N-gram is contiguous occurrence of n words in a sentence.

In this study we have used following three metrics of ROUGE:

1. ROUGE-1
2. ROUGE-2
3. ROUGE-SU

ROUGE-1 operates on unigrams. ROUGE-2 operates on bi-grams and it checks the occurrence of two adjacent terms in a sentence. ROUGE-SU is Skip Bi-grams and unigram-based co-occurrence statistics [24]. It allows gap between pair of words. We have set 4 as maximum gap between two words. The optional parameter -u includes unigram in ROUGE-S. ROUGE-S has been tested in automatic MT evaluation and the initial evidence have shown that ROGUE-S is better than NIST, BLEU, ROUGE-L (LCS-based ROUGE) and ROUGE-N (ngram based ROUGE) [25].

Table 7.4 encompasses the links of news sites containing match summary. The match commentary of these sites was used as reference summary for automatic summary evaluation.

Table 7.4: Links of reference summary used for automatic summary evaluation for all three matches

Match	**Reference Summaries**	**Reference Summaries link**
IPL Final, 27^{th} May, 2012	espncricinfo	http://www.espncricinfo.com/indian-premier-league-2012/engine/current/match/548381.html
	indiatoday	http://indiatoday.intoday.in/story/ipl-5-final-chennai-vs-kolkata-live/1/197822.html
UCL Final, 19^{th} May 2012	uefa	http://www.uefa.com/uefachampionsleague/season=2012/matches/round=2000267/match=2007693/postmatch/commentary/index.html
	skysports	http://www1.skysports.com/football/live/match/259018/commentary
UEFA EURO, 11^{th} June, 2012, France Vs England	espnfc	http://espnfc.com/commentary?id=334181s&cc=4716
	skyports	http://www1.skysports.com/football/live/match/253545/commentary

Figure 7.4 - 7.6 are showing IPL final cricket match results of ROUGE-1, ROUGE-2, ROUGE-SU, respectively. In all three metrics clustering using

Euclidean has highest value of F-measure and Recall. In figure 7.4 and 7.5 it has second highest Precision, whereas, in figure 7.6 it has less Precision than that of Cosine Similarity and of Manhattan. But it has high Precision than that of Hybrid TF-IDF and Phrase Reinforcement. In all three Manhattan has lowest F-measure and Recall. But in all three Hybrid TF-IDF and Phrase Reinforcement has low performance than that of Euclidean and Cosine Similarity.

