

# Meliorating usable document density for online event detection

Manisha Samanta, Yogesh Kumar Meena, Arka Prokash Mazumdar, Girdhari Singh,  
Dinesh Gopalani

Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, Jaipur, India

## Article Info

### Article history:

Received Dec 24, 2021

Revised Mar 8, 2022

Accepted Mar 26, 2022

### Keywords:

Anaphora

Event clustering

Online event detection

Pronoun resolution

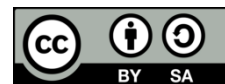
Social media

Streaming data

## ABSTRACT

Online event detection (OED) has seen a rise in the research community as it can provide quick identification of possible events happening at times in the world. Through these systems, potential events can be indicated well before they are reported by the news media, by grouping similar documents shared over social media by users. Most OED systems use textual similarities for this purpose. Similar documents, that may indicate a potential event, are further strengthened by the replies made by other users, thereby improving the potentiality of the group. However, these documents are at times unusable as independent documents, as they may replace previously appeared noun phrases with pronouns, leading OED systems to fail while grouping these replies to their suitable clusters. In this paper, a pronoun resolution system that tries to replace pronouns with relevant nouns over social media data is proposed. Results show significant improvement in performance using the proposed system.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Manisha Samanta

Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur

Jaipur, Rajasthan, India

Email: 2015rcp9524@mnit.ac.in

## 1. INTRODUCTION

The recent few years have seen a boom in the number of users on social media platforms and the amount of data produced in them. In 2017, there were a reported 288 million [1] social network users. The data produced from all these users vary from being promotions to thought sharing, or sometimes reporting about an incident or event that the user experienced directly or indirectly. All of the information posted is subject to their own experience, understanding, research, and findings. Monitoring events over social streams have many applications, such as crisis management and decision making. Owing to human nature and psychology, this information is subject to social media sites as soon as the incident occurs or even as the incident is occurring. In contrast to the traditional sources of information and news (such as TV stations, radio stations, and newspapers) that have to undergo a background check on the incident before reporting that usually takes time, monitoring the information on social media through users' posts may ensure that not only will the information be available quicker, but it is also possible getting firsthand information from a larger number of [1] social media participants. For instance, during the recent earthquake in Nepal, with the local media information unavailable to them, the international media took some time to validate the rumors and move in to report them. However, at the same time, in a time window of about a few hours, many tweets, Facebook status, pins on Pinterest were being shared with firsthand reporting of the events. If the information in these posts can be analyzed, a lot of firsthand information will be available in a short time.

This property of social media reporting gave rise to a new area of research, named online event detection (OED) [2]–[4], where a system tries to identify a possible event as quickly as possible through the recent posts. Most of these systems generally work in three phases. First, it continuously accumulates

documents (the user posts) from a particular social media platform. These systems then perform clustering based on an existing or proposed similarity metric. Finally, they indicate the significant clusters as possible events. For the second phase, most of these systems try to group the user posts based on the textual and contextual similarities among the documents [2], [4]–[6]. This implies that the words used in these documents are of utmost importance for correctly identifying their association with a particular event. A closer look at the popular social media platforms can show that one of the common properties across the platforms is the reply. This particular property can adversely affect the performance of such systems as the reply documents, at times, tend to replace important words with pronouns. As a document replaces a particular word with a pronoun, adding the document does not increase the number of occurrences of that word in the cluster, thereby not increasing the strength of this word. Therefore, the impact of this document reduces with respect to the cluster in question.

This article focuses on this particular property of social media to increase the usability of such documents and inherently increase the density of usable documents for the purpose. The critical aspect here is to update these documents such that they are appropriately clustered in their relevant events. It can be seen that the existing articles ignore the same and try to compensate with a more significant number of documents, where a higher number of documents also indicate added delay and higher memory requirement. In this paper, we present a method to augment the density of the event clusters by performing pronoun resolution to each incoming documents before applying OED algorithms. The primary contributions of this article are as: i) the proposed work aims to perform a parts-of-speech-based pronoun resolution between the reply document that contains the set of pronouns and the main document to which this document was posted. The resulting document can greatly affect the similarity measurement between the main and the reply document as more similar words will be present; ii) there can be a hierarchy of main-reply documents, and the pronoun may not correlate to the direct parent. The approach, therefore, traverses the hierarchy to identify the most appropriate word for the pronoun in question; and iii) as the system aims to perform in run-time, searching a considerably large hierarchy of documents for antecedent words may significantly delay the outcome. Therefore, we propose a selection window-based antecedent search so as to reduce the possibility of unaccounted delay.

We use the Twitter platform for simulation purposes as the data from this platform are easily downloadable, and many such datasets are also available. The experimental results demonstrate that the resulting dataset from the proposed approach significantly increases the performances of the existing OED algorithms compared to the raw-data usage. The rest of the article is organized as follows. Section 3 discusses the background and related works of OED algorithms and pronoun resolution techniques. In section 3, we propose our information augmentation technique for pronoun resolution for the streaming data. The results of the proposed algorithm are analyzed and discussed in section 4. Finally, the section concludes the paper and provides future directions in the area.

