# Extracting and Classifying Events from Social Media Posts for Life Story Generation

Robee Khyra Mae J. Te
De La Salle University
Manila, Philippines
robee_te@dlsu.edu.ph

Janica Mae M. Lam
De La Salle University
Manila, Philippines
janica_lam@dlsu.edu.ph

Camille Alexis T.R. Saavedra
De La Salle University
Manila, Philippines
camille_saavedra@dlsu.edu.ph

Alden Luc R. Hade
De La Salle University
Manila, Philippines
alden_hade@dlsu.edu.ph

Ethel Ong
De La Salle University
Manila, Philippines
ethel.ong@delasalle.ph

## ABSTRACT

People use social media, in particular Facebook, to share stories about themselves and the things that interest them. However, a Facebook user's posts by themselves cannot provide a concise narrative of events to tell a complete life story. Story generators can be designed to utilize these events extracted from posts that users share about themselves into a life story. Before this can be achieved, the story generator needs to be able to classify posts based on their textual content. Such techniques are already available in email, where it is possible to classify messages into categories. In this paper, we describe our approach in using automated classifiers to categorize a user's posts, focusing on those that describe travelling, eating, dining and celebrating events. We then show the performance of our classifiers. Once events are identified, we extract details from these posts and store them into an event model. We end our paper with a short discussion of our story generation process that utilizes the topical and temporal relations that exist among the events.

## KEYWORDS

Social Media; Life Event Classification; Story Generation

## 1 INTRODUCTION

The world of storytelling has evolved from oral tradition to digital and online technologies. In an attempt to build machines that can mimic a human's abilities to understand and generate stories, the past decade has seen a community of NLP researchers who combined AI with text generation techniques to develop automated story generation systems that can produce stories of various genre and for varying purposes.

A main requirement for computers to be able to tell stories is the availability of storytelling knowledge that enables intelligent machines to know concepts and events about the real world. These led researchers to explore the use of different resources in the generation of stories, including commonsense ontology [7][14], character models [17], event models [2], narrative structures [18], model of affective responses [13], and corpus of stories [5][10].

The prevalence of social media has enabled storytelling to move to the online digital platform, providing a multi-modal environment for people to share their life experiences using a combination of text, photos and videos. It also paved the way for increased participation and exchange of information from the community, as friends share similar related events and personal experiences to enrich the contents of the stories.

The most popular social media platform, Facebook, is emerging as a near-universal storytelling method. It contains numerous stories, facts and events from users all around the world. Users can create a complete story about themselves, from their birth to the current day by posting events in their respective Timeline. Simple features such as Likes and Shares enable posts to spread quickly. The Comments feature allow friends to pitch in information which may or may not be related to the post.

Facebook adopts a free-form nature in allowing users to share information. It does not limit the content to one or two media types, nor does it limit the length of text-based posts. Thus, a post can contain simple text that describes what the user is currently doing or feeling; to images and videos of any size and length that illustrate their experiences. These small acts of posting and updating one's status about personal life events, triumphs, failures, wishes and goals, can be considered short stories or snippets of one's life that are arranged chronologically similar to a storyline [24].

Social media, by its very nature, is very hard to deal with because of the presence of noisy user generated data. Our main contribution in this paper is the classification of posts from an individual user's Facebook account according to their event types, and the subsequent extraction of event details. This can be a first step needed by a smart computer to understand a person's life. Software agents can use the information from Facebook data to make sense of a person's activities and experiences leading to a better understanding of people, both as an individual and as a whole community, and opening up possibilities of customization and personalization in computer-based support systems. We first give a short review of related work in classifying social media data in Section 2. In Section 3, we describe the structure of an event that we are aiming for in relation to the life story to be generated. We then detail our approach in classifying events in Section 4, as well as the performance results. In Section 5, we briefly examine the feasibility of generating a life story containing events from these classified posts. We end our paper with a discussion of further work to improve our classifier and generation algorithms.

## 2 RELATED WORK

With the volume of data on social media platforms, NLP researchers have worked on putting some structure to organize text-based data to provide a more appealing interface [19]; to discover themes in disaster-related tweets [21]; to find patterns and glean community sentiments in election tweets [23]; and to detect life events [3].

The work of Kinsella, Passant, and Breslin states that social media, because of its informal and brief nature, presents a unique challenge for topic classification [6]. There is also the frequent reliance on hyperlinks to external sites to give context to a conversation. Their study investigated the usefulness of metadata such as those hyperlinks in order to better understand the topic of a particular post. They found out that including object metadata, not necessarily hyperlink metadata, outperforms classification that is based solely on the post's original text content.

The work of Setty, Jadi, Shaikh, Mattikalli, and Mudenagudi involves dynamically classifying a Facebook user's news feeds into categories such as life events, entertainment and liked pages as a "better representation of data on the user's wall" [19]. The life event posts were further classified based on their sentiments as happy, neutral and bad feelings.

