2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Characteristics Analysis of Moving Conversations to Detect Events on Twitter

Hansi Senaratne
Earth Observation Center
German Aerospace Center
Oberpfaffenhofen, Germany
hansi.senaratne@dlr.de

Dominic Lehle
Advanced Research & Technologies
Avira
Tettnang, Germany
derlehle@gmail.com

Tobias Schreck
Inst. of Computer Graphics & Knowledge Visualization
Graz University of Technology
Graz, Austria
tobias.schreck@cgv.tugraz.at

***Abstract*—A conversation is an exchange of thoughts, news, or ideas about a particular topic between two or more people. On Twitter, hashtags allow its users to collate all conversations pertaining to a particular topic. The progressions that occur in such conversations through the geographic space, the time, or the thematic contexts, create *trajectories of conversations* on Twitter, and they can give us valuable insights into interesting events that take place around us. In this paper we develop an approach based on data analysis and visualisation, to (1) construct such conversation trajectories for chosen popular hashtags, (2) analyse the various geospatial- and content- characteristics of the conversation trajectories (e.g., distance variance, speed of propagation, topic diversity, or credibility) to determine co-located events, and (3) rank and sort the resulting conversation trajectories according to a user-defined interestingess-measure, to narrow down the search space for *interesting* conversation trajectories. Our approach is among the first to introduce the usage of movement of conversations across geographic space and time for the exploratory detection and analysis of events, whereas most existing works use keyword-based text analysis to detect events on Twitter. All the three stages of the approach (construct, analyse, rank & sort) are presented in a visual-interactive interface that allows us to explore Twitter text data without extensive prior knowledge, and benefit from the pure exploratory capabilities of the tool. The usefulness of our approach is demonstrated as a proof-of-concept to detect sports-related events, where we were able to identify the outcome of a contest for Major League Baseball sportsmen on Twitter.**

***Index Terms*—Visual knowledge discovery, time varying data, trajectory characterisation & ranking, event detection**

I. Introduction

With over 500 Million Tweets generated every day, Twitter data have become a useful source to determine the *what*, *where*, and *when* of events happening around us. This is mostly done using keywords or hashtags to filter the live stream or static datasets to gather data relevant to an event, and by leveraging their temporal and spatial references. Due to the large amounts of inherent noise in Twitter data, or the unconventional nature of the language used in Tweets, such keyword-based approaches to filter the data also return data that are highly irrelevant to the event being investigated. We propose in this paper an exploratory approach that detects events based on the *movement trajectory of conversations* on Twitter and their geospatial and content characteristics. A spatial trajectory is defined by [1], as a trace generated by a moving object in geographical space, usually represented by a series of chronologically ordered points $(p_1,p_2,...p_n)$, where each point consists of a geographical coordinate set and a time stamp, such as $p = (x, y, t)$. Based on this, in our previous work [2], we developed an approach to construct trajectories for moving conversations on Twitter, by clustering Tweets to identify the consecutive geographical coordinate sets of cluster centers, and averaging the time in each cluster to identify the chronology of the trajectory. As a conversation moves through these space-time modalities, it also changes in terms of for e.g., the diversity of the topics discussed, the direction of spatial movement (linear, back and forth, cyclic etc.), the speed with which the conversation propagates, and also the sentiment polarity, certainty and the credibility of information. These characteristics mirror the different inherent characteristics of events being discussed around us- for e.g., sports- or politics- related events would have a high diversity of topics, high polarity of sentiments, or high speed of propagation of Tweets. Through detecting these characteristics, events that are of interest can be detected.

The contributions of this paper are therefore as follows: based on the approach by [2] we construct the conversation trajectories, and (1) we derive their geospatial characteristics: *distance variance*, *the geometric linearity* of the trajectories, and *the speed variance* of conversation propagation. These geospatial characteristics are described in detail in Section III-A. (2) we derive content characteristics of the conversation trajectories: *topic diversity*, *sentiment linearity*, *certainty* and *credibility variance*. These content characteristics are described in detail in Section III-B. Characterisation of these conversation trajectories based on their geospatial and content structures is useful to describe them and identify interesting events. These characterisations are used as our grouping strategy for filtering out meaningful and interesting conversation trajectories. As a third contribution (3), we introduce a method to *rank and sort* the conversation trajectories based on a user-defined *interestingness measure*. This is presented in Section III-C. These conversation trajectory extraction, characterisation, and ranking / sorting are further presented as a proof-of-concept for detecting sports-related events in a visual-interactive environment.