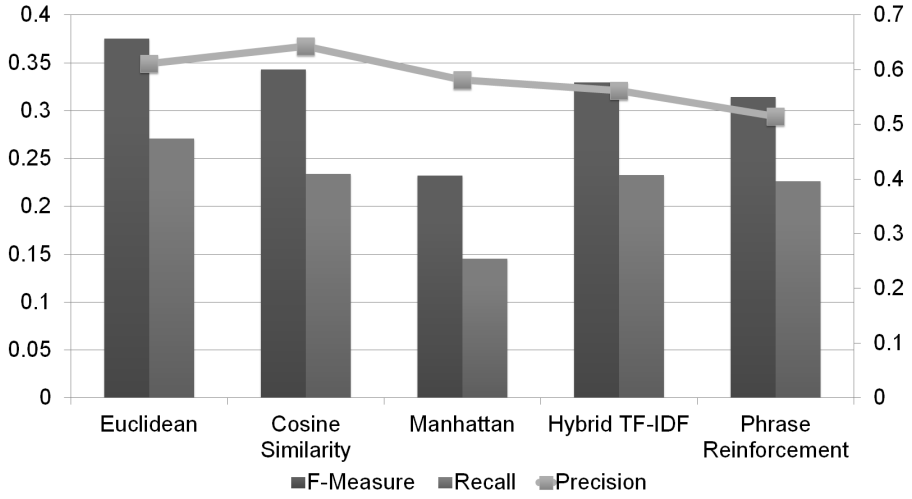

Figure 7.4: ROUGE-1 of IPL final cricket match, 2012

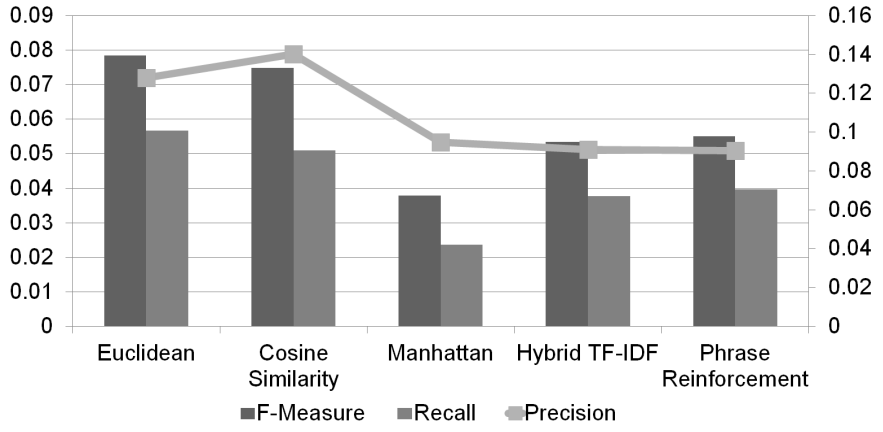

Figure 7.5: ROUGE-2 of IPL final cricket match, 2012

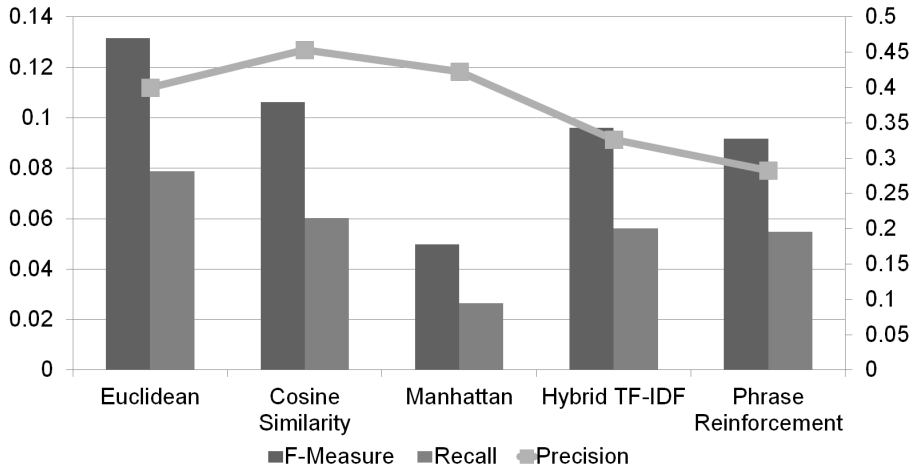

Figure 7.6: ROUGE-SU of IPL final cricket match, 2012

Figures 7.7-7.9 are showing UCL final football, 2012 results of ROUGE-1, ROUGE-2, ROUGE-SU, respectively. In all three metrics Euclidean has highest f-measure and Recall. It has Precision lower than that of Cosine Similarity and Manhattan. But it has high value of Precision than that of Hybrid TF-IDF and Phrase Reinforcement. In all three Manhattan has

highest Precision but lowest Recall. Again Cosine Similarity has the second best performance in all three parameters.