Cavalin, Moyano, and Miranda looked at detecting the travelling event by analyzing Tweets gathered from multiple users and written in both English and Portuguese [3]. They used multiple classifiers and several different feature sets such as uni-grams, bi-grams, co-occurrence of n-grams in conversations. They also performed oversampling to account for the unbalanced dataset that is a characteristic of social media data.

These studies, however, have focused on finding patterns and trends from data coming from posts and tweets of multiple users over a certain period of time. Choudhury and Alani further noted that most research works in detecting events from social media content have focused on world events such as earthquakes and elections, and entertainment news [4]. Focusing their own efforts on individuals, they detect common personal life events from Twitter to identify those that are interesting and important and can therefore be used to form part of a personal digital story book.

Developing classification systems that center on a single user's posts over a period of time may enable a software agent to make sense of a person's daily activities and experiences, which can potentially lead to an increased ability of intelligent machines to understand us and our world better. However, personal events posted by an individual user, whether through Facebook or Twitter, are not high volume and usually contain "short, informal and noisy content" that may be scattered across multiple posts [4]. NLP systems such as story generators need to deal with these characteristics when processing textual data from social media.

## 3 LIFE STORIES AND EVENTS

A life story is a personal narrative about the significant events and experiences in a person's life [22]. This is in contrast with an autobiography or a memoir, which is an account of a person's life written by themselves. A life story is also non-fictional. While most story generation systems have focused on the production of stories with a specific genre, such as children's stories [10] [20], in our study, the story generator has to be given personal life events from which it can generate an individual's life story.

Life stories can include a person's birthday, family members, childhood events, educational background, work experiences and significant contributions, a photo or likeness of the person [25]. The identification of these elements is necessary to enable the story generator to determine how to organize the contents of a life story.

### 3.1 Events in Stories

Events form an integral part in a story. While multiple definitions of events exist in literature, a common way of describing an event is to refer it as "any situation that can happen, occur or hold" at a particular location during a particular time [12].

Fiction-based story generation systems define an event to be any action that a character performs in the story world. Such actions are usually denoted as verbs in the narrative text, e.g. play (in the park) and read (a book). Non-verbal events can also take place in stories, such as parties and celebrations; and naturally-occurring phenomena like flood and earthquake.

We define an event as *anything that happens, especially one of importance*. Importance is a subjective quality; different readers of a story may have different perceptions of when an event is considered important. In this paper, we take on the assumption that posts describing events about an individual's life celebrations, travel and eating experiences are important, thus necessitating that our classifier be able to identify such events.

The selection of these categories was based from a status update feature that Facebook introduced in 2013. Called *predefined activities*, this status update feature allows users to easily specify what they are feeling or doing, using readily available prompts such as feeling, listening to, watching, playing, reading, celebrating, eating, and attending. The actual task of organizing and sequencing these events into a coherent story text is left to the story generator, which is described in Section 5.

### 3.2 Event Structures

In narrating an individual's life story, the story should consist of text describing one or more events involving him/her. These events are taken from posts found in the person's Timeline. One way to detect them is by identifying the verbs in the posts. Some posts do not contain verbs, but instead use nouns to describe a certain action, e.g., a "lunch date" refers to an activity on eating.

Posts may contain *metadata* or descriptive details such as the date the event was held, the location of the event and the people the user was with during the occurrence of the event. These details have to be tracked along with the events. Table 1 shows a representation of an event.

A post can contain multiple sentences, with each sentence equating to zero or more actions. The user can describe how he/she managed to travel, dined with his/her friends and celebrated his/her birthday, therefore narrating three events in a single post. Events from such posts are identified individually by splitting the post into independent sentences and linking them by their postID.

Multiple relations may exist between two events, be it temporal (time sequence) or topical (topics or events, e.g. travelling). Temporal relations are resolved by checking the timestamps attached to the post. This is necessary when sequencing story events based on the time of their occurrence in the user's life.

Though causal relations are also common, they are usually found in stories where an action may lead to another action, such as falling down (cause) and then getting hurt (effect). In the case of Facebook, such descriptions may be found in the post itself, in the Comments section of a post, or even be shared as another post. Detecting causal relations is currently not included in this study.

**Table 1: Representation of an Event**

| Field | Description |
|---|---|
| postID | Reference to a specific FB post where the event was extracted from |
| postType | Category of the post, i.e., celebration, travel, or eating |
| sentence | The text of the post itself |
| verb | Verb or verb phrase from the post |
| noun | Noun or noun phrase representing the direct object from the post |
| tagged | List of friends that has been tagged in the post; these friends may or may not actually have been involved in the event described in the post |
| location | Place where the event took place |
| date | Date when the event took place |

Given a sample post *"Going to the mall."* posted on *"June 13, 2017"* at *"SM Mall of Asia"* with tagged friends *"Janine Tan and Bianca Regala"*, the corresponding event representation is shown in Table 2. The fields *postID*, *sentence*, *tagged*, *location*, and *date* will be filled with data from the post. The other information will be obtained after performing extraction of the event details found in the post.

