# An LDA-lexical syntactical approach for events and features extraction of earthquakes from Spanish and English tweets

**Enrique Valeriano Loli, Juanjosé Tenorio Peña, Rodrigo López Condori**
Pontificia Universidad Católica del Perú
Lima, Perú
{enrique.valeriano,juanjose.tenorio,a20112387}@pucp.pe

## Abstract

In the last few years, social networks like Twitter have been a very useful resource for tracking the events that happened before, during and after an earthquake. Several studies of this topic have applied different techniques like Clustering or Temporal models for extracting these events from Twitter. In this paper, however, we propose a new approach for extracting not only the events that happened in the earthquake but also some of its most prominent features like intensity, epicenter and affected places. We performed a lexical syntactical analysis of Spanish and English tweets in order to find the events that happened, in addition to a semantical analysis using statistical metrics and models like Pointwise Mutual Information(PMI) and Latent Dirichlet Allocation(LDA) for extracting the features of the earthquake. Our results show that, by considering the semantics and syntactics of the tweets, we can extract important events and features of an earthquake, which can be used for online detection and tracking of similar disasters.

## 1 Introduction

Twitter is one of the most popular social networks in the world. As of February 2017, it has 319 million active users[1]. Because of its privacy policy, Twitter is vastly used for reporting information of events (Quan-Haase and Young, 2010). That's why, over the years, there have been many studies that used this information for reconstructing and extracting events and features from disasters like earthquakes (Doan et al., 2011).

Several studies have been done about this topic using different kinds of algorithms and approaches. Some of the most prominent algorithms found in the literature are probabilistic models like LSA and Spatial-Temporal models (Weiler et al., 2016), however there are several downfalls with these approaches. First, Weiler (2016) found that at least 75% of the works are case-oriented, that is, they are oriented to an specific earthquake or disaster. Furthermore, literature doesn't consider Spanish tweets (Bontcheva and Rout, 2014), which could be very useful for countries with high earthquake activity like Peru or Chile. Finally, from the studies that consider Latent Dirichlet Allocation(LDA) models, there are only real time Event Detection but not Feature Extraction like the epicenter of an earthquake or the affected places (Weiler et al., 2016).

In that sense, this paper propose an approach that use a LDA topic model with NLP techniques like Named Entity Recognition and Pointwise Mutual information for not only extracting the events related to an earthquake but also extracting its prominent features like epicenter, intensity and affected places.

## 2 Related Work

### 2.1 Event Concepts

There are some concepts that have to be considered in order to perform a proper Event Detection. Weiler (2016) found out that at least 60% of the related work considered Ground Truth as an evaluation metric. In Event Detection, Ground Truth is a metric for evaluating the quality of the extracted events by making sure that these events really happened (Weiler et al., 2016). Authors

[1] Fortune Magazine: http://fortune.com/2017/02/09/twitter-q4-2016/

like Li (2012) and Martin (2013) used a manual Ground Truth, by comparing manually the extracted events with news sources in order to test their validity. Other authors like Osborne (2010) automatize this process and used different APIs[2] from sources like Wikipedia for comparing the extracted events with relevant articles. In this work, we consider the manual Ground Truth because is the most accepted approach and provides better results depending of the source; Weiler (2016) observed that at least 70% of Event Detection works use the manual version.

### 2.2 Event Detection Approaches

Bontcheva (2014) classifies Event Detection approaches into 3 types: Model Based, Clustering Based and Based on signal's processing. On the other side, authors like Farzindar (2015) categorize the approaches based on application domains and evaluation metrics. We follow Bontcheva's proposal, because we observed many summary authors like Weiler (2016) and Winarko (2013) found out that many of the works considered this type of classification. From that point, we notice that at least 50% of the works used LDA as a method or baseline for detecting events. For example, Aiello (2014) compared six Topic Extraction methods for Event Detection and showed that LDA was the algorithm that performs better. In addition, these works used supervised learning, like the work of Takeshi, Okazaki and Yutaka (2010) which propose a technique for extracting events using a huge labeled dataset that represents events of an earthquake. Furthermore, popular tools like Twitinfo (2011) and Twevent (2012) also use labeled data for extracting real time events.