The remainder of this paper is organised as follows. In Section II we review the related works for trajectory analysis using Twitter data. The trajectory characterisation and feature-based ranking methodology for identified trajectories are described in Section III. In Section IV we demonstrate the usefulness of the developed methods within a proof-of-concept. We conclude our findings, limitations, and future perspectives in Sections V and VI.

II. Related Work

Movement detection and the observation of their trajectories are important for gathering insights about human / animal

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

behaviour, or other physical entities. Their various characteristics such as direction of movement, speed, similarity etc., can be used in applications such as city planning, urban dynamics, transportation, or logistics, as can be seen in the works of [3], [4], and [5]. Movement detection, specifically using data sources such as Twitter is not trivial. Some of the main reasons for this are that the data is unstructured, uses unconventional language, lacks orthography, and are mostly ambiguous.

However, many works in the state-of-the-art overcome these challenges through various approaches. [6] constructed trajectories of Twitter users from Tweet locations by computing the trajectory medoid (i.e., the cluster point of a dataset whose average dissimilarity to all objects in the cluster is minimal) for each spatially referenced Tweet. In another work [7] introduced a method for spatial generalisation and aggregation of movement trajectories by extracting only the significant points in a trajectory, that also retains the essential characteristics of the movement. Through parameterisation of the movement model they allow enough leeway to the user to control the extent of abstraction. They further introduce quality metrics for assessing the quality of the generalisation. [8] have been working on large complex time-dependent data, introducing time-dependent movement analysis features particularly for group movement, and methods to automatically analyse and filter interesting sub-parts of a dataset for in-depth inspections.

[9] analyse irregular, anomalous movement patterns of humans during crisis situations within a visual analytics environment. They extract the Tweeter locations from Twitter meta data and organise them in chronological order to define the Tweeters' movement trajectories. The incompleteness of the data are supplemented by using a shortest-path algorithm that fills the locations in-between recorded locations. The line segments of the extracted trajectories are further clustered using a modified partition-based clustering model, and the Hausdorff distance function to find common patterns of movements during a crisis situation. By referring to historical movement data, the abnormality of current movements is determined. With previous knowledge of the crisis events the user can navigate her/himself within the visual-interface by selecting the area and the time-frame of the events for analyses. [10] use Tweets to extract movement trajectories of Tweeters by also following a chronological order of Tweeter locations. They further develop a method to observe the activities conducted by moving Tweeters within the course of their movement trajectory, by looking into the types of Points of Interest along the trajectory where the users stopped at. The authors suggest that the semantics of the trajectories could be extracted by following their approach. [11] present a movement trajectory mining and analysis approach called the TravelDiff. The authors follow an approach similar to [9] and [10] to extract the movement trajectories of Tweeters using the Twitter user locations. In a subsequent step they use these trajectory patterns to infer about the seasonal changes and events that occur in the regions considered within their use-cases. Using movement trajectories of Tweeters as a baseline, [12] derive demographic characteristics of Twitter users in Chicago. By analysing the surname, race/ethnicity, age, gender, and the "home" and "activity center" locations that Tweeters have visited, the authors are able to classify the Tweeters based on their spatio-temporal distribution across Chicago, and further analyse their mobility characteristics. In contrast to the works above which use the geotag of Tweeter locations to extract their movement trajectories, [13] use the mentioned locations in the Tweets. The authors first run an algorithm that detects place names based on a Bayes model that has been learned by going over the locations mentioned in the Wikipedia page for China. In a subsequent step, the locations are further verified based on the locations mentioned in the friends' profiles within their social network. These locations along with their semantics are taken into consideration for the movement trajectory extraction. [14] further show through their preliminary works how semantics such as stay points and activity types can be extracted through a fusion of Twitter-based trajectory data and Open Street Map data.