**Table 2: Representation of an Event given a Facebook post**

| Field | Event Details |
|---|---|
| postID | 1 |
| postType | |
| sentence | Going to the mall. |
| verb | |
| noun | |
| tagged | Janine Tan, Bianca Regala |
| location | SM Mall of Asia |
| date | 06-13-2017 |

## 4 CLASSIFYING POSTS

Although Facebook's predefined activities feature is designed to enable users to easily classify their individual posts according to content, there are currently no available tools that can support the extraction of relevant elements from posts that use this feature. Furthermore, most Facebook users still prefer the traditional methods when crafting a post, i.e., typing text, and optionally combining photos and videos.

Given this limitation, we resort to using available tools for gathering posts from an individual user's Facebook account, preprocessing the posts, classifying posts according to their event types, and then extracting event details.

### 4.1 The Dataset

Posts comprising the dataset were gathered with full disclosure from the source user accounts and by utilizing Facebook's Graph API. This is a low-level, HTTP-based API developed by Facebook to be used primarily to access data and information from Facebook's platform [1]. The dataset consists of 21,412 posts from 216 user accounts. Manual inspection of these posts showed that 193 are about *eating*, 53 are regarding *drinking*, 409 are on *travelling*, and 643 described *celebrating* events.

Posts gathered from Facebook cannot be used directly in generating story text. Echoing the findings in the work of Kinsella et al., social media posts are usually brief, with users having a tendency to post snippets of incomplete, context-based data to show glimpses of their lives [6]. Moreover, the discussions that occur are more often than not informal, and there is a common tendency to resort to hyperlinks for context. These characteristics were evident in our dataset. Posts containing foreign characters, emoticons, laughter and hashtags abound. During preprocessing, these were removed as they currently have no relevance to the classification and the generation tasks.

Most posts are missing the actors or doers, in which case, the user who owns the account is assumed to be the actor. In certain cases, a post may contain multiple sentences. During preprocessing, Stanford CoreNLP takes care of splitting such post into its constituent sentences, and classification is performed on the individual sentences.

### 4.2 Extracting Event Details

As shown in Table 1, all events, regardless of their category, have the same basic elements. Existing NLP tools, such as Stanford CoreNLP, are used to identify POS tags, and to generate a constituent and dependency representation. From this output, syntactic analysis is performed to extract the necessary event details [9].

Consider the same post *"Going to the mall."*, relevant elements are extracted following these steps:

(1) *Extract the verbs* that signify the activity described in the post, and the *objects* or the recipient of the action, which may be another person or object. In this example, the verb is *"going"* and the object is the noun phrase describing the destination, *"to the mall"*.
(2) *Apply lemmatization* to transform words to their lemma in order to increase the accuracy of the classifier. In this case, *"going"* is lemmatized to *"go"*.

---

For posts that contain a verb, the fields *postType*, *verb*, and *noun (or direct object)* in the event representation can now be updated with the extracted information. The integer value for *postType* is determined based on the verb. Conceptually-related verbs for *Celebrating* posts will have a *postType* value of 1; the *postType* value 2 is assigned to the *Drinking* category; while for *Eating* posts, 3 is assigned; *Travelling* category gets the value 4; and for the other categories or no event posts, 0 is assigned. Table 3 shows the updated event representation that will be used later in life story generation.

[1] https://developers.facebook.com/docs/graph-api

**Table 3: Updated Representation of an Event after Extracting Event Details**

| Field | Event Details |
|---|---|
| postID | 1 |
| postType | 4 |
| sentence | Going to the mall. |
| verb | go |
| noun | to the mall |
| tagged | Janine Tan, Bianca Regala |
| location | SM Mall of Asia |
| date | 06-13-2017 |

However, the above steps do not work in all cases, such as in posts that have no explicit verbs used to describe the event. Given a sample post – *"Happy Birthday to the best dad ever!"*, the object is *"best dad"*; but there is no verb that can be used to determine its category. Posts with no verbs will have to rely on the event classification algorithm to determine its category, in this case, it is a *celebrating* post because of the keyword "birthday". In cases like this, only the field *noun (or direct object)* in the event representation will be updated. The last two fields, *postType* and *verb* will need to be updated again once the classification module determines the post's category.

## 4.3 Keyword-based Classification

A reference table (shown in Table 4) containing predefined keywords commonly associated with each event category was derived through manual inspection of the dataset. For *Celebrating* events, words which usually indicate special events such as *birthdays* and *Christmas* are used. For posts on *Travelling*, synonyms as well as methods of travelling are used. For *Eating*, aside from synonyms, the meals of the day are also used as indicators.

