GROUP 8 PROJECT REPORT

1. Team Members

Sarvender Dahiya	U70551903	Research, Code Implementation, Report Review
Dharmik Patel	U34190712	Research, Data Collection, Report Proofreading
Tianlin Zhang	U28119465	Data Collection, Report Drafting & Proofreading

2. Problem Statement

Social media plays an important role in the development of modern sports within the last 20 years(Özsoy, 2011). The UEFA Champions League(UCL) has a large fan-base and its events are discussed quite regularly on Twitter. We try to use tweets from the Twitter social media platform to estimate the date and time of a UCL match.

3. Research Outcome

Is it possible to determine the date and time of an event accurately within a few hours of the actual occurrence using the social media data?

Our work suggests yes, it is possible. For UCL matches we were able to determine kick-off time of 26/32 (81.25%) matches within a 3-hour window.

Week	Correct Predictions	Wrong Predictions
Nov 23 - Nov 30	13	3
Nov 29 - Dec 05	13	3

Table 1: Predictions of match time from tweets. A 3-hour period that overlaps with the time when the match is being played

4. Methodology

4.1. Identifying Matches in Tweets & Collecting Official Data

The most common way to tag a match in tweets is to use hashtags with 3-letter names of the teams with the home-team's name first, e.g. OLY vs MCI will be #OLYMCI. The match date along with the kick-off time in UTC was collected from the Official UCL website: https://www.uefa.com/uefachampionsleague/

4.2. Collecting and Analyzing Tweets

4.2.1. Tweet Collection: We used the twitter API to get the tweets containing the hashtags we collected.

Week	Nov 23 - Nov 30	Nov 29 - Dec 05
Tweets	40885	31378

Table 2 Number of tweets collected

4.2.2. Filtering Data: Tweets for each match were filtered separately, minutes and seconds dropped. Clubbed tweets into 3-hour batches

4.2.3. Plotting Graphs: Graphs plotted for no. of tweets vs time of tweet for analysis

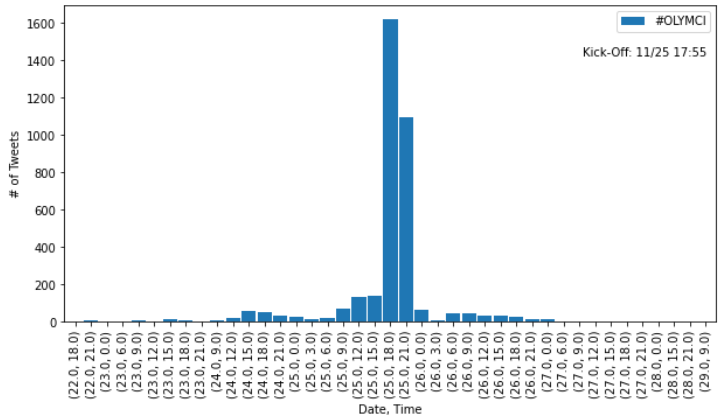

Figure 1 Tweet distribution for match: OLY vs MCI

4.3. GitHub Link:

https://github.com/sarvenderdahiya/match-times-from-tweets/tree/master

5. Discussion

Volume of tweets was higher towards the end of matches than during or before the match. This is expected as results tend to be discussed after a match. For the 4 out of the 6 wrongly predicted matches, it turns out that the teams involved in are very popular. Tweepy collects tweets starting from current time and goes back in time. The tweet-limit we set (10,000) for these matches was encountered before the date of the match could be reached. We believe that raising this limit will result in correct prediction for these matches. The other 2 didn't have enough tweets for any meaningful analysis. We believe that this can be resolved by incorporating hashtags from other languages too.

We could accurately predict the day of the match while the match time was predicted within a range of 3 hours overlapping with the match which lasts for about 2 hours. Further work can be done by collecting a more comprehensive dataset and more granular analysis of tweet times could yield even more accurate predictions. With larger dataset, manual inspection might not be feasible, hence we believe we can use statistical methods to identify if there is a significantly higher volume of tweets and make the predictions based on those parameters.

References

Conover, M. D., Ferrara, E., Menczer, F., & Flammini, A. (2013). The digital evolution of occupy wall street. *PloS one*, *8*(5), e64679

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication*, *5*, 31.

Özsoy, S. (2011). Use of New Media by Turkish Fans in Sport Communication: Facebook and Twitter. *Journal of Human Kinetics*, *28*(1). https://doi.org/10.2478/v10078-011-0033-x