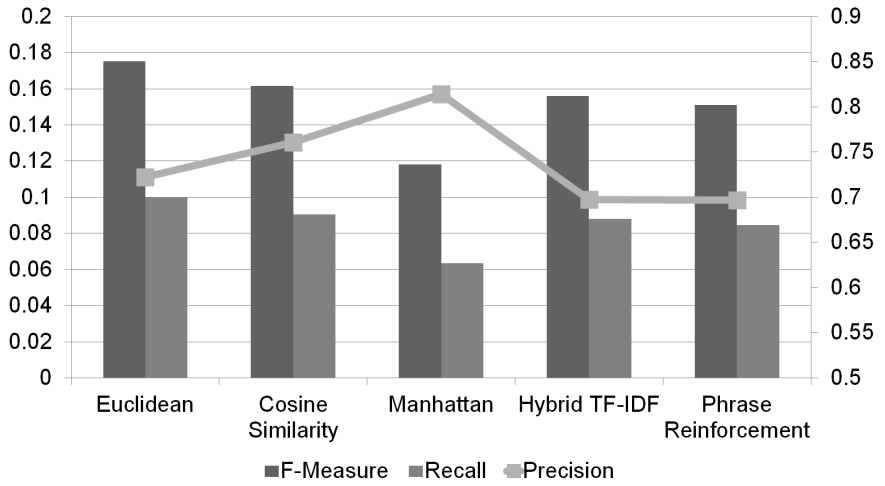

Figure 7.7: ROUGE-1 of UCL final football match, 2012

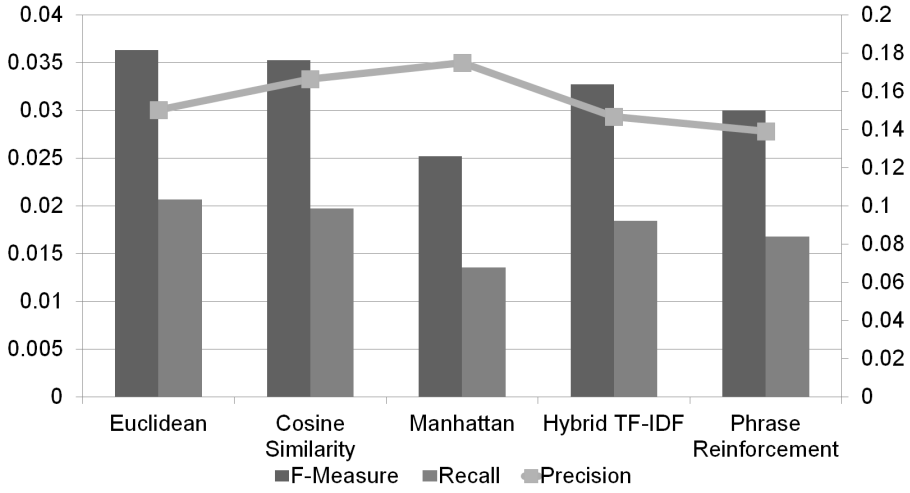

Figure 7.8: ROUGE-2 of UCL final football match, 2012

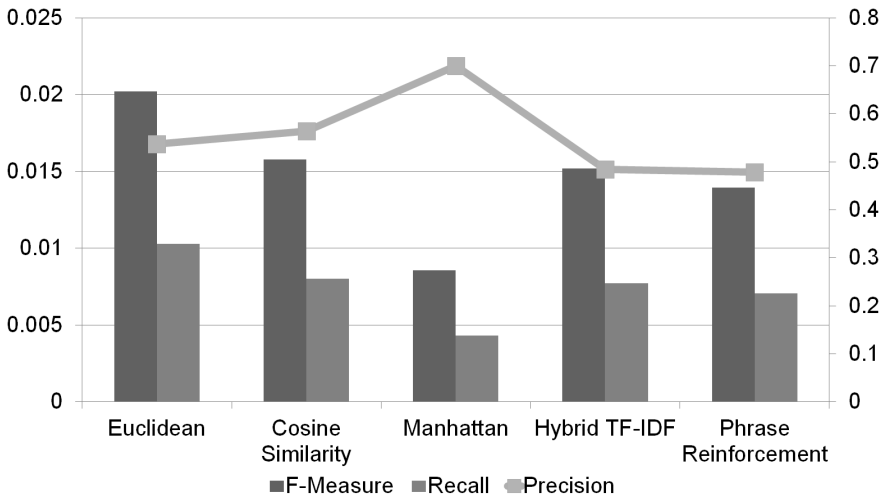

Figure 7.9: ROUGE-SU of UCL final football match, 2012

Figures 7.10-7.12 are showing UEFA EURO 11th June 2012 football match results of ROUGE-1, ROUGE-2, ROUGE-SU, respectively. Results of this match are slightly different than of IPL cricket match and of UCL football match. In all metrics Euclidean has highest values of F-measure and Recall. But Hybrid has second best value of F-Measure and Recall and it has high Precision than of Euclidean. Again in EURO football match as in UCL final match Manhattan has highest Precision but lowest Recall. This time Cosine similarity has low performance than of Hybrid TF-IDF in terms of F-Measure and Recall.