## 2. BACKGROUND AND RELATED WORK

The term 'Event' carries a diverse range of meanings depending on the context, but the idea behind it remains unchanged. Anything important, unfamiliar, or abnormal happenings in a normal context can be treated as an event. This could be a part of a chain of incidents as an effect of a preceding or as the cause of succeeding incidents. The process of discovering these happenings or incidents through social media data is termed 'online event detection'. An event detection technique [7], [8] can be either proactive, where categories and properties of possible event types are known, or reactive where the systems assume that an event type may be previously unknown. Primarily, any such system has four major processing phases: gathering raw data, preprocessing, term weight estimation, and grouping documents based on their similarity. Traditionally, events were detected by gathering [3] historical data such as news articles, radio broadcast articles, and classified columns. Due to the advent of social media streams, an enormous amount of real-time data has become obtainable [1]. The present situation is when an event occurs, a majority of times they are first and foremost posted on various social media sites (SNS), and then the news follows. Hence, highlighting the importance of social media streams in event detection, as a result, SNS gained popularity as a perfect source of event detection. Recent works such as Dakle *et al.* [9] on co-referential methodology performed on emails related text, Wright-Bettner *et al.* [10] on cross-document co-reference, where the goal of authors in [11] specifically on mission-related objects, locations, and actors, therefore, annotate a dataset of reference links, with the inclusion of coreferences.

### 2.1. Online event detection

Discovering topics from a document, or any kind of contextual source of information was initiated by topic detection and tracking (TDT) [3] project. This concept further moved on to integrate the time-

varying property [12] by incorporating online data from social streams. Data gathering plays an important role in OED. Without hampering the fundamental meaning of the input data streams, upgrading it is the core task of our work. The choice of data sources moved gradually from traditional media (news articles, T.V. broadcast, and article stories) to blogs, e-mails, and micro-blogs [4], [13]–[16]. due to the limited dataset availability on specific topics. The notable changes in data sources can be observed with the appearance of SNS [4], [5], [16], [17]. These steers drastic changes in event detection techniques due to their distinct characteristics. Twitter gained the highest popularity [1], [5] in the OED research field as it is very convenient to get the metadata based on certain requirements and can easily be gathered through their application program interfaces (APIs). To design a well-shaped input stream, text cleaning plays a significant role as raw data are highly unstructured, flooded with irrelevant texts, unimportant words, and spelling mistakes as people replicate the spoken language in textual form. Hence, it is essential to filter out those words or a set of words to reduce further processing complexity.

Becker *et al.* [4] detected potential events from Twitter data through incremental clustering by estimating the similarity using the *tf-idf* term weighing metric. They performed basic cleaning operations, such as stop-word removal, and stemming on tokenized terms. Hasan *et al.* [5] perform event clustering using two modules, firstly, the search module identifies the unique tweets, and secondly, an event clustering module that processes tweets that are not unique for the existing clusters by estimating the similarity between them. Elimination of user names, mentions, stop-words, and web links are done as pre-processing before applying the proposed event detection method. Nguyen *et al.* [1] first proposed to normalize tweets to get all potential terms, then to monitor the occurrences of terms, generate signals for each of them, and finally extract features to estimate the similarity of tweets to create potential clusters. To normalize the tweets, all the terms are converted to lower case letters, extra spaces are eliminated, removed multiple repetitions of letters within words, user mentions, URLs, and hash symbols. Guille and Favre [18] produced a list of events, where each event is described by three different things-the main word, a set of weight-related words, and the time-span and magnitude of events. Statistical methods and external sources, excluding social media streams are used to improve the accuracy. They also applied stop-word removal techniques using a normal stop-word list for text processing. Nikolaos *et al.* [15] talked about long-term and short-term event detection scenarios from the perspective of time. They redefined the inverse document frequency (IDF) score for the time-varying scenarios with fuzzy representation. Using natural language toolkit (NLTK) in python, their proposed technique counted the term frequencies, synonyms and abbreviations are handled using the lexical database of WordNet. It also removed URLs, slang words, common verbs, and nouns. The concept of high utility pattern mining [17] has been taken to detect event topics in this work. For pre-processing, the raw tweets of all the characters are converted into lower-case letters, performed tokenization, stop-word removal, dealt with HTML tags, removed URLs and special symbols. The method in [16] detects events through incremental clustering by checking similarity among documents. A new similarity metric is introduced embedding structural property with the textual property. To pre-process, these basic natural language processing (NLP) steps such as, stemming of words, and removal of stop-words are performed. A latent event and a category model are developed and proposed in [19] to discover the events from Twitter data. NLP tools, such as word-tokenizer, part of speech (POS) tagger, stemming, named entity recognition (NER) mapping of ‘today’, ‘tomorrow’ with the published date of tweets are performed.

We observed that, though the existing works for OEDs may follow different methodologies to identify or discover the events, the initial tasks of pre-processing are very much similar for almost all of them. Moreover, it can also be seen that the OED system mostly uses a term weighing metric for the clustering process, thereby establishing the importance of common words to be used in all the relevant documents so that they can be correctly clustered. Therefore, the documents where the prominent noun phrases are replaced with pronouns will suffer from reduced term weights, resulting in wrongly clustering these documents. Most of the OED systems would eventually suffer from this problem as they do not resolve the pronouns, and also, they remove these pronouns in their pre-processing phase, thereby eliminating the possibility of pronoun resolution further. A system that can resolve these pronouns before pre-processing, may therefore increase the efficiency of any such OED system by providing more relevant words in the documents. In the field of linguistics, this technique is termed anaphora resolution (AR). So far anaphora resolution is used in the retrospective datasets and dialogue textual documents. The motivation of our work is to increase the data density of streaming data through pronominal resolutions which make the process more challenging. The next section is a brief description of existing works of pronominal resolution found in the literature.

## 2.2. Identification of subjective words

In linguistics, both coreference resolution and anaphora resolution techniques are related in the context of identifying an antecedent or a subjective term of the referred term such as pronoun [20]. The choice of computation of anaphora resolution has shifted to the machine learning approach from heuristics

approaches due to the advancement of statistics in linguistics and publicly available annotated corpora. The corpora, that are retrospective or historic [21] in nature, consist of news stories [22], technical manuals [23], or conversational text [24]–[26]. Next, we will emphasize these areas.