**Table 4: Rudimentary Classification of Events based on Keywords**

| Category | Keywords from Manual Inspection |
|---|---|
| Celebrating | birthday, celebrate, congratulations, congrats, God bless, bless, wish, happy, merry, party |
| Travelling | go, travel, at, visit, drive, road, place, far, run, walk, adventure, bucket list |
| Drinking | bar, glass, wine, beer, milk, thirst, water |
| Eating | cook, eat, dine, breakfast, lunch, dinner |

Since the current dataset is by no means exhaustive of all Facebook user accounts, the list of keywords is not complete, and needs to be expanded to improve the classification process. We explored the use of existing knowledge resources, specifically WordNet [11] and ConceptNet [8], to form our keywords list. The seed words fed to these two resources were the categories itself: *Celebrating*, *Drinking*, *Eating*, and *Travelling*. The semantic relations such as *"IsA"*, *"MadeOf"*, *"DefinedAs"*, and *"InstanceOf"* were utilized to derive related contexts for the categories. After integrating these two knowledge bases, 1,687 keywords have been derived across all event categories. Table 5 shows a breakdown of the number of keywords per event category and the sources. Table 6 shows some of the keywords derived from ConceptNet and WordNet.

**Table 5: Keywords Count per Event Category from External Resources**

| Event Category | Keywords from WordNet | Keywords from ConceptNet | Total Count |
|---|---:|---:|---:|
| Celebrating | 9 | 350 | 359 |
| Drinking | 24 | 492 | 516 |
| Eating | 17 | 400 | 417 |
| Travelling | 16 | 389 | 405 |
| **TOTAL** | **66** | **1631** | **1697** |

**Table 6: Updated Keywords List Derived from ConceptNet and WordNet**

| Category | Keywords from ConceptNet and WordNet |
|---|---|
| Celebrating | victory, Christmas, firework, toast |
| Travelling | traveler, journey, passport, fun, explore, pack |
| Drinking | toast, booze, liquid, bottle |
| Eating | feed, consume, chew, swallow, plate |

Because a single sentence can contain multiple verbs or words signifying events, the first iteration of our automated classifier classified this into multiple event categories. For example, the sentence *"Walking around the streets of Rome while eating delicious gelato."* was classified as an *eating* event and a *travelling* event. However, to avoid redundancy and the loss of context in downstream tasks, the classification algorithm has been revised to use a scoring system, and the sentence is assigned the category with the highest score. If multiple event categories bear the same score, a bias scheme based on the hierarchy of *celebrating* => *travelling* => *eating* => *drinking* will be followed. The hierarchy is based on the frequency count of each event category's occurrence in the dataset.

A threshold value of 2 was also set to minimize the occurrence of misclassification. Most posts contained at least two keywords such as "*drink coffee*", "*eat food*", "*happy birthday*", and "*Merry Christmas*". Setting the threshold value to 1 increases the likelihood of misclassification. However, increasing the threshold value to 3 would limit most of the posts resulting to under classification which means getting a high false negative rate. Consider the sentence *"I'd love to take a walk on the park someday."*. The presence of the word *walk* in the list of keywords led the no-score classifier to consider this sentence as a *travelling* event, when it should not have been the case. In the score-based classifier, only sentences such as *"I'm going on an adventure to check off one from the bucket list"*, which has a score of 3 because of the words *"going"*, *"adventure"* and *"bucket list"*, would be categorized as a *travelling* event.

Going back to the sample post *"Happy Birthday to the best dad ever!"*, using the keywords, this will be classified as a *Celebrating* post. However, even after determining its event type, the post still does not contain a verb. Thus, the verb will be derived based on the event label. Posts classified as *drinking* is assigned the verb *"drink"*, *"eat"* is used for posts classified as *eating*, *"travel"* for *travelling* posts, and *"celebrate"* for *celebrating* posts.

## 4.4 Performance Results

The dataset was subjected to manual labelling. Using the classification scheme in Table 4, 21,412 posts or 6.06% were tagged as actual events. The complete dataset, including the posts with no events, is then fed to the automated classifiers, without scoring and with scoring, to assess their performance.

In Table 7, the no-score automated classifier achieved a precision of 21.92% (the number of correctly classified event divided by the total number of classified events) and recall of 37.75% (the number of correctly classified events divided by the total number of actual events). The score-based classifier, on the other hand, has a precision of 45.02% and recall of 8.01%. While instances of misclassified events have been reduced, the recall is drastically low for the score-based classifier because of the use of a threshold value.

**Table 7: Results of Event Classification**

| Classifier | No-Score | Score-based |
|---|---|---|
| Precision | 21.92% | 45.02% |
| Recall | 37.75% | 8.01% |
| Accuracy | 88.08% | 93.83% |

After updating the keywords list based on the output of ConceptNet and WordNet, the no-score automated classifier achieved a precision of 9.58% and recall of 55.16%, as shown in Table 8. On the other hand, the score-based classifier achieved a precision of 10.60% and recall of 26.96%. Even though the recall has increased because of the additional keywords resulting to more posts being classified correctly, the precision decreased. While the new keywords list were derived directly from ConceptNet and WordNet, the list was not pruned to validate the relevance of the keywords to the category. For instance, some keywords in the *travelling* category, such as *businessman* and *scientist*, are not related to the category.