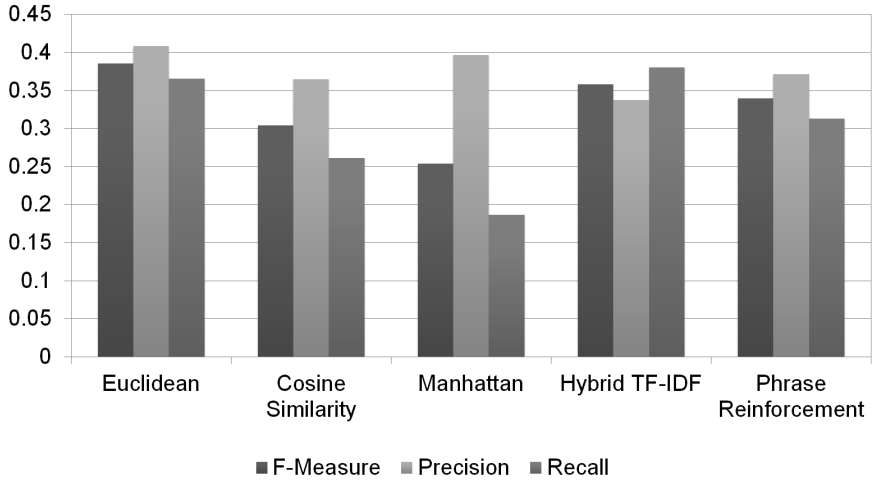

Figure 7.10: ROUGE-1 of UEFA EURO 11^{th} June, 2012, football match

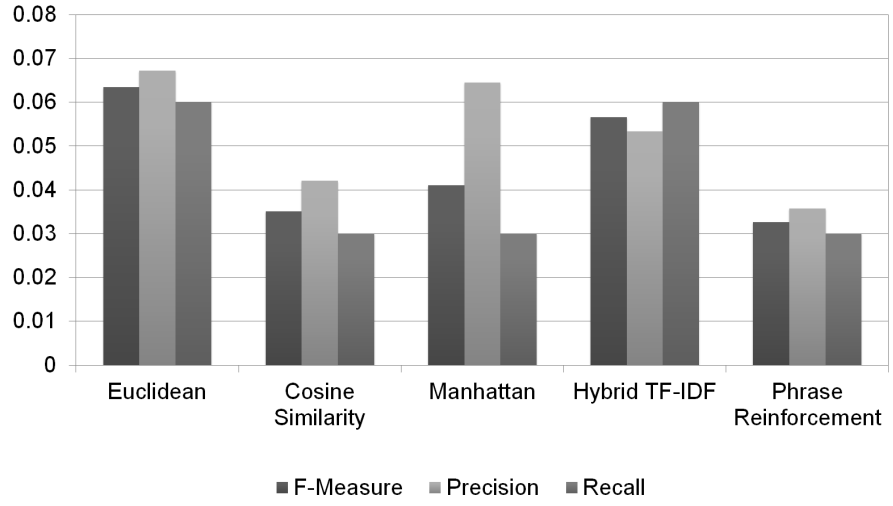

Figure 7.11: ROUGE-2 of UEFA EURO 11^{th} June, 2012, football match

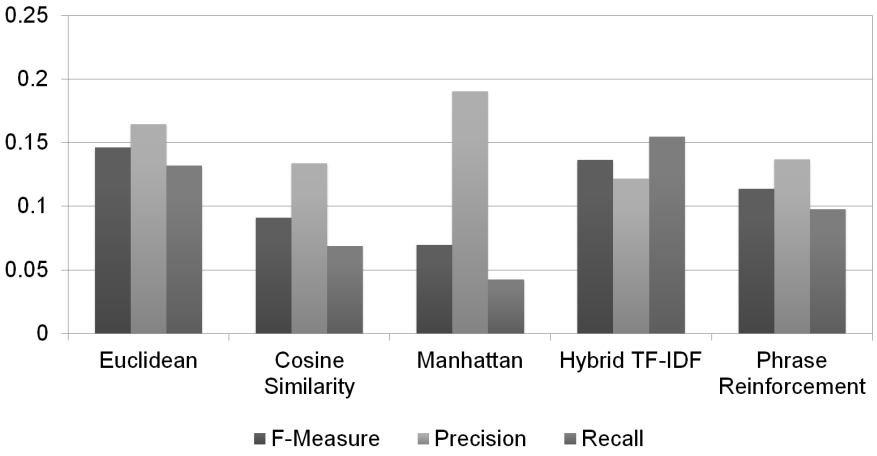

Figure 7.12: ROUGE-SU of UEFA EURO 11^{th} June, 2012, football match

7.3 Content Analysis

After using ROUGE, automatic summary evaluation, we were curious why Cosine Similarity is having highest Precision but lowest Recall and f-measure in IPL cricket match and why Manhattan has highest Precision but lowest Recall and F-measure in UCL football final and in UEFA Euro football match 11th June 2012. To answer our curiosity we did content analysis on all summaries produced by different algorithms and Figure 7.13-7.15 are showing the results. Primary axes are showing the total unigrams and common unigrams of each summary produced by different algorithms. Total unigrams are the number of words in each summary after removing the stop words and number of common unigrams is the words common between post summary and references summary. On secondary axes the results of Unigram Precision and unigram Recall are shown.

Unigram Precision of Post summary is calculated as:

$$P = \frac{w}{t}$$

Where, w is the number of words in post summary that are also in references summary and t is the total number of words in post summary [26]. Unigram Recall of post summary is calculated as:

