Post Summarization of Microblogs of Sporting Events

Mehreen Gillani
National University of Sciences and Technology
Islamabad, Pakistan
10msitmgillani@seecs.edu.pk

Muhammad U. Ilyas
University of Jeddah
Jeddah, Saudi Arabia
usman@ieee.org

Saad Saleh
National University of Sciences and Technology
Islamabad, Pakistan
saad.saleh@seecs.edu.pk

Jalal S. Alowibdi
University of Jeddah
Jeddah, Saudi Arabia
jalowibdi@uj.edu.sa

Naif Aljohani
King Abdulaziz University
Jeddah, Saudi Arabia
southtel2011@gmail.com

Fahad S. Alotaibi
King Abdulaziz University
Jeddah, Saudi Arabia
fsalotaibi@kau.edu.sa

ABSTRACT

Every day 645 million Twitter users generate approximately 58 million tweets. This motivates the question if it is possible to generate a summary of events from this rich set of tweets only. Key challenges in post summarization from microblog posts include circumnavigating spam and conversational posts. In this study, we present a novel technique called lexi-temporal clustering (LTC), which identifies key events. LTC uses k-means clustering and we explore the use of various distance measures for clustering using Euclidean, cosine similarity and Manhattan distance. We collected three original data sets consisting of Twitter microblog posts covering sporting events, consisting of a cricket and two football matches. The match summaries generated by LTC were compared against standard summaries taken from sports sections of various news outlets, which yielded up to 81% precision, 58% recall and 62% F-measure on different data sets. In addition, we also report results of all three variants of the recall-oriented understudy for gisting evaluation (ROUGE) software, a tool which compares and scores automatically generated summaries against standard summaries.

1. INTRODUCTION

1.1 Background and Motivation

Social networking sites have seen high growth rates in users the past few years. By 2013, 1.73 billion people were using various social networking services worldwide [5]. Facebook and Twitter are among the leading social network services transforming the world into a global village. The success of social networking sites is attributed to the ease in information dissemination, communication and entertainment. In 2014, Twitter had more than 645 million users generating 58 million tweets per day and 2.1 billion search queries daily [30]. These millions of tweets are a major source of news in which posts / information that receives more votes from users (in the form of 'favorites', 'retweets' and 'quotes') is pushed to the forefront of more people's Twitter feeds [16]. In this paper, we investigate a new approach to summarizing long running events, such as sporting matches, by selecting from millions of Twitter status updates. Such a summary of tweets related to any subject is referred as post summarization of microblogs.

Post summarization of microblogs is a challenging task because the incoming stream of tweets is very noisy due to the presence of tweets that are poorly composed, using non-standard language, personal conversations and / or contain irrelevant information (spam taking advantage of trending tags). A study by Pear Analytics [14] highlighted the low signal-to-noise-ratio of twitter by showing that 40.1% of tweets contained pointless information, 37.6% of tweets were conversational while only 3.6% of tweets highlighted mainstream news. Therefore, one of the key challenges lies in denoising the incoming tweetstream, i.e., separating tweets that will contribute towards the making of a good event summary from those containing little to no information of interest.

1.2 Limitations of Prior Art

Several tools, such as `Twitterfall.com`, `keyhole.co` and Twitter's own search tool, already summarize bodies of tweets one way or another. First and foremost, Twitter's own search tool returns results containing the given search terms and favoring tweets with high 'retweet' and 'favorite' counts. Other summarization tools like `Twitterfall.com` provide a steady stream of tweets containing the given search terms at an approximate rate of one tweet every 2-3 seconds. Another summarization tool, `keyhole.co`, returns tweets from a specified period of time using one of three possible prioritizations: 1) favoring tweets with higher retweet counts, 2) favoring recent tweets or 3) favoring tweets by Twitter users with high Klout scores [11]. These results are augmented by some descriptive statistics derived from the metadata of all discovered tweets containing the given search terms.

The majority of studies on post summarization limit their research to the importance of tweets using tweet length, choice of words, spikes in traffic volume and relevance to a given topic. Khan *et al.* [15] separated tweets using the latent Dirichlet allocation (LDA) algorithm and used words (unigrams) to determine their significance. Tweets were

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3038912.3038914

collected at periodic time intervals. Khan *et al.* reported achieving precision of up to 81.6% and recall rates of up to 80%. Chakrabarti and Punera [3] and Marcus, Bernstein, Badar, Karger, Madden and Miller [18], achieved up to 63% precision and 61% recall rate. However, like Khan *et al.*, the scope of these studies was limited to post summarization at periodic intervals. These above techniques are ill suited to post summarization of sporting and other events in which moments of interest may occur at any time during the event, not at regular intervals. Furthermore, only limited performance was achieved by these preceding techniques because they all estimate the importance of a tweet from characteristics of the text of the tweet itself.

1.3 Proposed Approach

We propose a novel approach called lexi-temporal clustering (LTC) to generate post-event summaries of sporting events from microblogs. LTC produces a summary of sporting events composed of individual microblog posts put in chronological order, each describing an important *moment* that occurred during the event. LTC follows three steps: First, the occurrence of critical moments is detected which are identified by local maxima in the posting rate function of microblog posts. We set a significance threshold based on the mean and standard deviation of microblog posting rate function over the duration of the sporting event. Times at which the microblogging rate function exceeds the threshold indicate the occurrence of key moments that are potential candidates for inclusion in the summary. Second, microblog posts are denoised by filtering irrelevant and low information posts and perform lemmatisation, stemming and normalization. Finally, the event is summarized by selecting microblog posts representative of the most important moments using k-means clustering. We explore several distance measures in k-means clustering including Euclidean, Manhattan and cosine similarity. Summarization is also performed using Hybrid TF-IDF [19] and phrase reinforcement [24] approaches to test the performance of k-means clustering with previous state of the art schemes. The performance of LTC is evaluated by comparing against manually created standard summaries from the mainstream press, A) in terms of detection and false alarm rates of moments during events, and B) automatic post summarization evaluation tools ROUGE-1, ROUGE-2 and ROUGE-SU [17] which report recall, precision and F-measure rates.

