# SEDTWik: Segmentation-based Event Detection from Tweets using Wikipedia

**Keval Morabia**, Neti Lalita Bhanu Murthy, Aruna Malapati and Surender Samant

Birla Institute of Technology and Science (BITS), Pilani, India

## Abstract

Event Detection has been one of the research areas in Text Mining that has attracted attention during this decade due to the widespread availability of social media data specifically twitter data. Twitter has become a major source for information about real-world events because of the use of hashtags and the small word limit of Twitter that ensures concise presentation of events. Previous works on event detection from tweets are either applicable to detect localized events or breaking news only or miss out on many important events. This paper presents a tweet-segmentation based system for event detection called SEDTWik, an extension to a previous work, that can detect newsworthy events occurring at different locations of the world from a wide range of categories. The main idea is to split each tweet and hash-tag into segments, extract bursty segments, cluster them, and summarize them. We evaluated our results on the well-known Events2012 corpus and achieved state-of-the-art results.

**Keywords:** Event detection, Twitter, Social Media, Microblogging, Tweet segmentation, Text Mining, Wikipedia, Hashtag.

## Introduction

### What is an Event?

- Some unique thing that happens at some point in time
- A real-world occurrence $e$ with an associated time period $T_e$ and a time-ordered stream of Twitter messages $M_e$, of substantial volume, discussing the occurrence and published during time $T_e$
- Categories of events: Sports, Politics, Entertainment, Science & Technology, etc.

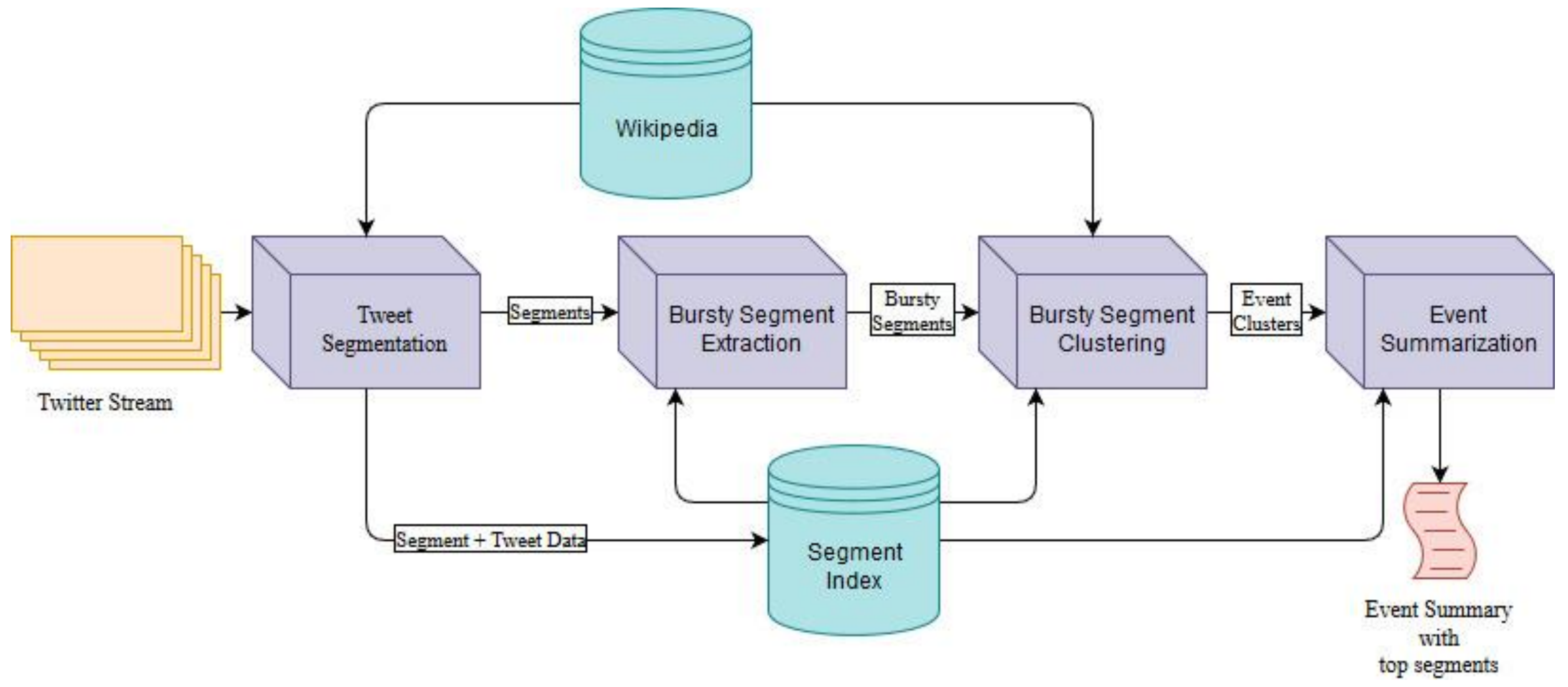
### Why Twitter can be used for Event Detection?