$$R = \frac{w}{r}$$

Where, w is the number of words in post summary that are also in reference summaries and t is the total number of words references summary [26].

In figure 7.13 Euclidean has highest value of number of common unigrams. For that reason it has highest Recall value. Cosine Similarity has highest Precision because the ratio of total unigrams and number of common unigrams is highest. But it has low Recall than of Euclidean because the number of common unigrams of Cosine Similarity is low than of summary produced by Euclidean. Similarly Manhattan has good value of Precision but it has lowest Recall because the summary produced by Manhattan is shortest among all. Total unigrams of Phrase Reinforcement summary are almost equal of Hybrid Term frequency but the number of common unigrams is less than of Hybrid TF-IDF. For that reason it has low Recall and Precision value than of Hybrid TF-IDF.

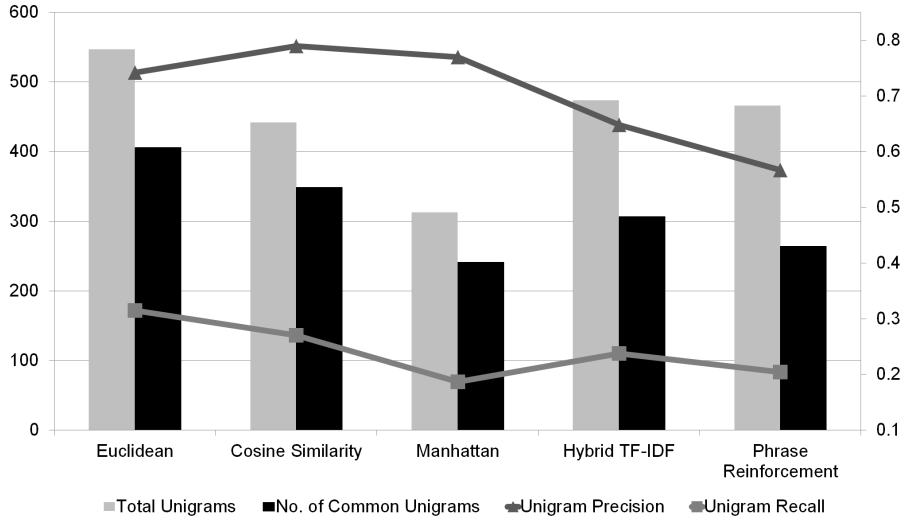

Figure 7.13: Content analysis of IPL 27^{th} May 2012, cricket match

Again in Figure 7.14 it is shown that post summary produced by Euclidean distance has highest number of common unigrams and Recall. It has produced the longest post summary among all. Manhattan has the highest Precision but lowest Recall and it has produced the shortest summary. Again Cosine Similarity has good balance of total unigrams and number of common unigrams. For that reason it has good Precision and Recall value.