Our approach makes a first attempt to detect the movement trajectories of *conversations* on Twitter, and use these trajectories to explore co-located events within a visual analytics environment. We rely on a grouping strategy to query, and thereby filter the Twitter dataset based on the *geospatial and content structure* of Tweets, rather than a single keyword-based approach as noted in the many related works. These structural analyses help us to detect changes in evolving conversations through the spatial, temporal, and thematic modalities. The key advantage of our approach is that it caters to wider analysis and exploratory possibilities that don't necessarily require us to possess prior knowledge of the events. The details of our approach are discussed in the following section.

III. Detection, Characterisation, and Ranking of Conversation Trajectories

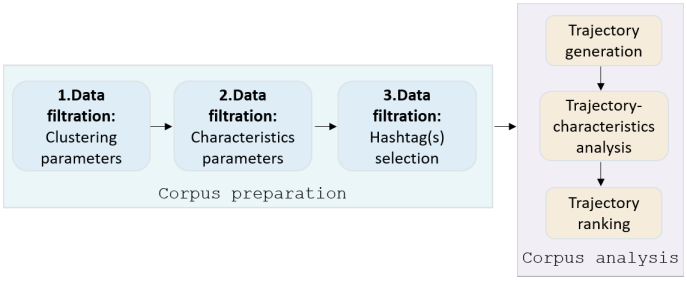

Fig. 1. The work-flow of the extraction and analyses of moving conversations.

Our grouping strategy for Tweets was to collate all Tweets pertaining to specific hashtags based on certain criteria. This allowed us to collect conversations of news, thoughts, and ideas about a trending topic that is circulated by the use of hashtags on Twitter. Therefore, the grouping strategy of the available data was implemented in three stages. As can be seen in Fig. 1, these are (1) identify the clustering parameters, (2) identify the characteristics parameters, and (3) select hashtags from the resulting ordered list to generate the conversation movement trajectories. To group similar hashtags based on their relative distribution in space and time, we performed a density-based sequential DBSCAN clustering [15] on each hashtag. The key advantages of using DBSCAN are one, it does not require to specify the number of clusters, and two, it can find non-linearly separable clusters. Due to the episodic nature of conversations, we used the extended *getNeighbors* function of the DBSCAN algorithm to add a maximum temporal distance (*tf*). This helped us to cluster together

TABLE I
SUMMARY STATISTICS OF THE TWITTER DATA.

Total no. of Tweets	60,000,000
Total no. of geotagged Tweets	8,607,490
Total no. of Tweets discarded	51,392,510

hashtags that are closer in time as well as in space. Further, we set a minimum number of Tweets (*MinTws*) as the minimum points in the cluster and a geographic radius as the distance threshold *eps* for the parameter settings of DBSCAN. Furthermore, as also described in [16] this sequential DBSCAN algorithm places unique cluster assignments by assigning border points (non-core points, but density reachable) to the first cluster they are reachable from, thereby making the results deterministic, and desirable for our approach. The average time in each of these hashtag clusters are connected sequentially to derive the episodic conversation movement trajectory (more on this can be found in [2]).

For a meaningful analysis of these conversation movement trajectories, we have identified several characteristics based on their geospatial structure and the content structure. Characterisation of conversation trajectories is helpful to filter out interesting and meaningful patterns, as well as to distinguish them from noise in the data. To reduce the search-space and sort the resulting trajectories, we developed a ranking technique based on an *interestingness measure* according to the task at hand. These characteristic features and the ranking technique are discussed in detail in the following sections. A Twitter dataset for 19 - 20 February, 2014 was used in all of the following sections. This dataset was created by [17], where the Twitter Streaming API with *Gardenhose* access was used to harvest a randomly sampled subset of data that consisted of 10% of the live stream (this level of access is no longer available). In their incoming stream, over 2.5 million Tweets on average per hour were harvested. To maximise the geotagged Tweets in the dataset, [17] merged the 10% stream with five additional geographically filtered streams (one stream per one area of interest in the world), into one stream of Tweets with no duplicates. The summary of data used in this study is in Table I. All geotagged Tweets with the *Tweet location* (where Tweeters geotag their content on the fly) were used in the following analyses, with no further filtrations.