**Table 8: Results of Event Classification after Updating the Keywords List**

| Classifier | No-Score | Score-based |
|---|---|---|
| Precision | 9.58% | 10.60% |
| Recall | 55.16% | 26.96% |
| Accuracy | 65.72% | 81.78% |

Table 9 shows the performance of the no-score classifier for each category of events using the manually derived keywords and keywords from ConceptNet and WordNet. In the no-score classifier, except the others category, *celebrating* events achieved the highest precision (10.73%) and recall (42.12%) among the four event types. Events tagged as *celebrating* are more explicitly stated compared to the other types of events, as seen in the sample posts *"Happy anniversary to my parents."* and *"Merry Christmas!"*. Events under *drinking* and *eating* have low precision because posts in these categories are usually implied through the use of proper nouns, such as the name of a drinking place or the food, instead of the actual action. Since the list of keywords does not contain any proper nouns, our two classifiers cannot tag sentences such as *"At Yellow Cab!"* as *eating* and *"Enjoying my daily cup of Starbucks."* as *drinking*.

**Table 9: No-Score Classification Performance Results on Event Categories**

| Classifier | Precision | Recall |
|---|---|---|
| Travelling | 5.95% | 31.54% |
| Eating | 3.63% | 32.64% |
| Drinking | 4.72% | 28.30% |
| Celebrating | 10.73% | 42.12% |
| Others | 96.14% | 70.13% |

Table 10 shows the confusion matrix of the no-score classifier. Because the keywords list is not exhaustive, majority of the posts were tagged as *Others* even though the posts contain an action. In addition, *Eating* and *Drinking* posts were usually classified as *Travelling* or *Celebrating*. Users who posted *Eating* and *Drinking* activities are usually celebrating an event, such as birthdays or New Year; they can also be travelling somewhere. Consider the post *"Eating Media Noche to celebrate New Year!"*; the score for *Celebrating* category, 2, is higher than the *Eating* category, 1, because the keywords list does not contain *"Media"* and *"Noche"*.

**Table 10: No-Score Classification Confusion Matrix where C - *Celebrating*, E - *Eating*, D - *Drinking*, T - *Travelling*, and O - *Others***

| | C | E | D | T | O |
|---|---|---|---|---|---|
| **C** | 270 | 20 | 14 | 26 | 2187 |
| **E** | 30 | 63 | 5 | 65 | 1573 |
| **D** | 11 | 7 | 15 | 11 | 274 |
| **T** | 40 | 17 | 7 | 129 | 1975 |
| **O** | 290 | 86 | 12 | 178 | 14107 |

Table 11, on the other hand, shows the performance of the score-based classifier for each category of events using the manually derived keywords and keywords from ConceptNet and WordNet. In the score-based classifier, still not including the *Others* category, *Celebrating* events still achieved good precision and recall values. The threshold did not affect the classification because most posts contained at least two of the *Celebrating* keywords, such as *"happy"* and *"birthday"*. This time, events under *Eating* and *travelling* have low precision following the same problems identified previously. Should the post be stated as *"eating pizza at Yellow Cab"*, the threshold would have been met with the keywords *"eating"* and *"pizza"*.

Table 12 shows the corresponding confusion matrix of the score-based classifier. Similar to the observations in the no-score classifier, many posts were still classified as *Others*. On the other hand, misclassification of *Eating* and *Drinking* posts as *Celebrating* and *Travelling* have been reduced. Only keywords with the closest relation to the categories were used and keywords that are far in relation to the category were removed, thus improving the classification.

As 13.89% of *Eating* and 12.50% of *Drinking* posts without including those misclassified as *Others* are misclassified as *Celebrating* posts, future works should consider combining these three into one category to reduce the overlap and to increase the chances of correct labels.

**Table 11: Score-based Classification Performance Results on Event Categories**

| Classifier | Precision | Recall |
|---|---|---|
| Travelling | 7.85% | 17.36% |
| Eating | 7.53% | 30.05% |
| Drinking | 11.82% | 24.53% |
| Celebrating | 12.70% | 32.35% |
| Others | 95.40% | 85.32% |

**Table 12: Score-Based Classification Confusion Matrix where C - *Celebrating*, E - *Eating*, D - *Drinking*, T - *Travelling*, and O - *Others***

| | C | E | D | T | O |
|---|---|---|---|---|---|
| **C** | 208 | 15 | 13 | 26 | 1376 |
| **E** | 13 | 58 | 4 | 19 | 676 |
| **D** | 3 | 6 | 13 | 2 | 86 |
| **T** | 12 | 7 | 0 | 71 | 815 |
| **O** | 407 | 107 | 23 | 291 | 17161 |

Table 13 shows example posts and their classifications. The first post is classified correctly, from the keywords *"happy"* and *"birthday"*. The second post is also correctly classified, from the keywords *"drinking"* and *"tea"*. The third post, however, was misclassified by the score-based system because it has insufficient keywords to satisfy the threshold. The last post was misclassified by both classifiers due to the keywords *"drive"* and *"adventure"* which pertain to *travelling*, but the context of the whole post is not.