1.4 Experimental Results and Findings

We collected three data sets from Twitter to test and validate our approach. These data sets include tweets of a cricket match of the Indian Premier League (IPL), a soccer match of the UEFA Champions League (UCL) and a soccer match of the UEFA European Championship (EURO). Our performance metrics of recall, precision and F-measure suggest that k-means clustering technique using Euclidean distance gives the best performance. Up to 58% recall and 81% precision are obtained for our LTC based post summarization. Performance results over various data sets verify that LTC based post summarization outperforms all other strategies.

1.5 Key Contributions

The major contributions of this paper are as follows:

1. We proposed a novel lexi-temporal clustering (LTC) employing k-means clustering using three different distance metrics, including Euclidean, cosine similarity and Manhattan distance. Post summarization is also performed with hybrid TF-IDF and phrase reinforcement approaches to summarize important moments of any event (Sec. 3).

2. We report results of manual evaluation metrics on LTC to detect significant moments during matches, using manually authored match summaries as benchmarks. Furthermore, we also report several automatic evaluation metrics from ROUGE-1, ROUGE-2 and ROUGE-SU tools to evaluate LTC's automatically generated summaries against the same benchmark summaries (Sec. 4).

2. PROBLEM FORMULATION

The problem considered in this paper is, *"How to build a short chronological list of microblog posts from a very large set that accurately describe the most important moments of an event?"* Important events are the critical moments of the match. Important events for all matches are available on news websites, through which we can compare our results. The problem is formulated as follows:

- Given an event E that was tweeted about in a set of P microblog posts, let k be the number of important moments during the event E.
- If p_i and p_j are microblog posts, select a set of k tweets P_k from P such that, $p_i \neq p_j$ where, $p_i, p_j \in P_k$.
- P_k is the set of tweets summarizing k moments in E.

3. TECHNICAL APPROACH

In this section, details of our proposed scheme lexi-temporal clustering (LTC) are presented. LTC includes the identification of critical moments, cleaning of the data set, feature selection, feature extraction and microblog summarization.

3.1 Data Set Collection

For the collection of microblog posts we chose Twitter because it has a relatively open policy with regards to data collection and has a large number of users throughout the world. We used the Python tweetstream 1.1.1 package for data collection, which provides access to Twitter's streaming Application Programming Interface (API) [9]. We collected tweets for three popular sporting events using their respective event hashtags namely $\#ipl$, $\#ucl$ and $\#Euro2012$, respectively. The details of the events for which tweets were collected are listed below:

1. Indian Premium League (IPL) T20 cricket final match between KKR (Kolkata Knight Riders) and CSK (Chennai Super Kings) on 27^{th} May, 2012, at MA Chidambaram Stadium, Chepauk, Chennai, India[1].

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

[1] `http://www.iplt20.com/`

[2] `http://www.uefa.com/uefachampionsleague/`

3. UEFA European Championship (EURO) football match being held on 11^{th} June 2012, between France and England at Donbass Arena, Donetsk, Ukraine[3].

Ground Truth Data Set: Summaries produced by various news sources have been used as the reference summaries of IPL, UCL and EURO matches [8], [12], [31], [28], [27], [7]. Performance of LTC and other summarization techniques has been compared with these reference summaries , both manually and using ROUGE, to compute performance metrics.

Tab. 1 lists some descriptive statistics of the data sets including total tweets, maximum and minimum tweets per minute, unique words and unigram statistics using Euclidean, cosine similarity, Manhattan, hybrid TF-IDF and phrase reinforcement. Comparison over various matches shows that IPL has a lower number of total tweets, maximum and minimum tweets per minute compared to UCL and EURO. We explain the difference in the number of tweets during the IPL cricket match and the two football matches by the difference in Twitter penetration rate in India on the one side, and Germany and Ukraine on the other.

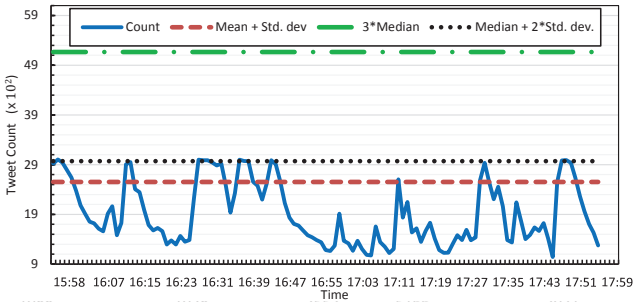

Figure 1: Minimum threshold using different methods for EURO match.

3.2 Identifying Important Moments

Important moments in a match were identified by local maxima in the number of tweets posted per minute. Volume surges in Twitter activity correlate with the occurrence of important moments in a match *e.g.* for a cricket match: a wicket, a six, end of first inning etc. Similarly, in football these peaks have been observed to correlate with important moments such as goals, half time, match start and end etc. Let $T(i)$ denote the number of tweets during minute i, then a local maxima occurs at a minute i if, $T(i-1) < T(i) > T(i+1)$.

In addition to local maxima, several studies suggest that a maxima needs to rise above a threshold to qualify it as an important moment in a given event. This threshold can be defined in terms of the median (M), mean (μ) and standard deviation (σ) of traffic volume [20][22]. Nichols *et al.*'s [20] approach of using $M \times 3$ was tested but almost all of the moments went undetected. Another approach of using a threshold of $M + 2 \times \sigma$ was tested but only a small number of important moments were detected [20]. Finally, $\mu + \sigma$ was tested which identified maximum number of important moments [22]. All important moments identified using various preceding thresholds were compared with the manual evaluation of important events (see Tab. 2).

Fig. 1 shows the minimum threshold by all three methods on the tweet rate function graph of the EURO match. There were 18 important moments in the EURO match. For a threshold level of $\mu + \sigma$, 12 important moments (local maxima above threshold level) were identified, while for a threshold of $M \times 3$ and $M + 2 \times \sigma$, 0 and 7 moments were identified, respectively. Refer to Tab. 2 for the other IPL and UEFA matches. Similar trends were obtained for IPL and UCL matches. This is not surprising since the $\mu + \sigma$ threshold is the lowest of the three considered here. Based upon these findings, we decided to use $\mu+\sigma$ as the threshold level across all data sets. Important moments missed by our threshold affect our precision, recall and F-measure.