In 1972, the first pronominal resolution approach was proposed by Winograd [27]. All the preceding noun phrases are considered as a candidate of an antecedent and rated based on their syntactic positions. Another early work on the syntactic constraint-based approach is Hobb's algorithm [28]. The algorithm searches the possible antecedent of the pronoun on a tree. The algorithm evaluated a corpus of news articles. The concept expanded and was enhanced gradually by appending several other features. A multi-stage approach [29] considered multiple knowledge of sentential syntax, case frame semantics, dialogue structure, and general world knowledge to resolve anaphors. The identification of antecedent of third-person pronouns (he and she) and lexical anaphors are taken into account by Lappin and Leass approach [30]. A revised and updated version of this approach is developed by Kennedy and Boguraev [22] as the algorithm runs on the output of the POS tagger instead of full, in-depth syntactic parsing. The robust approach by Mitkov [20] approach identifies antecedent noun phrases within two preceding sentential distances from the anaphor. As an input, a text is passed through a POS tagger through the antecedent scoring system antecedents are identified. The machine learning approach first introduced by Connolly *et al.* [31], aims to resolve the referred word. To identify two coreference noun phrases, a binary classifier or mention pair model has been proposed. The classifier can be trained by any off-the-shelf learning algorithm such as support vector machine (SVM), maximum entropy, and deep neural network. To maintain the transitivity property of coreference resolution, a separate clustering (agglomerative or graph partitioning) is performed to select the closest preceding antecedent (by closest first clustering) or the highest coreference likelihood (best first clustering) antecedent. The graph partitioning approach in [21], [32] also employed for this purpose where the nodes represent antecedents, and the edges represent the possible weights of connected antecedents. To train the classifier MUC6, MUC7, and ACE5 corpus are taken. Hence, the nature of the dataset is static or retrospective.

Turn structure is an organization of dialogue structure where people speak in an alternative manner, one by one. This feature gained huge attention [26], Stent and Bangalore [25] worked on a specific conversation-based dataset to improve the relative performance. A detailed study of the turn structures is discussed in detail by [33], using a specific dataset of dialogues on the tutoring system. The procedure considered the location information of the candidate antecedent, to analyze the corpus. Ritter *et al.* [34] used named entity recognition (NER) on tweets as the pre-processing task which may lead to better results in performance, but the linking between pronoun with its corresponding subjective nouns are not familiar. Resolution of first-person pronouns (I, me, and mine), and second-person pronouns (you and yours) can be observed in [24] on a static dataset. To resolve pronouns with non-nominal antecedents, switchboard corpus is used in [35], however, they fail to highlight the need of using such features.

### 3. INFORMATION MELIORATION IN OED

During clustering of the documents, a lot of information is lost or is not clustered properly due to the absence of relevant nouns or other words in the document. All the main posts related to an event mostly have a similar matter or carry similar information. New information is added as new documents or reply messages and is pushed into the relevant cluster. People use reference terms such as pronouns or abbreviations to refer to the main subject while writing responses/replies. Hence, the significance of nominal terms can be enhanced by establishing a link with their pronouns. Generally, pronouns are eliminated during pre-processing or data cleaning by a stop-word removal technique; sometimes abbreviations are also filtered out at this phase. Due to this phase, many of these cleaned documents are wrongly clustered. Before removal of stop-words, finding the association between reference words in the reply documents and their subject in the main posts may enhance the clusters' information density. So far, very few existing works in the field of event detection have given importance to this. Therefore, the main aim of this work is the augmentation of information density of event clusters by establishing the relation between reply posts with their main posts for streaming data from micro-blogging online social networking sites. For convenience, we consider Twitter as datasets that are available and can further be downloaded as streaming data, though the proposed system can be applied to any such platforms that allow posting replies.

Existing AR pronoun resolve approaches are based on static data [23], [25], [26]. These algorithms can process each document from the beginning to the end of the dataset in each of the iterations. However, in streaming data this is not feasible, making it not suitable for streaming data. To apply these methods, buffering of streaming documents become necessary before the process, thereby inducing a delay in processing and huge memory requirement (on avg. 350,000 tweets posted per minute). Therefore, the need for an A.R. algorithm that is less dependent on historical documents is needed for storing data. Moreover,

rule-based A.R. approaches are best suited for the purpose as they are less dependent on large corpus by performing pronominal resolution on a limited and fixed number of antecedent statements.

### 3.1. Pronoun resolution to meliorate document density

The proposed approach is of two folds. Firstly, it prepares a text window of a limited number of reply messages for each of the primary documents, and secondly, it tries to resolve pronouns of the reply messages within this text window. Social media text datasets, a dataset from Twitter, for example, represent streaming texts that are knowledge-poor by nature, that is, the texts are not POS tagged. Here we propose an antecedent tracking technique that can be applied to a set of all possible candidates found in the text window of a particular number of previous tweets. An aggregate score is then assigned to all the possible candidates for the antecedent of a given pronoun after the successful execution of multiple scoring criteria. The candidate, that achieves the maximum aggregate score, is treated as the antecedent. Therefore, POS tagging plays a major role in this scoring system.

All the nouns of the text frame and all pronouns of the current post are identified first. The set of all nouns and pronouns, or in other words, referencing terms, are listed out. A list of all nouns is taken as the input of the scoring system after checking for animacy, gender, and number agreement. As we are considering streaming data, the number of previous documents to consider can prove to be very large, therefore it is important to limit the preceding reply window for antecedent tracking. We propose a window size  $W$  to set up the window size of the preceding related reply tweets depending on the applicative nature of implementation. The value of  $W$  may affect the antecedent tracking system in the following way. A large  $W$  will cause increased delay, however, may have more candidates for tracking. On the other hand, a smaller  $W$  may reduce the candidates, but will cause smaller delays, which may prove beneficial for real-time tracking. The study of  $W$  is out of the scope for this article and for the rest of the paper, we assume the value to be 3. We assume a smaller value since in social media the readers normally can view only the most recent replies therefore, they tend to write replies on these visible documents. Even if suitable tweets for the window are not found, intra-sentential referencing has to be performed and proceed next to maintain the online environment. Here, we use the term 'tweet', 'post', and 'sentence' interchangeably.