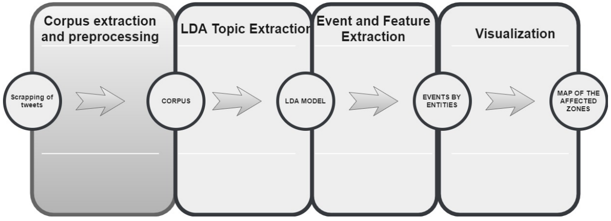
### 2.3 Conclusions from related work

From the state of the art's analysis, we conclude that the current tools and best approaches for Event Detection used labeled datasets, so there is a great opportunity for studying the semantic and syntactics of the text without having a labeled dataset beforehand. Also, we decided to evaluate our results using a manual Ground Truth, which is widely used for these works.

## 3 Methodology

Our approach consist of 4 main steps. These steps are described in Figure 1 and are explained in greater detail in Section 4:

Figure 1: Methodology for event and feature extraction of earthquakes


Corpus extraction and preprocessing
LDA Topic Extraction
Event and Feature Extraction
Visualization
Scrapping of tweets
CORPUS
LDA MODEL
EVENTS BY ENTITIES
MAP OF THE AFFECTED ZONES

First we identify the earthquake and get a corpus for a fixed ranged of dates. Next we generate a LDA model for determining clusters of terms for a fixed number of topics. Then we used the Stanford Named Entity Recognizer for identifying the entities and places of the earthquakes. With those entities and places, we performed a Lexical and Syntactical analysis to identify the events based of the related tweet for the identified entity.

Finally, we identified some features of the earthquake using Pointwise Mutual Information(PMI) between each pair of words in the clusters of the LDA model. Then, we select the best ”n” relationships, that are the pairs that have greater PMI. With all this information, we plotted the events in a map of the country using the date and hour of the resulting events from our analysis.

## 4 Experimentation

We developed 5 experiments with corpus from different earthquakes around the world; these can be seen on Table 1.

Table 1: Earthquakes used on the experiments

| Country | Range of dates |
|---|---|
| Chile | 16/09/15 - 30/09/15 |
| Chile | 20/12/16 - 31/12/16 |
| Ecuador | 15/04/16 - 30/04/16 |
| New Zealand | 13/11/16 - 09/12/16 |
| Japan | 10/04/16 - 20/04/16 |

In the next subsections, the results for the earthquake that happened on Chile on September 16, 2015 are shown. Then, there are tables with the results of entities and places for the other 4 earthquakes, so that the nature of these results can be explained.

[2] API: Application Programming Interface.

### 4.1 Corpus Description

For the extraction phase, we used the open source application GetOldTweets[3] for extracting tweets from more than 2 weeks. Different corpus of tweets were generated, one for each earthquake analyzed. Corpus of Chile's and Ecuador's earthquakes are in Spanish. Meanwhile, corpus of Japan's and New Zealand's earthquakes are in English. The information of the Chile 8.4 earthquake corpus can be seen on Table 2.

Table 2: Corpus information for Chile's 8.4 Earthquake

| Information | Value |
|---|---|
| Country | Chile |
| Number of tweets | 11873 |
| Language | Spanish |
| Range | 16/09/15 - 30/09/15 |
| Magnitude | 8.4 |

In the same manner, the corpus generated for each earthquake had a different quantity of tweets which are shown on Table 3.

Table 3: Information of each earthquake's corpus

| Earthquake | Year | Tweets |
|---|---|---|
| Chile 8.4 | 2015 | 11873 |
| Chile 7.6 | 2016 | 3789 |
| Ecuador 7.8 | 2016 | 8461 |
| New Zealand 7.9 | 2016 | 5611 |
| Japan 7.0 | 2016 | 3897 |

### 4.2 Latent Dirichlet Allocation (LDA)

For supressing the noise of the corpus, we applied regex and stopwords filtering for getting rid of the links, hashtags and stopwords.

Once the cleaning was finished, we used LDA in order to obtain the topics that were represented by different subsets of words as an initial way of identifying some words or terms that could be relevant on earthquakes domain of knowledge.

We experimented with different values for the amount of clusters generated as well as the words per cluster. We ended up choosing to generate 10 clusters with 20 words each and keep the 14 more relevant words of each cluster since after this threshold almost every word started being unrelated to the earthquakes.

Table 4: Clusters for Chile's 8.4 degree earthquake

| Clusters | Related words |
|---|---|
| 1 | 8,4,terremoto,chile,autoridades, santiago,sismo,dos,magnitud, siete,temblor,genes, chilenas, |
| 6 | terremoto,chile,tsunami,4,8, alerta,grados,fuerte,richter,graus, atinge,6 |
| 7 | chile,8,terremoto,4,muertos, grados,menos,evacuados,magnitud, tras,mill,deja,5 |
| 4 | terremoto,chile,8,4,tras, tsunami,alerta,costa,magnitud, olas,grados,toda,7 |

On Table 4 we can see some words related to this specific earthquake, such as ”chile” and ”santiago”, which relate to the place where the earthquake took place, as well as 8 and 4, which in this case indicate the degree of the earthquake: 8.4.