**Table 13: Sample Posts and their Classification (NS – no-score classifier; SB – score-based classifier; Act – actual classification)**

| Post | NS | SB | Act |
|---|---|---|---|
| "Happy birthday to my favorite sister!" | Celebrating | Celebrating | Celebrating |
| "Drinking tea on a Sunday morning" | Drinking | Drinking | Drinking |
| "Drinking Swiss Miss on a cold day." | Drinking | No Event | Drinking |
| "I'd love a good drive as an adventure." | Travelling | Travelling | No Event |

## 4.5 Working with Noisy Data

In comparison to previous studies that used machine learning approaches in classifying social media posts [3] [6] and tweets [4], to achieve a 55% to 80% precision and 80% to 90% recall, the low precision and recall values of our classifiers can be attributed to a number of problems when dealing with Facebook data. Examples of posts that posed challenges to our classifier are found in Table 14. Post #1 contains mixed language; #2 is an excerpt of a song; #3 refers to a restaurant; while #4 is referring to a different context.

**Table 14: Sample posts that caused classification problems due to the presence of noisy data.**

| Original Post | Classified As |
|---|---|
| #1 Starbucks after school :) KAPE PA | Not an Event |
| #2 If I go there's just no telling how far I'll go | Travelling |
| #3 Here at Eat and Go! | Eating, Travelling |
| #4 Cooking with Chef Curry | Eating |

First, many people do not explicitly state an activity that is needed to detect an event, as exemplified by posts #1 and #3 wherein only the restaurant name or food name is given. Both posts should have been categorized as *Eating*. The missing verb in post #1 and the presence of the keyword *"go"* in the restaurant name in post #3 caused the classification problem.

The second challenge is the use of mixed languages (post #1), which is a common practice in countries where English is not the first language. Similar observations were made by Pippin et al. [15] and Andrei, Elson, and Zarrella [1] in their studies.

A third challenge is context. A *"have a nice day"* post, for example, can have a positive or a negative (sarcastic) sentiment which is only clear to the sender and the receiver of the message. Post #4 in Table 7 is another example where *"cooking"* is not the literal meaning; the post is referring to an informal basketball term. Context can also be sourced elsewhere, such as object metadata from accompanying photo or video, and comments [6] .

Lastly, as pointed out by Cavalin et al. [3], only a very small fraction of the dataset is identified as event posts. In our dataset, only 6.0% are tagged as events under the category of eating, drinking, travelling and celebrating. Other events, such as listening, watching and reading, are not included in our current study. A large volume of non-events posts is on a variety of things that include quotes, song lyrics (post #2 in Table 14), and re-shared posts and videos from others that the user found interesting. Because of this, relying on Facebook posts alone to extract relevant data as a means to generate one's life story may not be enough to paint a very clear picture of the subject.

# 5 GENERATING LIFE STORIES

Once the events have been classified and relevant details have been extracted from posts retrieved from a single user account, we then attempt to generate a life story to investigate the sufficiency of Facebook posts as a data source for story generation. We briefly present our story generation process that follows the NLG pipeline of Reiter and Dale [16], consisting of content determination, story planning, and surface realization.

Our description of a life story in Section 3 leads us to divide its structure into three main parts - the introduction, the body and the conclusion. Content determination for the *Introduction* and the *Conclusion* involve deriving data supplied by the users through their

Facebook profile. These include information about themselves in the *About Me* section (birthday, family members, educational and work background); and and the types of Facebook pages that they *Liked*, such as *Local Business or Place, Company Organization or Institution, Brand or Product, Artist, Band or Public Figure, Entertainment and Cause or Community.* Story planning for these two parts is minimal and relies on a predefined story grammar or script to sequence these data into a coherent story text that describe the user as the main story character.

The bulk of the work for the story generator is in producing the text for the body of the life story. Content determination involves utilizing the events that were derived from processing and classifying the posts. Story planning handles the organization and sequencing of the events into a coherent story plan by taking into account the temporal and the topical relations of events. Topical relations are used to generate paragraphs, wherein one topic (or event category) equates to one paragraph. Within each paragraph, events are ordered based on their temporal relations, which are determined from the timestamps attached to each post and linked to the corresponding events. Surface realization converts the verb entries in the story plan into a sentence to express the date(s) of occurrence, as well as the people and places involved in each event. Surface realization also handles the aggregation of related events together, i.e., those with closer temporal relations or those with the same people involved. *Closer* temporal relations mean either the same date, the same month or the same year.