In the proposed method, we consider the '@' symbol as users directly refer to other persons by tagging them with this special character. We also consider the 'hashtag' (#) symbol, as most of the '#' tagged words are noun phrases, and are generally used to highlight some significant matter in social media. The following method is performed on the text window  $W$  for a particular document. Let us consider the window of document  $d$  as  $W_d$ . First-person singular pronouns like 'I', 'me', and 'my'. can be replaced with the 'username' of the post. Similarly, the first-person plural pronouns like 'we', 'our', and 'ours', can be resolved if the '@' symbol is present in the current tweet. These pronouns should be resolved by the 'username', who has posted the tweet, and all the user names who are mentioned in the tweet using the '@' symbol. The pronoun 'it' usually refers to non-living entities, and can be resolved according to the animacy agreement [23]. To resolve second person pronoun like 'you' in case no '@' symbol is found in the current tweet, it can be resolved to the user in the tweet for which the current tweet is a reply. On the other hand, if '@' is present in the tweet, then the pronoun 'you' can be resolved by all the usernames mentioned using the '@' symbol. The phrases mentioned with the symbol '#' are considered as the candidates for the noun phrases to which the pronouns can be resolved, where the abbreviations are also considered as the candidates for the noun phrases. Next, we discuss the design, formation, and usage of the text window  $W_d$ .

#### 3.1.1. Designing the reply window

As mentioned before, the sentential text window of size  $W_d$  is proposed here. This window represents the combination of the main tweet and the corresponding recent reply tweets. The pre-processing steps, such as a basic cleaning like removal of punctuation, conversion of all the letters to lower-case letters, removal of URLs, and stop words are performed after fetching the first tweet. Through POS-tagging, noun phrases in all the tweets within the frame are identified. The output of POS-tagger is used for animacy, gender, and number agreement checking, afterwards, the antecedent indicator scoring system is invoked to assign scores as discussed before. Algorithm 1 provides the algorithmic description of the proposed approach. Algorithm1 first creates a list of most recent tweets that have parent-child relations, 'sentence [1]' being the latest and the target tweet. The functions 'preprocess( $d$ )' and 'fetchParent( $d$ )', perform pre-processing as discussed before and fetches the parent tweet for the document  $d$ . It should be noted that the function 'fetchParent( $d$ )' can be implemented in two ways. The simplest way would be to fetch parent tweets online when required and run all the steps as the algorithm suggests, however, this will result in a slower implementation as it will depend on the communication latency. The other way is to store the data locally after the resolution, for future usage. It can be reasoned that this process will attract huge memory requirements. The procedure then creates a list of pronouns for the target tweet and a list of nouns for all other parent tweets. These two lists are further used to resolve the pronouns to their relevant nouns through

the function *pronounResolver*(*p*, *S*, *t*) and further updated on *Sentence*[1] with the help of *replaceDoc*(*.*). The details of the function *pronounResolver*(*p*, *S*, *t*) will be discussed next.

---

**Algorithm 1. Increasing information density of tweets by pronoun resolution**


---

```

Result: Pronoun resolved tweets
1  Sentence [1:  $W_d$ ]  $\leftarrow$  NULL;
2  Insert d into the current window after fetching its parent;
3  Sentence[1]  $\leftarrow$  preprocess(d)
4  Listpn  $\leftarrow$  pronounList(Sentence[1]) ;
5  Listn  $\leftarrow$  nounList(Sentence[1]) ;
6  for  $i \leftarrow 2 : W_d$  do
7      Parent  $\leftarrow$  fetchParent(d) ;
8      if Parent is NULL then
9          Break;
10     end
11     Sentence[i]  $\leftarrow$  preprocess(Parent) ;
12     Listn  $\leftarrow$  Listn  $\cup$  nounList(Parent) ;
13     d  $\leftarrow$  Parent;
14 end
15 ResolvedDict  $\leftarrow$  NULL;
16 for each p in Listpn do
17      $S \leftarrow$  Dictionary( $k;v$ ) where  $k \in$  Listn and
18      $v \leftarrow 0$ ;
19      $R \leftarrow$  pronoun Resolver( $p;S;t$ ) ;
20     ResolvedDict  $\leftarrow$  ResolvedDict  $\cup$  { $p;R$ } ;
21 end
22 Sentence[1]  $\leftarrow$  replaceDoc(Sentence[1]; ResolvedDict);

```

---

### 3.1.2. Antecedent scoring system

The primary aim of the proposed approach can be termed as an antecedent scoring system that tries to search for a suitable noun phrase for a given pronoun that appeared in the candidate reply tweet. Though there are few pronoun resolution systems, such as [23], [33]–[35], that try to resolve the same through the contextual properties, and even fewer articles that use platform-specific structural properties such as [36], to the best of our knowledge there is no proposal in literature till date that try to use both structural and contextual properties for pronoun resolution for streaming data from social media. Therefore, in our proposal, we try to employ the structural properties that a social media platform provides, along with contextual properties, similar to [23], of the related sentences to achieve our aim. The Algorithm 2 explains the procedure for the method *pronounResolver*(*p*, *S*, *t*), where *p*, *S*, and *t* represent the pronoun to be resolved, set of all the candidate noun phrases for *p*, and the tweet where the resolution is to be performed, respectively. This algorithm first employs the structural properties of Twitter to perform possible pronoun resolution as can be seen until line number 17. Twitter provides much structural information that is used in this proposed algorithm, such as, author name of the tweet (*authorName*(*t*)), parent tweet id in case the document in question is a reply (*fetchParent*(*t*)), user mentions using '@' (*usernameSet*(*t*), returns set of all user mentions). Apart from these, the method *remMismatchGenderNumber* (*S*,*p*) removes the noun phrases in *S* that do not apply with gender and number agreement with *p*, and *tweetFormat*(*t*, {'it', verb}), returns *FALSE* if the format {'it', verb} is not present in the tweet *t*.