Phrase Reinforcement has lowest ratio between total unigrams and number of common unigrams.

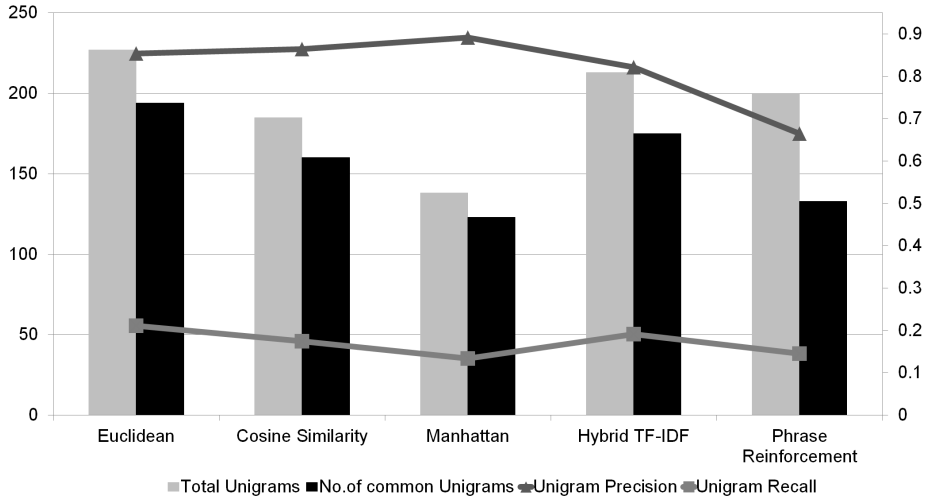

Figure 7.14: Content analysis of UCL 19^{th} May 2012, football match

In figure 7.15 it is shown that the summary produced by hybrid TF IDF is longest. But the ratio of total unigrams and number of common unigrams is better of Post summary produced by Euclidean. In Euclidean summary out of 136 90 words are common words. Whereas, in hybrid TF-IDF out of 167 94 are the number of common words. The number of common unigrams of Hybrid TF-IDF is slightly high than of post summary produced by Euclidean. For that reason Recall of Hybrid TF-IDF is slightly better than of Euclidean. Manhattan has highest Precision and lowest value of Recall, as it has produced the shortest summary among all. This time Cosine Similarity has not good ratio of total unigrams and number of common unigrams, as out of 106 60 words are common. It has low Precision than of Euclidean and Manhattan but it has better Precision than of summary produced by Hybrid TF-IDF and Phrase Reinforcement.

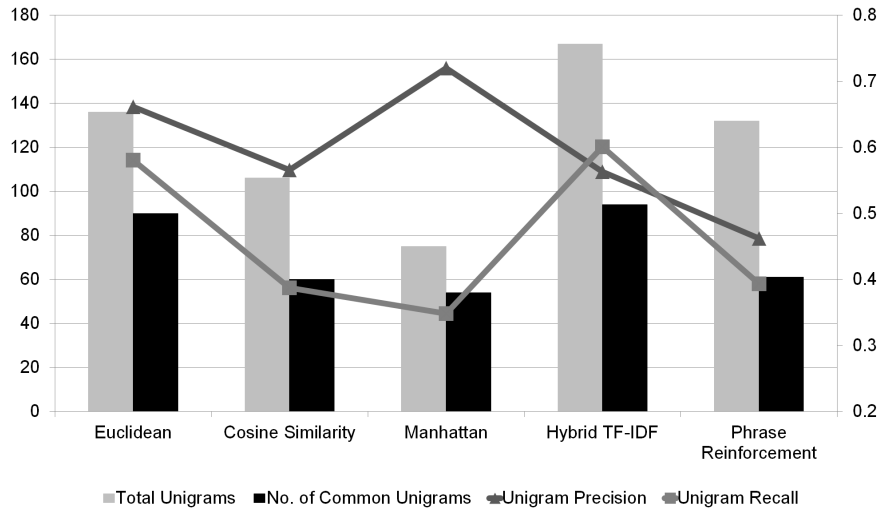

Figure 7.15: Content analysis of UEFA EURO 11^{th} June 2012, football match

Chapter 8

Conclusion and Future Work

In this chapter, conclusions drawn from this study has been presented. The future directions on micro-blog summarization have also been commented upon.