3.3 Data Cleaning

The collected tweet data sets contained a lot of noise in the form of chatter irrelevant to a match summary. For feature selection and analysis, the data set was passed through a number of cleaning stages to reduce noise. Various stages of data cleaning are described as follows.

Removal of Irrelevant Tweets: A number of tweets were filtered out of data sets based on their contents. Based on a detailed inspection of the data sets, we developed the following rules for filtering out irrelevant and low information tweets. Tweets falling into one or more of the following categories were removed: (1) Majority ($> 50\%$) non-English words, (2) Containing URLs, (3) Reply / direct message tweets *i.e.* tweets containing '@', and (4) Duplicate tweets. Clustering techniques were applied to data sets before and after tweet removal. Results suggested that irrelevant tweets are a major source of noise in post-event summaries and that denoising using the simple 4-step procedure above significantly improved the output. Similarly, all duplicate tweets were removed except the one with the latest timestamp.

Removal of Irrelevant Content: We removed stop words, # and other special characters in order to reduce the level of noise in the tweets.

Word Normalization: A number of tweets contained unnecessary repetition of words by users. To simplify the complications, we have normalized the words and removed unnecessary repetition of letters, *e.g.* 'goooooal' to 'goal.'

Lemmatization and Stemming: We performed lemmatization and stemming using the Natural Language ToolKit (NLTK) [2] and stemming.porter2 [4]. Lemmatization groups different forms of a word as a single word. Stemming replaces related words with a stem or root word. Unlike stemming, lemmatization chooses the appropriate lemma by considering the context and part of speech of a term.

3.4 Feature Selection

A number of features were observed in the collected data set. Based on the characteristics of important posts in an event, we suggest the use of following features to perform clustering, (1) Time stamp of tweet, (2) Words contained in a tweet. Only two characteristics have been found sufficient for detection of important events. The tweet timestamp plays a very important role. All tweets about an important moment tend to cluster immediately after that moment's occurrence. Thus in a lexical-temporal feature space, the temporal axis time serves to separate tweets describing similar moments that are, however, separated by time. Our

[3] `http://www.uefa.com/uefaeuro-finals/`

Match	Total tweets	Max tweets/ min.	Min tweets/ min.	Unique words	Unigram Statistics									
					Euc		CS		Man		HTI		PR	
					TW	CW	TW	CW	TW	CW	TW	CW	TW	CW
IPL	46370	1452	85	3423	547	406	442	349	313	241	474	307	466	264
UCL	226109	3029	488	2420	227	194	185	160	138	123	213	175	200	133
EURO	233288	3005	1044	3618	136	90	106	60	75	54	167	94	132	61

Table 1: Data set characteristics containing total unigrams (TU) and common unigrams (CU) for Euclidean (Euc), cosine similarity (CS), Manhattan (Man), hybrid TF-IDF (HTI) and phrase reinforcement (PR).

analysis shows that tweets describing an important moment in an event contain similar fields, *e.g.* for a cricket match, name of the player, wicket and the type of play, four, six etc. Simply put, among the tweets that contribute to a peak in traffic volume above the chosen threshold of $\mu + \sigma$, we seek to identify a tweet close to the cluster centroid in lexical feature space.

3.5 Feature Extraction

We perform clustering using tweet timestamps metadata and the words tweets are composed of. Tab. 1 tabulates, among other statistics, the total number of unique words for IPL, UCL and EURO matches. Thousands of unique words decrease the efficiency of clustering but act as a source of noise for detection of critical moments. So the most important words have been filtered based upon the words frequency. Fig. 2 is a semi-log plot of the histogram of unique words after cleaning the UCL match data. A threshold of 2.18 on a $\log_{10}$ scale has been selected to determine the lower bound of the local maxima.

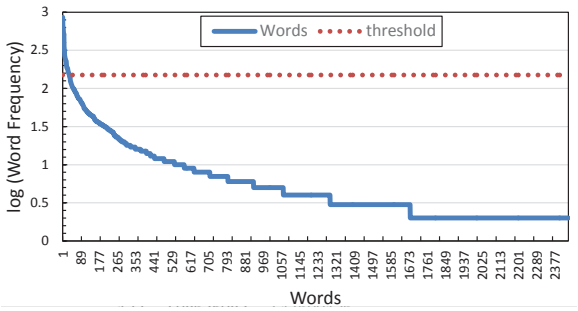

Figure 2: Word frequency of UCL match.

3.6 Microblog Summarization

Summarization of tweets based upon important moments is performed using k-means clustering and Hybrid TF-IDF and Phrase Reinforcement approaches. Various distance metrics used for clustering are (1) Euclidean, (2) Cosine Similarity, and (3) Manhattan. After clustering tweets, the 2 tweets closest to each cluster centroids are generated at the output. Tab. 4 presents the summaries generated using different clustering techniques for IPL matches. Results show that using Manhattan distance tends to favor the short tweets as outputs. Cosine similarity and Euclidean distance consistently identify medium length tweets at their outputs. Hybrid TF-IDF and phrase reinforcement produce the longest length tweets at their outputs, because they assigned more weight to tweet length than its contents. Only marked moments have been identified. In Tab. 4, moments of IPL match at time instants 15 : 41, 15 : 42 and 17 : 46 have not been identified by any clustering using any distance metric. We have not included tables similar to Tab. 4 for the UCL and EURO matches for brevity reasons. Similarly, moments of UCL match occurring at 18 : 45, 21 : 00, 21 : 27, 21 : 28 and 21 : 29 were missed. It is pertinent to mention that not all critical moments in close succession were detected because local maxima requires three values to identify the peak value. For the EURO data set, the yellow card moment goes undetected by all clustering schemes. The number of tweets for yellow card moments and other undetected single instant moments were very low, so moments were missed by all clustering schemes.