Finally, in case the structural resolution fails to find a match for *p*, the algorithm proceeds further to contextual properties to perform the same through the method *maxSimilarity*(*p*, *S*). Empirically identified subjective antecedents are used in this method to assign scores to the candidate nouns to identify possible noun phrases. These assigned scores are related to salience, structural matches and referential distance, and preference of terms. As previously mentioned, the structure of each noun phrase in *S* is {*k*, *v*} which indicates the noun phrase and its total score, respectively. The scores as mentioned next are added to the *v* for each *k* for a given *p*. The factors are assigned as follows. A 'definite noun phrase', replaced with 'demonstrative' or 'possessive' pronoun or by 'definite noun phrase', is assigned the score 0. The candidates, representing 'given information' or 'theme', are considered as antecedents, and a score of 1 is assigned. A score of 1 is assigned to the candidate if any specific list of verbs is satisfied. The noun phrases with reiterations are a good choice for antecedents, the score is assigned depending on the number of repetitions within the same text window. A score of 2 is assigned if the occurrence repeats minimum twice within the same window, if repeats once assign 1 else 0. Phrases that are preceded and succeeded by a prepositional phrase are penalized as -1, non-prepositional phrases are preferred, and secure 1. If the candidate carries alike collocation pattern with pronoun are preferred and secures 2, others score 0.

**Algorithm 2. PronounResolver (p, S, t)**


---

```

Result: Noun Phrase in S to which pronoun p resolves to.
1   $S' \leftarrow \text{remMismatchGenderNumber}(S; p);$ 
2   $U \leftarrow \text{unameSet}(t);$ 
3  if  $p \in \{i; me; my; myself; mine\}$  then
4      return  $\text{authorName}(t);$ 
5  else if  $p \in \{we; our; ourself; ourselves; ours; us\}$  then
6       $U \leftarrow U \cup \text{authorName}(t) \cup \text{authorName}(\text{fetchParent}(t));$ 
7      return  $U;$ 
8  else if  $p \in \{it\}$  &&  $\text{tweetFormat}(t; \{'it'; \text{verb}\}) == \text{TRUE}$  then
9      return  $\text{NULL};$ 
10 else if  $p \in \{you; yours; yourself; yourselves\}$  then
11     if  $\text{sizeof}(U) > 0$  then
12         return  $U;$ 
13     else if  $\text{fetchParent}(t) \neq \text{NULL}$  then
14         return  $\text{authorName}(\text{fetchParent}(t));$ 
15     else
16         return  $\text{NULL};$ 
17 end
18 else
19      $(k; v) \leftarrow \text{maxSimilarity}(p; S);$ 
20     return  $k$ 
21 end

```

---

If the noun phrases (N.P.), pronoun phrase (PNP) and verb (V1, V2, and V3) follow the structure of sentence "... (PNP) V1 N.P... conj (PNP) V2 it (conj (PNP) V3 it)", where, 'and', 'or', 'before', and 'after' are some examples of possible conjunctions. The N.P. succeeding V1 is quite a similar candidate for the antecedent of PNP 'it' of preceding V2, and therefore selected as the preferred term and awarded 2, else 0 is assigned. A referential distance is estimated for the candidate in the following order; assign a score of 2 if it belongs to the same sentence/line of the pronoun if it belongs to the preceded line then a score of 1, and a score of 0 if it resides in the pre-preceded line. N.P.s constitute terms that are better candidates as an antecedent, and thereby receives an additional score of 1.

A possible example of the algorithms presented here is as follows. Let us assume  $t$  as "*a month ago, you and I co-led a letter with over 100 colleagues of mine and sent this*" to another 45.", posted by the user 'ct\_turnip' in reply to a tweet posted by 'honor\_man' that states "*Together we will fight #DefendDACA*". Here, the list of pronouns to be resolved is {'you', 'I', 'mine', 'this'}. The first pronoun in the list, 'you', and can be resolved to the parent tweet's author, 'honor\_man',  $U$  is empty. The pronouns 'I' and 'mine' are both first-person singular pronouns, therefore, must be resolved with the author's name 'ct\_turnip'. To resolve pronoun 'this', a possible list of nouns is prepared as {'honor\_man', 'ct\_turnip', 'DefendDACA', 'month', 'letter', 'colleague'}. The initial score is set to 0 for all the candidates. The method,  $\text{maxSimilarity}(p, S)$ , the assigns scores to each of the noun phrases as discussed previously, and they are {'hono\_man': 0, 'ct\_turnip': 2, 'DefendDACA': 3, 'month': 2, 'letter': 3, 'colleague': 1}. We can observe that both 'DefendDACA' and 'letter' same score. Then, as antecedent the immediate reference candidate is considerable. Therefore, the pronoun 'this', is replaced by 'letter', and the final resolved tweet is "*a month ago, honor\_man and ct\_turnip co-led a letter with over 100 colleagues of ct\_turnip and sent letter to another 45*". It is to be noted that we used a certain pronoun resolution method in our proposed algorithm. There exist several similar methods in the literature that may improve the performance of the desired outcome. However, these algorithms are not explored here and are considered as the future scope of our work.