A. Characterisation based on the Geospatial Structure

Characteristics based on the geospatial structure of a trajectory path are important to identify how the episodic clusters are generated, the difference between trajectories, and to detect interesting changes in the trajectory path. These characteristics are detailed below:

1) Distance Variance: Distance variance calculates the distance between two consecutive clusters of a conversation trajectory. This is useful to determine the overall impact/coverage of a given hashtag on Twitter. To calculate the distance variance we calculate the great-circle distance between two cluster centroids using the Haversine formula [18].

$$\begin{aligned} a &= sin^2(\Delta\varphi/2) + cos\varphi_2 * sin^2(\Delta\lambda/2) \\ c &= 2 * arctan2(\sqrt{a}, \sqrt{(1-a)}) \\ d &= R * c \end{aligned} \tag{1}$$

φ and λ are the geographical coordinates- the latitude and the longitude respectively, and R is the radius of the earth. In Fig. 2 and 3 we demonstrate the distance variance for the hashtags "#melfest" which refers to the EuroVision song contest, and "#chinesenewyear". #melfest has a much lower distance variance than #chinesenewyear, as it has a lower global spread (audience is mainly coming from Europe). Whereas the Chinese new year is celebrated around the world, and it has a larger spread with clusters of tweets coming from many corners of the globe.

2) Conversation Trajectory Linearity: Conversation trajectory linearity indicates the directional characteristics of trajectories. We calculate the ratio of turning points of each trajectory segment to calculate the linearity. A segment of a trajectory is assumed to have a turning point when the bearing angle between the subsequent segments is higher than a predefined threshold of 90 degrees. Therefore, the ratio indicates how many times, more than 25%, a trajectory is heading in a different direction. In the examples in Fig. 2 and 3, #chinesenewyear has significantly less turning points, resulting in a more linear trajectory, as compared to #melfest Eurovision song contest trajectory, which has a more cyclic nature. The reason for this could be that the Eurovision contest is a live event very popular in Europe that takes place in the course of 4 hours, in contrast to the Chinese New Year that is celebrated all over the world at different time zones at the dawn of the new year.

3) Speed Variance: The speed variance determines how fast a particular hashtag on Twitter propagates between locations. While some hashtags have a high peak time soon followed by a drop, others propagate over a steady speed at a longer time interval. This characteristic can be used for e.g., to detect the virality of a hashtag on Twitter, and further analyse the content therein. The speed variance is calculated first by averaging the Tweet creation date of all Tweets in each cluster, and then by dividing the distance of each subsequent trajectory by the difference in time for each cluster.

B. Characterisation based on the Content Structure

Analysing the content of Tweets is important for context-aware information gathering. We use sentiment and topic analyses techniques for identifying the different characteristics based on Twitter content. While term-usage analysis is used to find general patterns and topic terms, keyword based analysis of Tweets help to find what people are talking about, where, when, and how often in the clusters. Using sentiments and topic analyses within large amounts of trajectories help us to determine which current conversation trajectories are worth exploring. These characteristics are detailed below:

1) Topic Diversity: The first step of analysing the diversity of topics in the Tweets, in the episodic clusters, is to determine the thematic categories that the Tweets fall into. Many techniques have been used in the state-of-the-art for topic classification, such as Named Entity Extraction (NER), or Latent Derichlet Allocation (LDA), which require to know the number of topics in advance. Due to the unconventional nature of Twitter language (use of abbreviations or slang), we need more multi-modal language features and a specific classifier training to achieve effective topic classifications. The Java library LingPipe [19], which relies on computational linguistics for topic classification, uses a Hidden Markov Model that is trained on complete sentences of over a

Fig. 2. Conversation trajectory visualisation for #melfest, which refers to the Twitter discussions on the Eurovision Song Contest in Sweden.