Performance of different clustering schemes for IPL, UCL and EURO matches is shown in Tab. 3. Results show that maximum performance for IPL match is obtained for Euclidean and Cosine Similarity which identify 20 of 24 moments. For UCL, all clustering techniques identified 10 out of 18 moments. For EURO, 7 out of 8 moments were identified for all clustering techniques. Comparison shows that clustering using Euclidean distance identified the most moments compared to other metrics.

Match	**Important Moments**	$\mu + \sigma$ [22]	**M×3** [20]	**M + 2** $\times\sigma$ [20]
IPL	25	23	1	6
EURO	18	12	0	7
UCL	8	7	0	5

Table 2: Important moments identified by various methods using mean (μ), standard deviation (σ) and median (M).

4. EVALUATION

In this section the effectiveness of LTC has been evaluated using manual and automatic summary evaluation of the data set of three matches from Twitter.

Baseline Techniques: ROUGE-1 and ROUGE-2 use 1-gram and 2-gram based co-occurrence statistics [17]. ROUGE-SU uses skip bigram and unigram-based co-occurrence statistics.

Performance Metrics: Recall, precision and F-measure are the key measures used for the evaluation of LTC. The following sections describe the results of LTC using ROUGE relying on reference summaries of the matches.

Match	Imp. moments	Euclid.	Cos. sim.	Manh.	Hyb. TF-IDF	Phrase reinf.
IPL	24	20	19	14	14	15
UCL	10	8	8	8	8	8
EURO	8	7	7	7	7	7

Table 3: Important moments identified by IPL, UCL and EURO.

Time	Important Moments	E	C.S.	M.	H.	P.R.
14:32	Match started	✓	✓	✓		
14:59	Lee to M Hussey, SIX.	✓	✓	✓		
14:59	Lee to M Vijay, FOUR	✓	✓			
15:25	Bhatia to Vijay, OUT, slower one and what a catch	✓	✓		✓	✓
15:33	Kallis to Raina, SIX	✓	✓	✓		
15:41	Pathan to Raina, SIX					
15:42	Pathan to Raina, SIX					
15:43	1. Pathan to Raina, FOUR 2. MEK hussey 50	✓	✓	✓	✓	✓
15:54	1. Narine to Raina, SIX, 2. Fifty Raina	✓	✓	✓	✓	✓
15:58	Kallis to Hussey, OUT	✓	✓	✓	✓	✓
16:06	Narine to Raina, SIX	✓				
16:11	Shakib Al Hasan to Raina, OUT	✓	✓	✓	✓	✓
16:12	End of first inning	✓	✓	✓	✓	✓
16:30	Hilfenhaus to Gambhir, OUT	✓	✓	✓	✓	✓
16:51	Ashwin to Bisla, SIX	✓	✓	✓		
17:16	Jakati to Bisla, FOUR (Twice)					
17:21	Ashwin to Bisla, SIX,					✓
17:36	Morkel to Bisla, OUT	✓	✓		✓	✓
17:45	Bravo to Shukla, OUT	✓	✓	✓	✓	✓
17:49	Bravo to Kallis, SIX	✓	✓		✓	✓
17:51	Ashwin to Pathan, OUT	✓	✓	✓	✓	✓
18:02	Hilfenhaus to Kallis, OUT	✓	✓	✓	✓	✓
18:06	9 required on 6 Balls	✓	✓		✓	✓
18:15	Bravo to Tiwary, FOUR, KKR won the IPL	✓	✓	✓	✓	✓
Total Detected Moments		20	19	14	14	15

Table 4: Important moments of IPL match identified by different methods.

4.1 Recall

In the context of information retrieval, recall is the fraction of correctly classified instances to the total number of identified instances. Tab. 5 shows the recall of key moments of IPL match. Results show that clustering using Euclidean distance has highest recall value for moments of all categories. Hybrid TF-IDF and phrase reinforcement have lower recall statistics at the moments capturing "Fours" and "Sixes". Clustering using cosine similarity gave lower recall statistics than Euclidian but higher recall than all other clustering techniques.

Key Moment	Euclid.	Cos. Sim.	Manhttn.	Hyb. TF-IDF	Phrase Reinf.
Match start	1	1	1	1	1
Innings End	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 5: Recall of key moments of IPL match using various clustering techniques.

Tab-6 shows the recall of key moments of UCL and EURO matches identified by different clustering techniques. Results show that almost the same recall statistics are obtained for all clustering techniques. "Game Start" moment has a recall of 0.5 because all methods have successfully identified the start of EURO match but the start of UCL match could not be identified. "Yellow card" moment has a recall of 0.33 because only a small number of tweets are available for this moment. No red card moment occurred in either UCL or EURO.

Key Moment	Euclid.	Cos. Sim.	Manhttn.	Hyb. TF-IDF	Phrase Reinf.
Goal	0.78	0.78	0.78	0.78	0.78
Red Card	✗	✗	✗	✗	✗
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
Disallwd. Goal	1	1	1	1	1
Match End	1	1	1	1	1

Table 6: Recall of key moments of UCL and EURO using various clustering techniques.

Fig. 3, Fig. 4 and Fig. 5 show the recall statistics from ROUGE-1, ROUGE-2, ROUGE-SU and manual evaluation (labeled just 'LTC') of LTC for IPL, UCL and EURO matches, respectively. LTC outperformed other techniques for all matches using any distance metric. A closer investigation of individual matches suggests that Euclidean distance based clustering gave better performance for IPL and UCL matches. For EURO, hybrid TF-IDF gave slightly better performance gain than Euclidean distance. Manhattan gave

the lowest performance for all matches in comparison to all other clustering schemes.