We also obtained some other words we might expect to find on any earthquake such as ”terremoto”, ”magnitud”, ”temblor” and ”sismo”.

### 4.3 Stanford Named Entity Recognizer

Afterwards, we proceeded to use the Stanford Named Entity Recognizer with previously trained models for Spanish and English, focusing on obtaining places relevant to each earthquake corpus.

Table 5: Places identified by Stanford Named Entity Recognizer for Chile's 8.4 degree earthquake

| Identified places | Frequency |
|---|---|
| Illapel | 89 |
| Santiago | 69 |
| Ecuador | 60 |
| Hawai | 56 |
| Coquimbo | 22 |
| Tsunami | 14 |
| Valparaiso | 8 |

On Table 5 the places with higher frequency are presented, aside from Chile which had the highest with 9303.

### 4.4 Lexical and Syntactical analysis

We performed a Lexical and Syntactical analysis as a way to retrieve relevant entities from the earthquakes. In those entities we expected to find some locations in similar manner to Section 4.3, but also

[3] GetOldTweetsrepository: https://github.com/Jefferson-Henrique/GetOldTweets-python

some groups, entities or words which were heavily involved in those events, such as affected people or assistance actions.

For this we used the UDPipe[4] parser in order to obtain the grammatical categories of each word on a tweet, as well as the syntactic tree.

The grammatical categories which were worked upon were the 'SUBJ' tag for subjects.

Following this procedure, each tweet was assigned to the subjects found inside it; this was used to obtain tweets grouped by subject and verb.

### 4.4.1 Pruning and clustering

If we just used all the words tagged as 'SUBJ' the result would be a large amount of entities. In order to reduce this amount and keep the most relevant entities, a pruning process was performed, where the entities that had between 10% and 100% of the maximum subject frequency where the ones that remained.

After the entities have been pruned, a clustering process occurs, in which the remaining entities that surpass a minimum threshold of similarity were clustered together and treated as the same.

Table 6: Entities for Chile's 8.4 degree earthquake

| Clusters | Entities |
|---|---|
| 1 | chile, chilee |
| 2 | terremoto |
| 3 | tsunami |
| 4 | richter |

The entities shown on Table 6 where clustered via 2 metrics:

1. Levenshtein distance: minimum replacements/insertions/deletions needed to turn one word into another.

2. Average Entity to entity distance in a tweet

The minimum Levenshtein distance in order to cluster two entities together was 20% of the maximum length between the two entities and the minimum average entity to entity distance was 1.5. Both of these values were obtained via experimentation and results reviewing.

Table 7: Example of a tweet per entity

| Entities | Related tweet |
|---|---|
| chile | #TerremotoChile Vdeo noticia: Chile sufre un terremoto de 8,4 grados el ms potente de este ao en el mundo http:// fb.me/3UG3lLxyA |
| terremoto | Terremotode 8,4 grados remece el centro de Chile terremoto sacudi esta tarde al centro de Chile , alcanzando... http:// fb.me/7uXzLS5tD |
| tsunami | Tsunami llega a las costas de Chile tras terremoto de magnitud 8,4 (Fotos) http:// bit.ly/1LjN4hi |
| richter | Un terremoto de 8,4 en la escala Richter sacude el centro de Chile http:// fb.me/MlWJrNqd |

On Table 7 a tweet for each entity is shown. Each entity has more tweets similar to these, being the entity "chile" the one that has the most.

The tweets per entity were classified as an event inside the earthquakes since they served the purpose of explaining what happened to each entity in the corpus.

## 4.5 Pointwise Mutual Information (PMI)

Pointwise mutual information(PMI) is an information measure for identifying how much related are two variables. In information retrieval, it's vastly used for identifying words that are very related in function of their co-occurrences in a document, that is, how many times those words appear together in the same document (Bouma, 2009). The PMI between two words is computed by the following formula:

$$\log_2 \frac{P(x,y) * N}{P(x) * P(y)}$$

where P(x,y) is the number of co-occurrences between the words 'X' and 'Y', P(x) and P(y) are the frequencies of the words X and Y respectively and N is the number of documents (Bouma, 2009).