8.1 Conclusion

Summaries produced from Twitter updates are shorter than summaries presented on main stream media. For that reason, results of automatic summary has low Recall value but the high Precision indicates that the summary contains more tweets describing important moments occurred in real event.

Tweets timestamp and tweet words are sufficient features for tweet clustering to identify important moments of a main event. By concatenating these moments an event summary can present highlights of real time event. Post summary produced by clustering using Manhattan distance and Cosine Similarity are short. For that reason they have high Precision but low Recall value.

Clustering using Euclidean distance outperforms Phrase Reinforcement and Hybrid TF-IDF. It improves F-measure in all matches and ROUGE-metrics.

Results showed that clustering using Euclidean distance outperforms Cosine Similarity and Manhattan distance metrics.

8.2 Future Work

In this work clustering using unigrams as features has been performed in future clustering using bigram and tri-gram as feature space can also be

applied.

Post summarization using domain knowledge of sports event and labeling different type of moments, a supervised learning approach can be explored.

Bibliography

[1] http://en.wikipedia.org/wiki/Twitter

[2] http://en.wikipedia.org/wiki/Machine_learning

[3] Beaux Sharifi, Automatic microblog classification and summarization, Master's thesis, University of Colorado at Colorado Springs, 2010.

[4] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita, Summarizing microblogs automatically. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10), 2010.

[5] Sharifi, B. and Hutton, M.A. and Kalita, J. Automatic summarization of Twitter topics, in National Workshop on Design and Analysis of Algorithms, Tezpur, India, 2010, pp. 121128.

[6] Beverungen, G. and Kalita, J. Evaluating Methods for Summarizing Twitter Posts, WSDM11, February 912, 2011, Hong Kong, China.

[7] Beaux Sharifi, Mark-Anthony Hutton, and Jugal K. Kalita, Experiments in Microblog Summarization, In Proceedings of

the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM '10).

[8] Inouye, D. Multiple Post Microblog Summarization, REU Research Final Report, UCCS, JULY 2010.

[9] Inouye, D.; Kalita, J.K.; , "Comparing Twitter Summarization Algorithms for Multiple Post Summaries," Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom) , vol., no., pp.298-306, 9-11 Oct. 2011.

[10] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using Twitter. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (IUI '12). ACM, New York, NY, USA, 189-198.

[11] OConnor, M. Krieger, and D. Ahn. TweetMotif: Exploratory search and topic summarization for Twitter, In Proceedings of ICWSM, 2010.

[12] http://en.wikipedia.org/wiki/Microblogging

[13] http://en.wikipedia.org/wiki/2012_UEFA_Champions_League_Final#Match_summary

[14] http://en.wikipedia.org/wiki/UEFA_Euro_2012_schedule

[15] http://en.wikipedia.org/wiki/Unsupervised_learning

[16] http://en.wikipedia.org/wiki/K-means_clustering

[17] http://www.python.org/

[18] https://pypi.python.org/pypi/tweetstream

[19] http://en.wikipedia.org/wiki/Euclidean_distance

[20] http://en.wikipedia.org/wiki/Cosine_similarity

[21] http://en.wikipedia.org/wiki/Manhattan_distance

[22] http://en.wikipedia.org/wiki/Lemmatisation

[23] http://www.berouge.com/Pages/default.aspx

[24] http://en.wikipedia.org/wiki/Bigram

[25] http://www-nlpir.nist.gov/projects/duc/RM0507/ksj/duc03wgreps

[26] http://en.wikipedia.org/wiki/BLEU

[27] http://en.wikipedia.org/wiki/Long_tail

[28] Shamma, D. A., Kennedy, L., and Churchill, E. F. Tweet the debates: Understanding Community Annotation of Uncollected Sources, In Proc. WSM 2009.

[29] http://en.wikipedia.org/wiki/Precision_and_recall#Definition_.28information_retrieval_context.29

[30] http://en.wikipedia.org/wiki/2012_Indian_Premier_League