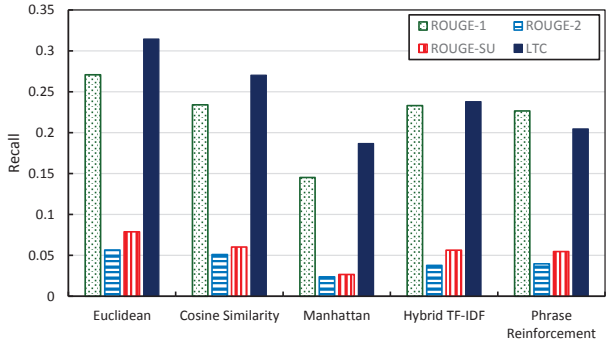

Figure 3: Recall of IPL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

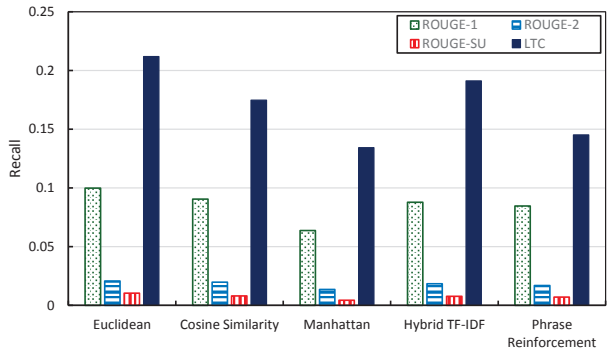

Figure 4: Recall of UCL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

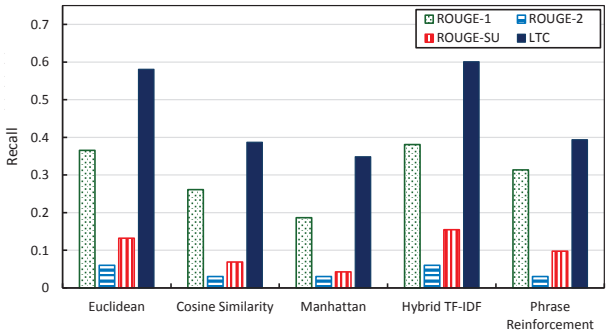

Figure 5: Recall of EURO match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

4.2 Precision

In the context of information retrieval, precision is the fraction of retrieved instances that are relevant. Fig. 6, Fig. 7 and Fig. 8 show the precision statistics from ROUGE-1, ROUGE-2, ROUGE-SU and manual evaluation (labeled just 'LTC') of LTC for IPL, UCL and EURO matches, respectively. LTC provided better performance compared to automatic post summarization schemes in all scenarios. Individual analysis of various matches shows that Phrase Reinforcement gives the worst performance compared to other distance metrics used for clustering. For IPL, cosine similarity gave highest precision of 79% followed by Euclidean distance clustering yielding 74% precision. For UCL, Manhattan gives 89% precision followed by cosine similarity and Euclidean providing 86% and 85% precision, respectively. For EURO, Manhattan provides 72% precision followed by Euclidean providing precision of 66%.

A comparison of precision and recall for various clustering schemes suggested that Euclidean distance gives best performance for all matches. High precision implies tweets selected for the post-event summary are covering the important moments of the event. Conversely, low precision suggests the presence of more noisy tweets in the post summary. Clustering using cosine similarity has the second highest value of precision in IPL and EURO matches.

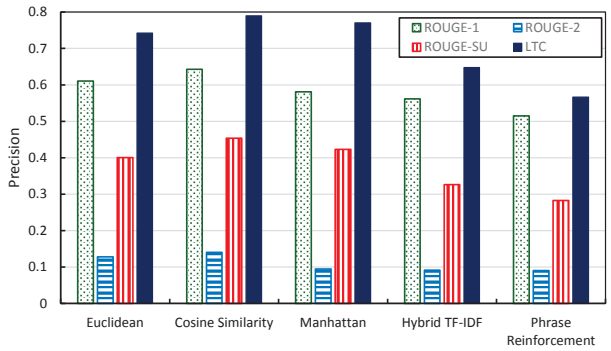

Figure 6: Precision of IPL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

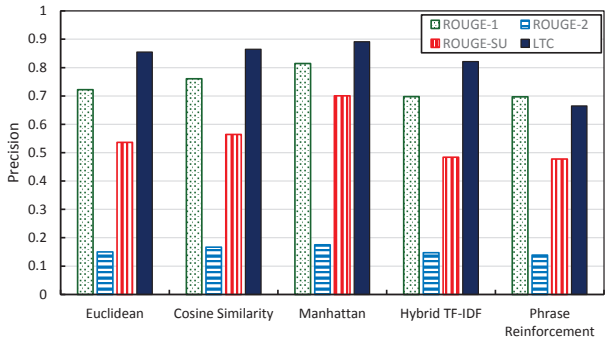

Figure 7: Precision of UCL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

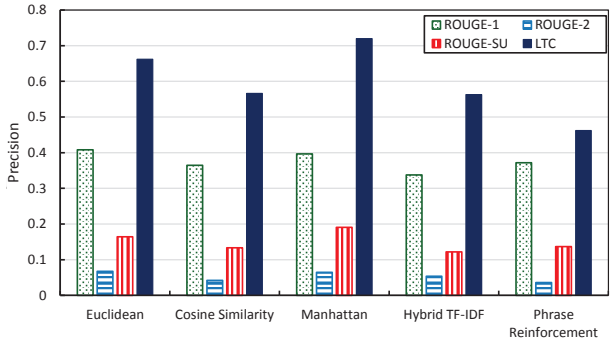

Figure 8: Precision of EURO match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

4.3 F-measure

F-measure is used to check the accuracy of the test based upon the precision and recall measures of various matches. Fig. 9, Fig. 10 and Fig. 11 show the F-measure statistics from ROUGE-1, ROUGE-2, ROUGE-SU and manual evaluation (labeled just 'LTC') of LTC for IPL, UCL and EURO matches, respectively. Results showed that automatic post summarization techniques gave poor performance relative to LTC. An investigation of various matches showed that Euclidean gave better performance for IPL with an F-measure of 44%, followed by cosine similarity with F-measure of 40%. For UCL, Euclidean and cosine similarity gave an F-measure of 34% and 29%, respectively. For EURO, Euclidean gave best performance with an F-measure of 62% followed by hybrid TF-IDF with an F-measure of 58%. Performance comparison of various distance metrics for clustering over different matches showed that Euclidean outperformed all other clustering schemes.