## 4. RESULTS AND DISCUSSION

In this section, we present our experimental results, analyses, and discussions. However, before that, we present the existing OED algorithms that are used here to perform our experiments, information on the dataset used, and the experimental setup of the evaluation systems.

### 4.1. Online event detection algorithm

For the evaluation purpose, we consider two major document pivot OED approaches in this area, proposed by Becker *et al.* [4] and Hasan *et al.* [5], as the document pivot techniques are focused on the clustering of documents by satisfying certain similarities between the documents. According to Becker *et al.* [4], an event, denoted by  $e$ , is a real-world occurrence and is associated with a period ( $t_e$ ), and a time-ordered streamed Twitter message set ( $M_e$ ) where all the messages in this set are posted in the period  $t_e$ . To fulfill the goal of event clustering through tweet streams in real-time, the approach uses an incremental online clustering algorithm. During this phase, a threshold value for similarity score is set empirically, where the

similarity of the incoming tweets and the existing clusters are estimated through cosine similarity. The incoming tweet will be inserted to an existing cluster that has the highest similarity score if the similarity value is within the previously set threshold, otherwise, a new cluster is formed. Each of these clusters is a potential candidate portraying possible new events. The authors, in this article, further proposed an SVM-based classification technique to classify these events into real-world events and non-events, which is out of scope for our article.

Hasan *et al.* [5] proposed a new approach TwitterNews+ [5]. It is an incremental clustering algorithm that has comparatively less computational complexity as it throws away the old clusters after satisfying a threshold of the period to occupy the new clusters. TwitterNews+ consists of two main components: 'search module' and the 'event cluster module'. From the set of latest tweets maintained by this system, the Search Module provides fast retrieval of similar tweets and a binary decision on the uniqueness of an input tweet. A tweet from the search module, tagged as 'not unique', is handed over to the event cluster module that searches for a candidate cluster where tweets can be assigned. This module has a de-fragmentation sub-module to merge small fragmented clusters. To get newsworthy events from the candidate event clusters, this system uses a set of different filters and word-level longest common subsequence. A notable pre-processing task is performed by TwitterNews+ that involves the removal of spam phrase tweets (such as 'free access', 'click here').

#### 4.2. Experimental set-up

For the experiments, an open-source dataset from the George Washington University website [37] is used hereafter hydration. The dataset consists of tweets over the period of August 2013 to January 2019. As we consider English tweets, all the non-English tweets are filtered out before use. However, due to incomplete information in the old tweets, we remove all the documents before 2017 to maintain consistency among the data points considered. For evaluation of our work, manually annotated ground truth for clusters is prepared. Given that the proposed work targets streaming social media, the proposed experimental set-up imitates this property by treating the database as time-series data through the timestamp associated with each hydrated tweet, and by not considering any document having a higher timestamp value than the document where the resolution is being performed.

The most commonly used evaluation metrics, namely, *precision*, *recall*, and *F-Measure*, are used here to evaluate the efficiency of implemented work. The evaluation is done by considering the common cluster membership of object pairs in clustering. This common cluster membership is used to calculate recall and precision. To run the experiments, we run both the existing procedures as explained in section 4.1 separately. For each of them, we first run the procedure with the previously mentioned dataset, then the same algorithm is run again with the dataset while resolving the pronouns for each document first through the algorithms as explained in the previous section. For evaluation for these outcomes, each of them is compared with the previously formed ground truth. Finally, the proposed metrics are measured based on the true-positives ( $t_p$ ), false-positives ( $f_p$ ), true-negatives ( $t_n$ ), and false-negatives ( $f_n$ ). Apart from these three parameters, as we intend to demonstrate that the number of documents properly clustered should increase, we propose a new comparison metric, the average number of true positives for event clusters, which will be able to demonstrate the average increase or decrease in document density of the event clusters.

#### 4.3. Evaluation and result analysis

We begin the evaluation by examining the procedure as proposed by [4]. on the given dataset along with (named base data), and with our proposed pronoun resolution (named P.R. data) algorithm. Similarly, we also examine the TwitterNews+ algorithm with these two corpora and present the results through the parameters discussed above. The results for the precision, recall, and F-measure parameters are shown in the Table 1 and Table 2 respectively for the mentioned algorithms.

From Table 1 and Table 2 we can see that the approaches proposed in both the articles achieve better precision values with the pronoun resolved corpus. Where we observed a 2.79% increase of precision for the approach by [4], a 3.52% hike of precision can be seen for TwitterNews+ as proposed by [5]. The primary reason behind this increase in these values is related to their choice of weight metric selection. Both the papers, though they have different clustering techniques, have used a term weighing metric, *tf-idf*, where the weights are assigned based on the common terms in two documents. Considering the fact that many of the tweets in a corpus are replies to earlier tweets, they at many times are not clustered together for the lack of many common terms. This happens as the writer of the reply tweet may choose to replace words with pronouns that are already mentioned in the previous tweet. Due to the uses of pronouns in place of the nouns, the similarity of these two documents may reduce greatly, thereby affecting the clustering process. By resolving the pronouns and replacing them in place through the proposed method, the performance of the existing methods is improved as the weighing metrics perform better with the pronoun resolved data.



For the same reasons as discussed above, a similar trend can also be seen in the Recall and F-measure values for both the algorithms. One important point to be noted here is that, while preparing the ground truth of this corpus, if a tweet is marked as a part of a certain event, all its reply tweets are also categorized as part of the same event. However, it has been observed that most of the reply tweets are not necessarily contributing to the discussion and do not contain either the important words or pronouns pointing to those words. This leads to a huge decrease in the recall value of the clustering methods as the term weighing metrics is solely dependent on the re-usability of common words.