Issues encountered in the generation of the introduction and conclusion include the readability of the extracted dates, i.e., earlier iterations generated "5/25/1996"; the presence of special characters, such as those used in the names of some Facebook pages (i.e., "Samgyupsalamat - 삼겹살라맛"); and inconsistencies in the use of pronouns. To resolve these problems, the name of the month in a given date is used, the special characters are removed, and the gender of the user is used to determine the proper pronoun to be generated, respectively.

In generating the body of the life story, the problems encountered stem mostly from difficulty in parsing posts that are mostly informal in nature. Some posts are parsed incorrectly. Consider the sample celebrating posts in Table 15. The term *"thesismate"* in the third post was tagged as the object of the sentence, yielding the sentence *"Mae celebrated thesismate with Cam."* In another post, laughter (*"HAHA"*) was treated as a proper noun because of the use of capitalization, hence leading to the sentence *"Mae celebrated Jamie HAHAHA with Jamie."*

There are also numerous instances where the informal or brief nature of posts caused errors for Stanford CoreNLP's parser. One workaround for this is to add rules in the story planner to change the tags of some portions of the parse tree, i.e., *"thesismate"* should not be tagged as the event being celebrated, and *"18th"* should not be tagged as a verb.

## 6 CONCLUSION AND FURTHER WORK

In this paper, we described our approach in classifying a user's Facebook posts in order to identify events related to travel, dining, drinking, and celebrating. This is to determine the potential use of Facebook posts as a data source to generate an individual's life story, focusing on his/her profile, interests and events.

**Table 15: Sample Facebook posts classified as celebrating, and their Metadata**

| Original Post | Metadata | |
|---|---|---|
| Happy 18th Angie! | date created | 10/03/14 |
| Happy anniversary Jamie HAHA | date created | 02/07/17 |
| | user tagged | Jamie |
| Happy friendversary thesismate! | date created | 02/14/17 |
| | user tagged | Cam |
| Party party! | date created | 08/16/16 |
| | user tagged | Shane |
| | location | Manila, Philippines |

Our dataset, however, showed that only a very small fraction of the posts, specifically, 6.06%, contain events. Our results also show that there is no correlation between the number of posts per event category to its precision and recall values. The classification hinges largely on the textual content of the post.

We rarely encountered posts that provide sufficient data from which useful details about the event can be extracted. These posed numerous challenges to our classifier, which, despite achieving a reasonable accuracy of 88.08% and 93.83% for no-score and score-based approaches, respectively, also had low precision and recall values. This can be attributed to several factors. First, not many people state explicitly the verb to describe the activity they are doing. Second, short phrases and the use of mixed languages mess up the syntactic structure of the post, making it difficult for the parser to properly perform POS tagging. Third, the different context implied by a post further challenges the interpretation of a statement.

It was also observed that newer posts may have dependency to older posts. Future research can explore the use of graph models to represent the dependent relationships among the posts, such as common tagged friends, location and event date, which can then be utilized during the classification task. For example, the user first posts about riding a plane to take a vacation in another country. A few hours later, he/she may post about arriving at his/her destination. These two posts should both be classified as *Travelling*.

Relying on manually-built keywords also posed a knowledge base insufficiency issue. While we explored the use of existing knowledge resources, specifically WordNet and ConceptNet, to form our keywords list, the findings we reported here did not yield promising results. We may have to look into pruning the resulting keywords, and doing phrasal matching. Aside from focusing on building the keywords list, we can also look at how emojis could help in event classification. There is a growing number of Facebook users who use emojis that are related to the category of the post. For example, the use of emojis like *cake* or *gift* could determine that a post is a *Celebrating* event.

The insufficient details provided in the posts, combined with the inability of the extraction algorithm to derive correctly tagged elements, makes it difficult for the story generator to produce meaningful story text. The use of other sources to provide additional context to a post should be investigated to support this task. This includes processing metadata from embedded objects (photos and videos), and using online resources to determine if a given text is a quote from a book, a song lyric, or a line from a movie.

# REFERENCES


[1] Amanda Andrei, Sara Beth Elson, and Guido Zarrella. 2015. *Language and Emotion in Philippine Twitter Use During Typhoon Haiyan.* Technical Report. Mclean, VA. https://doi.org/10.13140/RG.2.1.4671.5923

[2] Karen Ang and Ethel Ong. 2012. Planning Children's Stories Using Agent Models. In *International Pacific Rim Knowledge Acquisition Workshop (LNCS 7457)*, D. Richards and B.H. Kang (Eds.). Springer, Verlag Berlin Heidelberg, 195–208. https://doi.org/10.1007/978-3-642-32541-0_17

[3] Paulo Cavalin, Luis Moyano, and Pedro Miranda. 2015. A Multiple Classifier System for Classifying Life Events on Social Media. In *Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW).* Institute of Electrical and Electronics Engineers, 1332–1335. https://doi.org/10.1109/ICDMW.2015.182

[4] Smitashree Choudhury and Harith Alani. 2014. Personal Life Event Detection from Social Media. *Workshop Proceedings, CEUR* (2014).