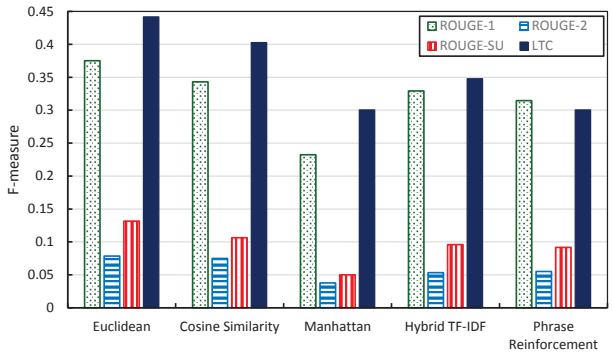

Figure 9: F-measure of IPL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

5. DISCUSSION

In this section, we perform content analysis to identify ambiguities in performance gains of different clustering techniques across different matches. Our results for IPL match from Sec. 4 suggest that cosine similarity has the highest precision with minimum recall and F-measure. For UCL and EURO matches, Manhattan has highest precision with lowest recall and F-measure. Individual content analysis of

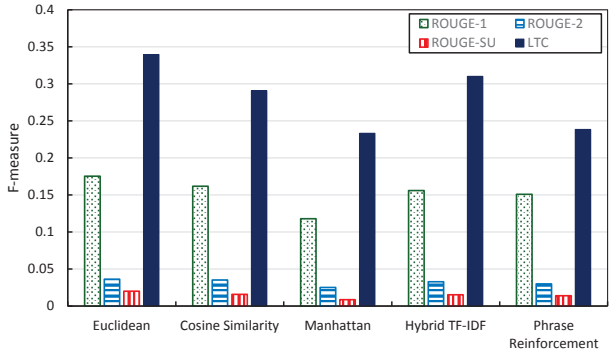

Figure 10: F-measure of UCL match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

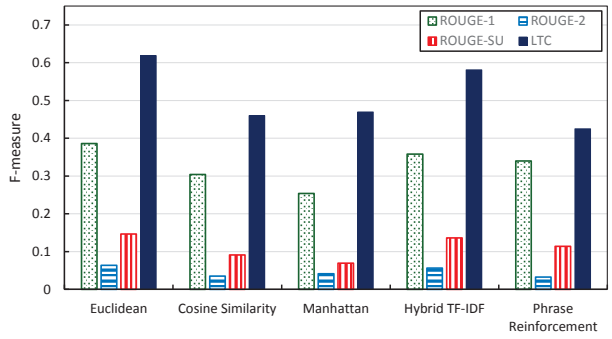

Figure 11: F-measure of EURO match using ROUGE-1, ROUGE-2, ROUGE-SU and LTC.

various matches can identify these anomalies. Observation of total unigrams and common unigrams for various matches have been found helpful in the investigation of such varying performance gains. Total unigrams represent the number of words (after data cleaning) in all tweets of a given match in the collected data set. Common unigrams represent the tweet words of the collected data set matching the tweets of ground truth data set of important moments. For analysis, unigram precision P and recall R are given by $P = w/t$ and $R = w/r$, respectively. Here, w represents common unigrams, t represents the total unigrams and r represents the number of words in the reference summary.

5.1 IPL match

Fig. 12 presents the total and common unigram statistics for the IPL match along with the precision and recall for different distance metrics. Results show that Euclidean has 406 common words out of 547 total words which yields a high recall. Cosine similarity has the highest precision because the ratio of total unigrams to common unigrams is highest (0.79). Recall of cosine similarity is lower than that of Euclidean because cosine similarity has only 349 common unigrams. Similarly, Manhattan has a precision comparable to other schemes but it has the lowest recall of 0.19 because Manhattan produces the shortest summary. Phrase reinforcement and hybrid TF-IDF produce summaries of roughly similar length, but the number of common unigrams are greater for phrase reinforcement than hybrid TF-IDF.

Hence, phrase reinforcement has lower recall and precision than hybrid TF-IDF.

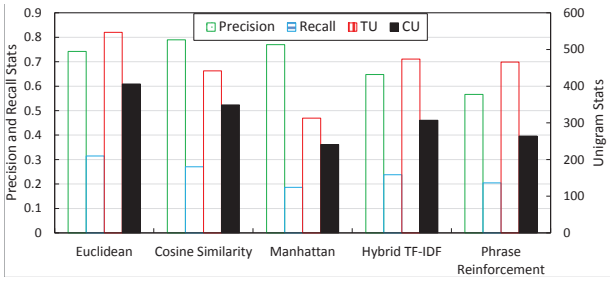

Figure 12: Content analysis of IPL match displaying Total Unigrams (TU) and Common Unigrams (CU).

5.2 UCL match

Total unigrams and common unigram statistics for UCL match along with the precision and recall are shown in Fig. 13. Similar to previous trends, the post summary produced by Euclidean distance has the highest number of common unigrams (194) which makes for a 21% recall. The post-event summary produced by Euclidean is the longest among all summaries. Manhattan has the highest precision but lowest recall because it produced the shortest summary. Cosine similarity has sufficient precision and recall statistics because total unigrams and common unigrams differ by only 25 unigrams. Total unigrams and common unigrams are in balance with each other for cosine similarity. Phrase reinforcement has the lowest ratio of 0.66 between total unigrams and common unigrams which gives only 14.5% recall and 66.5% precision.

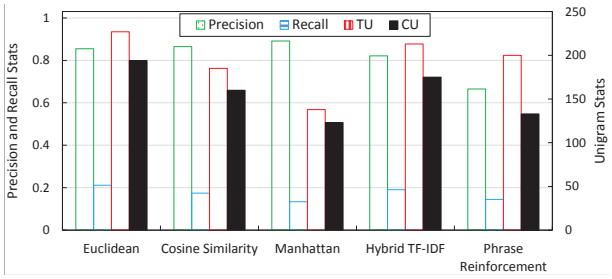

Figure 13: Content analysis of UCL match displaying Total Unigrams (TU) and Common Unigrams (CU).