Table 1. Result of Becker *et al.* [4] approach

Test dataset	Precision	Recall	F-measure
Base data	0.8113	0.2975	0.4354
PR data	0.8452	0.3264	0.4709

Table 2. Result of Hasan *et al.* [5] approach

Test dataset	Precision	Recall	F-measure
Base data	0.7140	0.3179	0.4399
PR data	0.7516	0.3500	0.4776

The results shown in Figure 1 focuses on the primary aim of our work, which is to increase the information density in the clusters. We can see from the Figure 1, for both the protocols, the number of true positives has increased manyfold on average for the identified event clusters. This is due to the same reason we have iterated before, that is, as many of the pronouns are resolved to their suitable noun phrases, the term weighing metrics can cluster the reply tweets more efficiently by assigning a higher score to them. In the results shown above, we considered a fixed value for  $W_d$  for simulation. In future, we will consider the effects of different values for  $W_d$ , and explore possibilities of developing adaptive  $W_d$  window size for this proposed work.

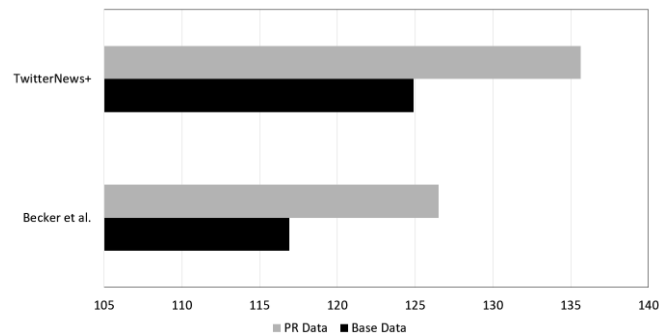


Figure 1. Increase in the average number of T. P. for event clusters

## 5. CONCLUSIONS AND FUTURE WORKS

In this paper, we presented a pronoun resolution algorithm for streaming data from social media platforms that can be used before applying various OED algorithms. Our results show that, by applying our algorithm before OED procedures, these procedures incur improved efficiency in clustering the relevant documents. Moreover, to resolve the pronouns, our method needs to fetch only a fixed number of previous documents in the reply chain of posts, rather than fetching all the previous documents. Though we have presented our results using Twitter data, it can be used in data from any social media platforms that have the reply feature. The selection of the size of the document window remains a future work for this work. Another area to focus on in the future is to explore other pronoun resolution algorithms that can further enhance the performance of this work.




## REFERENCES

- [1] D. T. Nguyen and J. E. Jung, "Real-time event detection for online behavioral analysis of big social data," *Future Generation Computer Systems*, vol. 66, pp. 137–145, 2017, doi: 10.1016/j.future.2016.04.012.
- [2] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, 1998, pp. 28–36, doi: 10.1145/290941.290953.
- [3] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: final report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 194–218, doi: 10.1184/R1/6626252.v1.
- [4] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: real-world event identification on twitter," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 438–441.