There has been a wide range of applications of PMI in Natural Language Processing. Rana (2016) found that is one of the most common metric for relationship extraction in Ontology Learning, because it considers the semantic relationship

[4] UDPipe library: https://ufal.mff.cuni.cz/udpipe

between entities, which is what we wanted for extracting the most prominent features of the earthquake.

We compute PMI between the topics and terms identified in the LDA Model. Based on that, we identified the pair of words with a greater PMI and mapped this pair with a feature of the earthquake. The results are shown on Table 8.

Table 8: Features extracted using PMI

| Pair | Feature | PMI |
|---|---|---|
| chile-8 | Intensity | 12.533 |
| tsunami - alerta | Tsunami | 12.122 |
| epicentro-illapel | Epicenter | 10.043 |
| santiago - septiembre | month | 10.588 |

We observed that we extracted interesting properties of the earthquake like its intensity and epicenter. In the case of the epicenter, we identified Illapel as the epicenter of the earthquake, which is true based of the news sources.

## 5 Results and Discussion

The identified places possibly related to the earthquakes previously mentioned (except the earthquake of Chile from 2015) are shown on Tables 9 ,10 , 11 and 12.

Besides, as it was mentioned on Table 5, all the places shown have the higher frequency without taking in consideration the respective country of each earthquake (Chile, Ecuador, New Zealand and Japan respectively) .

Table 9: Places identified for Chile's 7.6 degree earthquake

| Identified places | Frequency |
|---|---|
| Melinka | 40 |
| Sur | 38 |
| Isla | 25 |
| Quellon | 24 |
| Chiloe | 24 |
| Lagos | 20 |
| Ecuador | 12 |

Table 10: Places identified for Ecuador's 7.8 degree earthquake

| Identified places | Frequency |
|---|---|
| Colombia | 216 |
| Guayaquil | 142 |
| Chile | 100 |
| Pedernales | 97 |
| Quito | 90 |
| Colima | 75 |
| Esmeraldas | 66 |

Table 11: Places identified for New Zealand's 7.9 degree earthquake

| Identified places | Frequency |
|---|---|
| Christchurch | 818 |
| Island | 693 |
| South | 689 |
| Wellington | 127 |
| Amberley | 126 |
| Hanmer | 54 |
| Canterbury | 50 |

Table 12: Places identified for Japan's 7.0 degree earthquake

| Identified places | Frequency |
|---|---|
| Kumamoto | 663 |
| Kyushu | 253 |
| Ecuador | 136 |
| Prefecture | 36 |
| Island | 31 |
| Tsunami | 15 |
| Myanmar | 12 |

We observed many different places related to an earthquake, not only the country where it happened. This is possible due to different factors such as other countries that might be affected as well shown in Ecuador's possible places, which includes one of its neighbors Colombia. Also the date of the earthquake can affect the results, such as for Japan's earthquake, which includes Ecuador since both earthquakes happened around the same week.

Below, entities and tweets associated with each earthquake are shown on Tables 13, 14, 15, 16, in order to see their relevance.

Table 13: Entities for Chile's 7.6 earthquake

| Entities | Related tweet |
|---|---|
| Chiloe | Asi quedo Chiloe tras terremoto de 7,6 en Chile - teleSUR TV |
| Melinka | Terremoto en Chile 7.6 grados en la Zona de MELINKA. ZONA SUR DE CHILE |
| Chile | Terremoto de 7.6 grados Richter parte la tierra en Chile . Se descarta Tsunami |
| Lagos | Terremoto 7.6 ritcher en melinka, region de los lagos. chile . Es en el Sur. daddy_yankee |

Table 14: Entities for Ecuador's 7.8 earthquake

| Entities | Related tweet |
|---|---|
| Ecuador | Terremoto de 7 ,8: Mensaje de actor de doblaje mexicano de 'Dragon Ball Z' conmueve a Ecuador |
| Quito | As se sinti en Quito el terremoto de 7 ,8 en Ecuador |
| Colombia | Terremoto de 7 ,8 grados sacude la frontera entre Colombia y Ecuador |
| Esmeraldas | ECUADOR . Provincias de Manabi y Esmeraldas, las mas afectadas por el terremoto de 7 ,8 |

Table 15: Entities for New Zealand's 7.9 earthquake

| Entities | Related tweet |
|---|---|
| Christchurch | Christchurch, New Zealand Hit by 7 .9 Earthquake and Tsunami |
| Wellington | New Zealand Earthquake : Damage in Wellington after 7 .8 magnitude tremor |
| New Zealand | We need to pray for New Zealand and all of the islands nearby. 7 .4 to 7 .8 earthquake with tsunami's! |
| Canterbury | Thoughts and prayers to my New Zealand friends in the South Island in Canterbury where there was a 7 .8 magnitude earthquake . Stay safe |