5.3 EURO match

Fig. 14 presents the precision and recall along with the total and common words for EURO match. Results show that Hybrid TF-IDF produces the longest summary but the ratio of total unigrams and common unigrams is greater than the ratio of Euclidean. In Euclidean summary, 90 out of 136 words are common words. In hybrid TF-IDF, 94 out of 167 words are the common words. The number of common unigrams of Hybrid TF-IDF are slightly larger than the post summary produced by Euclidean which results a higher recall to Hybrid TF-IDF than Euclidean. Manhattan has highest precision and lowest recall because it produces the shortest summary. For EURO, cosine similarity has a small ratio between total unigrams and number of common unigrams, as 60 out of 106 words are in common. Cosine similarity has a lower precision than Euclidean and Manhattan but has better precision than hybrid TF-IDF and phrase reinforcement.

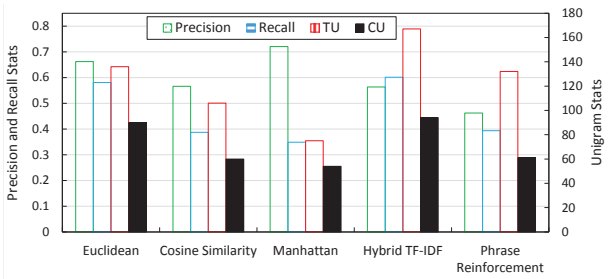

Figure 14: Content analysis of EURO match displaying Total Unigrams (TU) and Common Unigrams (CU).

5.4 Comparison with predecessor schemes

In this subsection, we present a comparison of previous state-of-the-art schemes with LTC. It is worth mentioning that performance metrics vary in different studies but we would focus on the precision, recall and F-measure for different studies. Tab. 7 presents a comparison of precision, recall and F-measure of LTC versus previous schemes. Performance of Marcus *et al.'s* [18] algorithm is dependent upon the data set which degrade to only 14% precision and 77% recall for certain data sets. Khan *et al.* [15] reported better precision and recall numbers for their given data set, but they sampled tweets at uniform intervals of 1 minute without bothering to identify important moments during an event. Performance of all other post summarization approaches is less than the LTC.

Approach	Precision (%)	Recall (%)	F-measure (%)
[15]	85	80	–
[3]	63	52	–
[18]	14-95	77-100	–
[25]	31	30	–
[20]	32	20	23
[10]	18-41	18-41	–
LTC	**81**	**58**	**62**

Table 7: Comparison of LTC with previous schemes

6. RELATED WORK

A number of studies have focused on the post summarization of microblogs using different approaches. Sharifi *et al.* [25] trained a naïve Bayes classifier to classify posts based on categories of Google directory (Art and entertainment, business, food and drink, computer game, health, politics, science, sports, technology, world). Sharifi *et al.* proposed two algorithms to perform automatic post summarization. The phrase reinforcement algorithm used frequently occurring terms to build graphs from posts. The path containing highest weight was selected as the post summary [23]. Another technique, the extended term frequency - inverse

document frequency (TF-IDF) algorithm was proposed for event summarization using normalization of long posts to get higher weights [24] [25]. The automatic summaries produced by these algorithms were compared with manual summaries and significant performance gains were reported [23] [20]. Random selections of posts and length of sentences were used as baseline results [23] [25]. Sharifi *et al.* reported the metrics of recall-oriented understudy for gisting evaluation (ROUGE) for evaluation [23] [24] [25]. Spam posts containing long tweets with irrelevant material are given higher weight, which degrades the precision and recall statistics.

Use of tweet contents has been a major focus for post summarization of important events. Clustering of microblogs is inherently difficult because of the use of non-dictionary words and limited number of features. This is particularly challenging on the Twitter microblogging platform because posts or *tweets* are limited to a maximum 140 characters. Beverungen and Kalita [1] explored the effects of post normalization, noise reduction and improved feature selection on clustering performance. They concluded that post normalization slightly degrades the clustering performance while gap statistic technique performs better when determining the correct number of clusters for posts. They showed that noise reduction improves clustering performance whereas features based on tri-grams outperform uni-grams, bi-grams and term expansion.

Several studies have focused on post summarization while achieving diversity of tweet topics inside tweet clusters. Tweets on a specific topic often contain several subtopics or themes. For multi-post summaries, Inouye and Kalita [13] used the hybrid TF-IDF algorithm to assign individual weights to each post and select the top N posts, where each post should add new information to the small corpus of selected tweets. They developed two algorithms, bisecting k-means and k-means++, for picking the most informative tweets to build the summary. However, k-means++ was not designed, nor has it been tested, for applications where the goal is to generate a summary of key moments during a long running event. To that end, Singhal [26] demonstrated that the cosine similarity measure yields good results in the detection of subtopics in a set of N posts. Simple frequency based algorithms *i.e.* hybrid TF-IDF and sum-basic summarizer give better performance for post summarization than traditional MEAN, TextRank [19] and LexRank algorithms [6]. These algorithms are limited in performance owing to their complexity when the number of tweets identified for the summary might be quite large.

Several investigations focused on spikes in tweet posting rate functions and tweet terms to identify important events. Nichols *et al.* [20] proposed an extended phrase reinforcement algorithm to produce journalistic summaries of special events using Twitter posts. The algorithm is based on the intuition that a sudden increase in number of tweets implies the occurrence of an important event. It selects highly weighted graph phrase from each spike of the event. O'Connor *et al.* [21] developed an exploratory search application for Twitter topics, which cluster tweets into topics based upon terms and keywords frequency. However, the overlapping use of terms in different contexts and a small number of tweets can significantly degrade the performance of this algorithm. Marcus *et al.* [18] developed the application "Twitinfo" to summarize important events using tweets. First, historical events were identified based on the number of tweets in reference to the historical mean. Next, important tweets pertaining to these events were filtered based upon the five tweet words common in the majority of the tweets. A sentiment score was also identified based upon the selection of words in tweets. Precision and recall over various data sets varied between $14 - 95\%$ and $77 - 100\%$, respectively.