[5] Angel Daza, Hiram Calco, and Jesús Figueroa-Nazuno. 2016. Automatic Text Generation by Learning from Litarary Structures. In *Proceedings of the NAACL Human Language Technology 2016 5th Workshop on Computational Linguistics for Literature.* Association for Computational Linguistics, 9–19.

[6] Sheila Kinsella, Alexandre Passant, and John G. Breslin. 2011. Topic Classification in Social Media Using Metadata from Hyperlinked Objects. In *European Conference on Information Retrieval, ECIR 2011 (LNCS 6611)*, P. et al. Clough (Ed.). Springer, Berlin Heidelberg, 201–206.

[7] Hugo Liu and Push Singh. 2002. MakeBelieve: Using Commonsense Knowledge to Generate Stories. In *Proceedings of the 18th National Conference on Artificial Intelligence.* Association for the Advancement of Artificial Intelligence, 957–958.

[8] Hugo Liu and Push Singh. 2004. Commonsense Reasoning In and Over Natural Language. In *International Conference on Knowledge-Based Intelligent Information and Engineering Systems (LNCS 3215).* Springer, Verlag Berlin Heidelberg, 293–306. https://doi.org/10.1007/978-3-540-30134-9_40

[9] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny R. Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics, 55–60.

[10] Neil McIntyre and Mirella Lapata. 2009. Learning to Tell Tales: A Data-driven Approach to Story Generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* Association for Computational Linguistics, 217–225.

[11] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[12] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures. In *Proceedings of the NAACL Human Language Technology 2016 4th Workshop on Events.* Association for Computational Linguistics, 51–61.

[13] Brian O'Neill and Mark Riedl. 2014. Dramatis: A Computational Model of Suspense. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Vol. 2. AAAI Press, 944–950.

[14] Ethel Ong. 2010. A Commonsense Knowledge Base for Generating Children's Stories. In *Proceedings of the 2010 AAAI Fall Symposium Series on Common Sense Knowledge.* 82–87.

[15] Michael M. Pippin Jr, Ron Jairus C. Odasco, Ronald E. De Jesus Jr, Miguel Angelo Tolentino, and Rex P. Bringula. 2015. Classifications of Emotion Expressed by Filipinos through Tweets. In *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, Vol. 2215. Newswood and International Association of Engineers, 292–296.

[16] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems.* Cambridge University Press, New York.

[17] Mark O. Riedl and Robert Michael Young. 2010. Narrative Planning Balancing Plot and Character. *Journal of Artificial Intelligence Research* 39, 1 (2010), 217–268.

[18] Elena Rishes, Stephanie M. Lukin, David K. Elson, and Marilyn A. Walker. 2012. Generating Different Story Tellings from Semantic Representations of Narrative. In *International Conference on Interactive Digital Storytelling (ICIDS 2013) (LNCS 8230)*, H. Koenitz, T.I. Sezen, G. Ferri, M. Haahr, D. Sezen, and G. Catak (Eds.). Springer International Publishing, Switzerland, 192–204. https://doi.org/10.1007/978-3-319-02756-2_24

[19] Shankar Setty, Rajendra Jadi, Sabya Shaikh, Chandan Mattikalli, and Uma Mudenagudi. 2014. Classification of Facebook News Feeds and Sentiment Analysis. In *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI).* Institute of Electrical and Electronics Engineers, 18–23. https://doi.org/10.1109/ICACCI.2014.6968447

[20] Candice Jean Solis, Joan Tiffany Siy, Emerald Tabirao, and Ethel Ong. 2009. Planning Author and Character Goals for Story Generation. In *Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity.* Association for Computational Linguistics, 63–70.

[21] Leif Romeritch Syliongka, Nathaniel Oco, Alron Jan Lam, Cheryll Ruth Soriano, Ma. Divina Gracia Roldan, Francisco Magno, and Charibeth Cheng. 2015. Combining Automatic and Manual Approaches: Towards a Framework for Discovering Themes in Disaster-related Tweets. In *Proceedings of the 24th International Conference on World Wide Web.* Association of Computing Machineries, New York, 1239–1244. https://doi.org/10.1145/2740908.2742125

[22] Jeff T. Titon. 1980. The Life Story. 93, 369 (1980), 276–292.

[23] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. In *Proceedings of the ACL 2012 System Demonstrations (ACL '12).* Association for Computational Linguistics, Stroudsburg, PA, USA, 115–120.

[24] Laura E. West. 2013. Facebook Sharing: A sociolinguistic Analysis of Computer-Mediated Storytelling. *Discourse, Context & Media* 2, 1 (2013), 1–13.

[25] Youse. 2005. Ten Elements of Biography. (2005). mybrary.wikispaces.com/file/view/Ten_Elements_of_Biography.ppt