Table 16: Entities for Japan's 7.0 earthquake

| Entities | Related tweet |
|---|---|
| Kyushu | Earthquake in japan 6- 7 Magnitude kyushu |
| Kumamoto | There was a big earthquake today in Kumamoto, Japan .Shindo,the unit of earthquake 's size is 7 , the largest. |
| Japan | Big earthquake in southern Japan , initial magnitude 7 .1 |
| Ecuador | Japan was struck by a 7 .0M earthquake Friday. Then yesterday, Ecuador was struck by a 7 .8-M quake- perhaps related to Mars & Pluto stations. |

We consider each of the tweets related to an entity as en event. For testing the validity of every event, we compute the manual Ground Truth, which can be seen as "How many of these events really happened?". The state of art suggest that having many news sources helps to improve the quality of the results, so we considered some of the most important news sources like CNN, BBC, New York Times and The Associated Press. Once the source news were established, we check every event with their articles that covers a particular earthquake, so if the event was verified by each of the source news, we considered him as valid. The results from this evaluation are presented on Table 17.

Table 17: Ground Truth Evaluation

| Location | Degree | Date | GT (%) |
|---|---|---|---|
| Chile | 8.4 | 16/09/2015 | 60% |
| Chile | 7.6 | 25/12/2016 | 50% |
| Ecuador | 7.8 | 16/04/2016 | 50% |
| New Zealand | 7.9 | 14/11/2016 | 80% |
| Japan | 7.0 | 11/04/2016 | 80% |

From the summaries of Weiler (2016) and Winarko (2013), we notice that the values of Ground Truth for many event detection works varies between 60% to 90%, so we got very good results comparing with the state of the art. This results, however, are not very accurate because of different factors such as the size of the corpus, the range of dates and so on. That's why, there is room for improvement in our evaluation measure.

## 6 Visualization

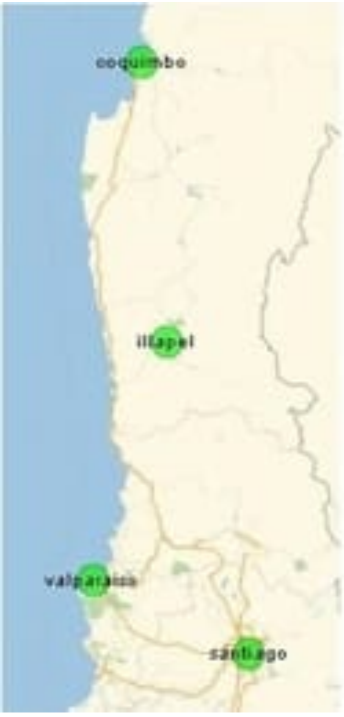
### 6.1 Motivation

Our motivation was to propose a new approach for earthquake detection and tracking systems, so we had to make sure that our approach can be adapted to those systems. That is the reason why we implemented a visualization module that can show the events and entities detected by our approach in real time.

### 6.2 Visualization Module

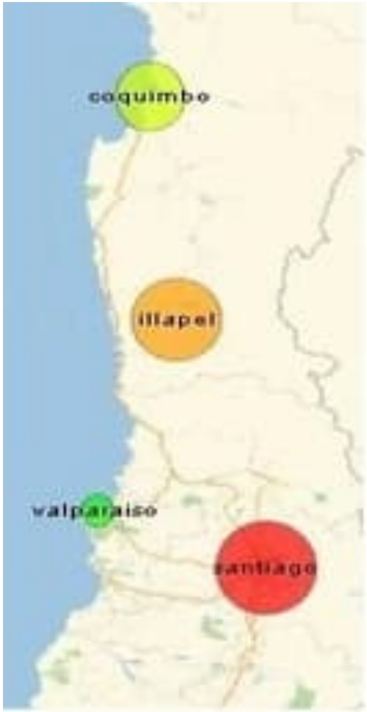
We used the information obtained from the tweets associated with each earthquake, in order to show in a world map different factors of the earthquakes such as the epicenter, intensity or how was the frequency of tweets at different times. For this purpose we use the software Mathematica[5], which is an incredible tool with helpful features, together with the collected information.
To begin with, we show in Figure 2 a regular picture of Chile with some cities that might be affected by the earthquake of 2015.