Recently, Khan *et al.* [15] resolved the post summarization problem using a graph retrieval algorithm. Multiple subtopics are separated using LDA algorithm. Then, the top two tweets are collected for each subtopic based upon the words found in each tweet. Khan *et al.* sampled the tweets at uniform intervals of time, irrespective of the event.

Hu *et al.* [10] dealt with a different but related problem of identifying user's interest in various tweet topics and developed an analytical model for topical clustering. Based on the most common words, top two tweets for every topic are generated along with the topic popularity. Tao *et al.* [29] analyzed a different but related problem of duplicate tweet detection. They presented a framework for duplicate tweet detection using external web sources and specified the similarity patterns in duplicate tweets. Chakrabarti and Punera [3] analyzed the tweet summarization problem by developing a hidden Markov model (HMM). They separated various tweet events into subtopics using tweet frequency and HMM to identify key moments.

In our research, we identify the critical moments, clean the data set, identify and analyze features, use various clustering metrics along with post summarization techniques. We use summaries available from popular news websites as reference summaries (ground truth) and compare LTC's results on the collected experimental data set using ROUGE-1, ROUGE-2 and ROUGE-SU.

7. CONCLUSIONS

We presented the novel idea of lexi-temporal clustering for post summarization of microblogs. LTC identifies significant moments of interest during long running events using information about the mean and standard deviation of the traffic rate function. It uses k-means clustering is used to identify representative tweets to describe moments. We evaluated the performance of LTC on three real world sporting events of IPL, UCL and EURO matches. Performance results show that LTC achieved up to 58% recall, 81% precision and 62% F-measure. We also evaluated the results against manual summaries using ROUGE-1, ROUGE-2 and ROUGE-SU. Results show that LTC provides upto 11.94 and 30.15 times increase in precision and recall, respectively in comparison to automatic post summarization schemes. Evaluation of various distance metrics suggests that Euclidean distance provides better recall and precision statistics than all other distance metrics.

8. REFERENCES

[1] G. Beverungen and J. Kalita. Evaluating methods for summarizing twitter posts. *In Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*, 11:9–12, 2011.

[2] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. OŠReilly Media Inc., 2009.

[3] D. Chakrabarti and K. Punera. Event Summarization Using Tweets. In *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[4] M. Chaput. stemming 1.0 : Python package index. *https://pypi.python.org/pypi/stemming/1.0*, 2017.

[5] eMarketer. Worldwide Social Network Users: 2013 Forecast and Comparative Estimates. Technical report, eMarketer, 2013.

[6] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[7] ESPN. ESPN Commentary. In *http://goo.gl/UHpQBO*, [accessed Jan-2016].

[8] ESPNcricinfo. Indian Premier League - Final, Kolkata Knight Riders vs Chennai Super Kings, Scorecard. In *http://goo.gl/vTpi3l*, [accessed Jan-2016].

[9] R. Halvorsen. Simple Twitter Streaming API access, tweetstream 1.1.1, `https://pypi.python.org/pypi/tweetstream`. Technical report, Pyhthon.org, 2011.

[10] Y. Hu, A. John, D. D. Seligmann, and F. Wang. What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds. In *Intern. Conf. on Weblogs and Social Media*, 2012.

[11] K. Inc. Klout | be known for what you love. *https://klout.com/, 2015*, 2015.

[12] Indiatoday. IPL 2012 Final Live: scores and commentary. In *http://goo.gl/UIhIkR*, [accessed Jan-2016].

[13] D. Inouye and J. K. Kalita. Comparing Twitter Summarization Algorithms for Multiple Post Summaries. In *Third IEEE International Conference on Social Computing (SocialCom)*, pages 298-306, October 2011.

[14] R. Kelly. Twitter Study Reveals Interesting Results About Usage, 40% is Pointless Babble. *http://goo.gl/DZea6f*, 2009.

[15] M. A. H. Khan, D. Bollegala, G. Liu, and K. Sezaki. Multi-tweet summarization of real-time events. In *Social Computing (SocialCom), 2013 International Conference on*, pages 128–133. IEEE, 2013.

[16] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *International Conference on Weblogs and Social Media*, 10:90–97, 2010.

[17] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[18] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Eevent Exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236, 2011.

[19] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP)*, volume 4, page 275. Barcelona, Spain, 2004.

[20] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 189–198. ACM, 2012.

[21] B. O'Connor, M. Krieger, and D. Ahn. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *International AAAI Conference on Web and Social Media (ICWSM)*, 2010.

[22] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10, 2009.

[23] B. P. Sharifi. Automatic microblog classification and summarization. *Doctoral dissertation, University of Colorado at Colorado Springs, 2010*, 2010.

[24] B. P. Sharifi, M. A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Human Language Technologies*, pages 685–688. Association for Computational Linguistics, 2010.

[25] B. P. Sharifi, M. A. Hutton, and J. K. Kalita. Experiments in Microblog Summarization. In *IEEE International Conference on Social Computing*, 2010.

[26] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[27] Skysports. European Championships Commentary. In *http://goo.gl/Wk3mR6*, [accessed Jan-2016].

[28] Skysports. UEFA Champions League Commentary. In *http://goo.gl/Df1NQo*, [accessed Jan-2016].

[29] K. Tao, F. Abel, C. Hauff, G. Houben, and U. Gadiraju. Groundhog Day: Near-Duplicate Detection on Twitter. In *Proceedings of the international conference on World Wide Web*, 2013.

[30] Twitter. Twitter Statistics. Technical report, available at `www.statisticbrain.com/twitter-statistics/`, Online; accessed Jan-2016.

[31] UEFAchampionsLeague. UCL 2012 Final Post-Match Commentary. In *http://goo.gl/LWift2*, [accessed Jan-2016].