- Concise text because of 280-char limit
- Not only publish about an event but also propagate by ***retweeting***
- Use of ***Hashtags*** that contain most information about an event. E.g. #RIP, #ElectionResults
- **Mention** entities associated with the event. E.g. *@iamsrk* for Shah Rukh Khan

### Challenges for Event Detection from Tweets

- Noisy data containing grammatical and spelling mistakes
- Informal writing in mixed language
- Large volume of data (**~500 million tweets/day!**)

## Methodology



SEDTWik Architecture

### 1. Tweet Segmentation

- **Why?** Phrases contains much more specific information than the unigrams in it
- **How?** Split a given tweet into non-overlapping meaningful segments, giving more weight to hashtags ($H$). Filter out words not present as a Wikipedia page title
- [vice presidential debate] vs [debate], [presidential] , [vice]
- Sample Tweet Segmentations with $H = 3$:

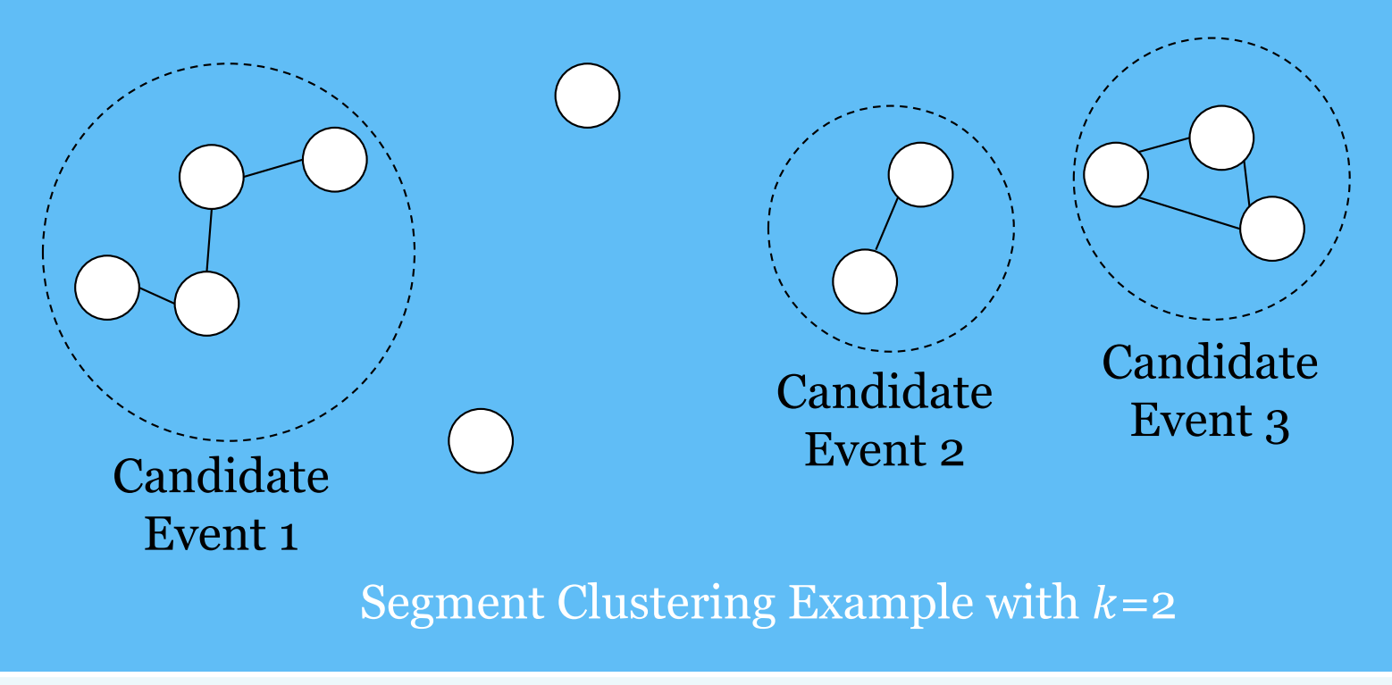
| **Tweet** | **Segmentation** |
|---|---|
| Joe Biden and Paul Ryan will be seated at the debate tonight #VpDebate | [joe biden], [paul ryan], [seated], [debate], [tonight], [vp debate]x3 |
| Amanda Todd took her own life due to cyber bullying #RipAmandaTodd #NoMoreBullying | [amanda todd], [cyber bullying], [rip amanda todd]x3, [no more bullying]x3 |