- [5] M. Hasan, M. A. Orgun, and R. Schwiter, "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework," *Information Processing & Management*, vol. 56, no. 3, pp. 1146–1165, May 2019, doi: 10.1016/j.ipm.2018.03.001.
- [6] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: news in tweets," in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 2009, pp. 42–51, doi: 10.1145/1653771.1653781.
- [7] M. Cordeiro and J. Gama, "Online social networks event detection: a survey," in *Solving Large Scale Learning Tasks. Challenges and Algorithms*, vol. 9580, S. Michaelis, N. Piatkowski, and M. Stolpe, Eds. Cham, Switzerland: Springer Nature, 2016, pp. 1–41.
- [8] M. Samanta, Y. K. Meena, A. P. Mazumdar, and M. C. Govil, "EDOF: an open framework for event detection systems," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Jul. 2017, pp. 1–7, doi: 10.1109/ICCCNT.2017.8204155.
- [9] P. P. Dakle, T. Desai, and D. Moldovan, "A study on entity resolution for email conversations," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 65–73.
- [10] K. Wright-Bettner, M. Palmer, G. Savova, P. de Groen, and T. Miller, "Cross-document coreference: an approach to capturing coreference without context," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 2019, pp. 1–10, doi: 10.18653/v1/D19-6201.
- [11] N. Skachkova and I. Kruijff-Korbayova, "Reference in team communication for robot-assisted disaster response: an initial analysis," in *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, 2020, no. 122–132.
- [12] J. Allan, *Topic detection and tracking*, vol. 12. Boston, MA: Springer US, 2002.
- [13] N. S. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: automated trend discovery for Weblogs," in *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004, vol. 2004.
- [14] X. Wan, E. Milios, N. Kalyaniwalla, and J. Janssen, "Link-based event detection in email communication networks," in *Proceedings of the ACM Symposium on Applied Computing*, 2009, pp. 1506–1510, doi: 10.1145/1529282.1529618.
- [15] N. D. Doulamis, A. D. Doulamis, P. Kokkinos, and E. M. Varvarigos, "Event detection in Twitter microblogging," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 2810–2824, Dec. 2016, doi: 10.1109/TCYB.2015.2489841.
- [16] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, Apr. 2012, pp. 624–635, doi: 10.1137/1.9781611972825.54.
- [17] H.-J. Choi and C. H. Park, "Emerging topic detection in twitter stream based on high utility pattern mining," *Expert Systems with Applications*, vol. 115, pp. 27–36, Jan. 2019, doi: 10.1016/j.eswa.2018.07.051.
- [18] A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach," *Social Network Analysis and Mining*, vol. 5, no. 1, p. 18, Dec. 2015, doi: 10.1007/s13278-015-0258-0.
- [19] D. Zhou, L. Chen, and Y. He, "An unsupervised framework of exploring events on twitter: filtering, extraction and categorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, vol. 29, no. 1, pp. 2468–2475.
- [20] V. Ng, "Supervised noun phrase coreference research: the first fifteen years," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 1396–1411.
- [21] J. Lu and V. Ng, "Event coreference resolution: a survey of two decades of research," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Jul. 2018, pp. 5479–5486, doi: 10.24963/ijcai.2018/773.
- [22] C. Kennedy and B. Boguraev, "Anaphora for everyone: pronominal anaphora resolution without a parser," in *Proceedings of the 16th conference on Computational linguistics -*, 1996, vol. 1, pp. 113–118, doi: 10.3115/992628.992651.
- [23] R. Mitkov, "Robust pronoun resolution with limited knowledge," in *Proceedings of the 36th annual meeting on Association for Computational Linguistics -*, 1998, vol. 2, pp. 869–875, doi: 10.3115/980691.980712.
- [24] P. Jain, M. R. Mital, S. Kumar, A. Mukerjee, and A. M. Raina, "Anaphora resolution in multi-person dialogues," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 2004, pp. 47–50.
- [25] A. J. Stent and S. Bangalore, "Interaction between dialog structure and coreference resolution," in *2010 IEEE Spoken Language Technology Workshop*, Dec. 2010, pp. 342–347, doi: 10.1109/SLT.2010.5700875.
- [26] M. Poesio, A. Patel, and B. Di Eugenio, "Discourse structure and anaphora in tutorial dialogues: an empirical analysis of two theories of the global focus," *Research on Language and Computation*, vol. 4, no. 2–3, pp. 229–257, Oct. 2006, doi: 10.1007/s11168-006-9005-z.
- [27] T. Winograd, "Understanding natural language," *Cognitive Psychology*, vol. 3, no. 1, pp. 1–191, Jan. 1972, doi: 10.1016/0010-0285(72)90002-3.
- [28] J. R. Hobbs, "Pronoun resolution," *ACM SIGART Bulletin*, no. 61, p. 28, Feb. 1977, doi: 10.1145/1045283.1045292.
- [29] J. G. Carbonell and R. D. Brown, "Anaphora resolution," in *Proceedings of the 12th conference on Computational linguistics -*, 1988, vol. 1, pp. 96–101, doi: 10.3115/991635.991656.
- [30] S. Lappin and H. J. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, pp. 535–561, 1994.
- [31] D. Connolly, J. D. Burger, and D. S. Day, "A machine learning approach to anaphoric reference," in *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, *ACL*, 1994, pp. 255–261.
- [32] M. Strube and C. Müller, "A machine learning approach to pronoun resolution in spoken dialogue," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, 2003, vol. 1, pp. 168–175, doi: 10.3115/1075096.1075118.
- [33] N. B. Niraula and V. Rus, "A machine learning approach to pronominal anaphora resolution in dialogue based intelligent tutoring systems," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Cham, Switzerland: Springer Nature, 2014, pp. 307–318.
- [34] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from Twitter," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1104–1112, doi: 10.1145/2339530.2339704.
- [35] C. Chen and V. Ng, "SinoCorefencer: an end-to-end Chinese event coreference resolver," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 4532–4538.
- [36] B. Aktaş, T. Scheffler, and M. Stede, "Anaphora resolution for Twitter conversations: an exploratory study," in *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, 2018, pp. 1–10, doi: 10.18653/v1/W18-0701.
- [37] J. Littman, "115th U.S. Congress Tweet Ids," *CORE*, 2017. <https://core.ac.uk/display/268509313> (accessed Mar. 06, 2021).

## BIOGRAPHIES OF AUTHORS






**Manisha Samanta**    is currently pursuing Ph.D. degree in Computer Science & Engineering from Malaviya National Institute of Technology, Jaipur, India. Her research interests are data mining, natural language processing, knowledge management, deep learning, software engineering. She can be contacted at email: 2015rcp9524@mnit.ac.in.






**Dr. Yogesh Kumar Meena**    is presently associated with Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur as an Associate Professor. He received his Ph.D. degree from Malaviya National Institute of Technology, Jaipur, India. He is having more than 12 years of academic experience. His research interests include issues related to data mining, natural language processing, pattern recognition, and knowledge management. He is the author of a great deal of research studies published at national and international journals, conference proceedings as well as book chapters. He can be contacted at email: ymeena.cse@mnit.ac.in.






**Dr. Arka Prokash Mazumdar**    is currently working as an Assistant Professor of Computer Science and Engineering, at Malaviya National Institute of Technology Jaipur, India. He has completed his Ph.D. from Indian Institute of Technology Patna, India. His Ph.D. research involved opportunistic communication and energy efficiency in wireless communication. Presently, he is carrying out his research in the areas of internet of things, social internet of things, information centric networks, and wireless communication strategies. He can be contacted at email: apmazumdar.cse@mnit.ac.in.



**Dr. Girdhari Singh**    is presently associated with Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur as a Professor. He obtained his Ph.D. from Malaviya National Institute of Technology Jaipur. He is having more than 20 years of academic experience. His research interests are software engineering and intelligent systems. He can be contacted at email: gsingh.cse@mnit.ac.in.



**Dr. Dinesh Gopalani**    is currently associated with the Department of Computer Science and Engineering, Malaviya National Institute of Technology Jaipur, India as an Associate Professor. He received his Ph.D. degree from Malaviya National Institute of Technology, Jaipur, India. He is having more than 20 years of academic experience. His research interests include issues related to aspect-oriented programming, compiler design, natural language processing, and knowledge management. He has published research papers at national and international journals, conference proceedings as well as chapters of books. He can be contacted at email: dgopalani.cse@mnit.ac.in.