Figure 2: Cities affected by Chile's 8.4 degree earthquake



Then with the collected information we can reflect in that picture how much information were related with each city during different moments in the range of dates of the related earthquake.
To show this we use different colors and sizes for the cities in Figure 3.

Figure 3: Chile during the 8.4 degree earthquake



The circles for each city are related with the frequency of tweets per day, a small circle means that the city was not mentioned regularly that day and a big circle means that the city is mentioned in a lot of tweets. Besides, the color scale start from green to red, with yellow being in the middle, so the colors can be a variation of those 3 colors, which depends of the quantity of tweets.
According to this, it seems that Valparaiso is not involved in the earthquake, meanwhile Coquimbo, which has some mentions, is really close to the earthquake. In the same way, Illapel is mentioned very frequently over time, which suggest that is the epicenter, and finally Santiago is mentioned a lot some days after the earthquake, because is very close to Illapel and most of the assistance and supplies came from there.

## 7 Conclusions and Future Works

From the experiments, we have shown that, by using NLP techniques and tools like Named Entity Recognition(NER) with a LDA model, it is possible to identify events and features from disasters like earthquakes that can be as good as the ones identified by the state of the art. In that sense, this

[5] Mathematica: https://www.wolfram.com/mathematica/

approach has the following advantages:

1. This approach considers Spanish language, which is poorly found in the literature.

2. By considering the semantics and syntactics of the tweets, better features and events were identified.

3. This approach can compete with the state of the art and may get better results if more data for each earthquake is provided.

4. Our approach can be adapted for an Online Earthquake Detection System, because LDA has an online version which is very used in this type of systems.

For future research, we consider to perform an adaptation of this approach for detecting events in real time. The advantage of this work is that LDA is very good for real time detection, like shown by Aiello (2014) in his summary of Event Detection works. Another improvement is on the visualization module, because if we use Google Maps API for automatic detection of the places, this module will be fully automatized. Finally, using a not-manual Ground Truth like the one proposed by Weiler (2016) may provide a better metric for testing our results.

## References

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence Journal* 31:132–164.

Kalina Bontcheva and Dominic Rout. 2014. Making sense of social media streams through semantics: a survey. *Semantic Web* 5:373–403.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference 2009*. Gunter Narr Verlag, pages 31–40.

Martin Carlos, David Corney, Ayse Goker, and Andrew MacFarlane. 2013. Mining newsworthy topics from social media. In *Proceedings SGAI Workshop on Social Media Analysis in conjunction with Intl. Conf. of the British Computer Societys Specialist Group on Artificial Intelligence (SGAI)*. pages 35–46.

Soan Doan, Huu Phuc Vo, and Nigel Collier. 2011. An analysis of twitter messages in the 2011 tohoku earthquake. In *Electronic Healthcare*. pages 58–66.

Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: Segment-based event detection from tweets. In *Proceedings for 2012 International Conference of Information and Knowledge Management (CIKM)*. pages 155–164.

Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings for 2011 International Conference of Human Factors in Computing Systems (SIGCHI)*. pages 227–236.

Arif Nurwidyantoro and Edi Winarko. 2013. Event detection in social media: a survey. In *Proceedings for 2013 International Conference on ICT for Smart Society*. pages 1–5.

Miles Osborne, Sasa Petrovic, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proceedings of 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT)*. pages 181–189.

Symeon Papadopoulos, David Corney, and Luca Maria Aiello. 2014. Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media. In Symeon Papadopoulos, David Corney, and Luca Maria Aiello, editors, *Proceedings of the SNOW 2014 Data Challenge*. pages 1–8.

Anabel Quan-Haase and Alyson L. Young. 2010. Uses and gratifications of social media: A comparison of facebook and instant messaging. *Sage Journals* 30:350–361.

Toquir A. Rana and Yu-N Cheah. 2016. Aspect extraction in sentiment analysis: comparative analysis and survey. *An International Science and Engineering Journal* .

Li Rui, Kin Hou Lei, Kevin Chen-Chuan Chang, and Ravi Khadiwala. 2012. Tedas: A twitter-based event detection and analysis system. In *Proceedings of 2012 IEEE 28th International Conference on Data Engineering*. pages 1273–1276.

Sakaki Takeshi, Makoto Okazaki, Huu, and Matsuo Yutaka. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. ACM New York, NY, USA 2010, pages 851–860.

Andreas Weiler, Michael Grossniklaus, and Marc H Scholl. 2016. Editorial: Survey and experimental analysis of event detection techniques for twitter. *The Computer Journal, 2016* .