### 2. Bursty Segment Extraction

- **Why?** Processing millions of segments is computationally expensive
- **How?** Score segments based on their bursty probability ($P_b$), and follower count ($fc$), retweet count ($rc$), and count of unique users using them ($u$). Select top $K = \sqrt{N_t}$ segments based on $Score$ ($N_t$ = total number of tweets)
- $P_b(s)$ measures how frequent a segment is occurring compared to its expected probability of occurrence
- $Score_s = P_b(s) \times \log(u_s) \times \log(\mathrm{rc}_s) \times \log(\log(fc_s))$

### 3. Bursty Segment Clustering

- **Why?** To group segments that are related to the same event
- **How?** Variation of Jarvis-Patrick Clustering algorithm. Segments considered as nodes in a graph and 2 segments belong to same cluster if both are in $k$-NN of each other
- Segment similarity: $tf-idf$ similarity between contents of tweets containing the segment



Segment Clustering Example with $k$=2

### 4. Event Summarization

- **Why?** List of segments not as informative as actual tweets
- **How?** Use all tweets containing segments of the event cluster and apply LexRank algorithm to them to obtain a summary of the event
- Use newsworthiness of segments and segment-similarity in a cluster to compute Event Newsworthiness
- Only summarize those events that have newsworthiness more than a threshold

## Examples of Events detected

Detected events from the Events2012 dataset (McMinn et al., 2013):

| **Date** | **Event** |
|---|---|
| Oct 11, 2012 | *[national coming out day], [national coming day], [lgbt], [coming day], [ncod]* → National Coming Out Day celebrated on this day |
| Oct 12, 2012 | *[nobel peace prize], [nobel], [european union], [peace prize]* → The European Union wins the 2012 Nobel Peace Prize |
| Oct 15, 2012 | *[justin bieber], [baabworldrecord], [vevo]* → Justin Bieber's music video Beauty and a Beat (BAAB) creates world record of most watched VEVO video in 24 hrs |
| Oct 16, 2012 | *[debate], [barack obama], [presidential debate]* → 2nd US presidential debate between Barack Obama and Mitt Romney |

## Experimental Results

Comparison of SEDTWik with Twevent (Li et al., 2012), a similar Tweet-Segmentation based method for events detected in period Oct 11 – Oct 17, 2012:

| **Method** | **No. of events** | **Precion** | **DERate** |
|---|---|---|---|
| SEDTWik | **79** | **88.12%** | **14.10%** |
| Twevent | 42 | 80.32% | 16.67% |

Metrics:

- *No. of events:* total no. of realistic events detected
- *Precision:* fraction of the detected events that are related to a realistic event
- *Duplicate Event Rate (DERate):* percentage of events that have been duplicately detected among all realistic events detected

## Acknowledgements

- We thank Dr. Yang Liu, head of Liulishuo's Silicon Valley AI Lab for providing valuable feedback during the pre-submission mentorship program
- We also thank Dr. Aixin Sun, Associate Professor at Nanyang Technological University (NTU), Singapore for providing the Wikipedia keyphraseness values that was used in calculating segment newsworthiness
- We also thank Microsoft Research Lab, India for proving partial travel grants to attend this conference

## References

- Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: Segment-based event detection from tweets. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, New York, NY, USA. ACM
- Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM '13, New York, NY, USA. ACM