Fig. 3. Conversation trajectory visualisation for #chinesenewyear.

million words, and has shown to be effective in topic classification (e.g., [20], [21]). Based on these works, we use a hierarchical feature subset selection algorithm to classify the Tweets. The training of this language model is done by categorising character sequences. For each classified topic, conditional and joint probabilities are calculated, and a score is given. We take the topic with the highest score to classify the Tweet. A character based n-gram is used to classify the Tweets, where n is set to the average length of a word in a Tweet. We use an n-gram the size of 5 [22]. During the labelling process, we filtered out URLs, unicode characters, usernames, punctuation etc., and stop words. Accordingly, the Tweets are classified into 12 topics: *computers & technology*, *education*, *family*, *food*, *health*, *marketing*, *music*, *news& media*, *pets*, *politics*, *recreation & sports*, *other*. The topics found in #skilledtrade are mapped to a colour scheme as seen in Fig. 4 and visualised accordingly in the clusters. To

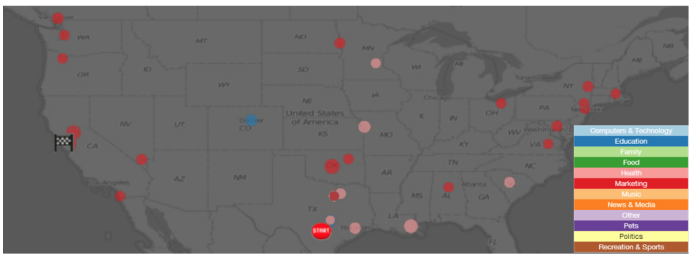

Fig. 4. The clusters of #skilledtrade. The circle radius indicates the number of Tweets in the cluster, and the colour indicates the most frequent topics observed in the cluster.

determine the diversity of topics in the dataset, we calculated the Simpson-Index [23] which assesses the probability of two Tweets from random clusters having the same topic. It is expressed as:

$$\lambda = 1 - \sum_{i=1}^{s} p_i^2 \tag{2}$$

p_i represents the relative amount of the topic i to the sum of all individual topics. This gives us an overall indication of the topic diversity along the conversation trajectory. Fig. 4 shows a low topic diversity for the #skilledtrade. The circles represent the clusters, and the circle radius represents the size of the cluster (Tweet density). The colour of the circles represents the topic category accordingly. Therefore, the topics covered in the clusters are marketing (red), health (pink), and education (blue). Evidently, #skilledtrade is used for job offers in skilled trades such as welders, electricians, machinists etc.

2) Sentiment Linearity: Sentiment analysis allows us to observe the majority attitude and opinion of people regarding a particular topic, brand, product etc., thereby enriching the content analysis process. In our previous trajectory analysis work in [2] we demonstrated how content analysis together with sentiment analysis helped to discover the cancellation of a concert tour in a particular city. In this paper, we look into the sentiment linearity which indicates the *changes* of sentiments in the course of a conversation trajectory. This is especially useful to detect controversial events, where people discussing these events have opposing opinions, surprise, or disbelief [24]. To calculate the sentiment linearity (S_l), we first calculate the number of positive, negative, and neutral Tweets in each episodic cluster. Next, we calculate a sentiment score for the subsequent cluster by using the following measure of contradiction by [25]:

$$S_l = \frac{\theta * \sigma^2}{\theta + (\mu)^2} \tag{3}$$

σ^2 is the variance, and μ is the mean of the sentiments in a given cluster, and θ allows us to add a small value that limits the level of contradictions when the aggregated sentiments is close

TABLE II
(UN)CERTAINTY SIGNAL WORDS AND THEIR CORRESPONDING SCORES.

Signal Words	Score
impossible: unthinkable, unreasonable, cannot, infeasible, unreliable	0.00
unlikely: odd, uneven, diverse, unsure, implausible, improbable	0.25
even chance: believe, estimate, guess, maybe, suppose, think, perhaps, eventual, assume, presume	0.50
likely: possibly, high chance, expected, expect, anticipated, potential, potentially, supposably, belike, presumably, reasonable, probable, plausible	0.75
certain: certainly, sure, safe to say, of course, confident, definitly, certainly, most likely, most probably, assured, reliable	1.00

to zero. Therefore, we set the value for θ at 0.05 (similar to [25]). To indicate how often significant changes occur between subsequent episodic clusters, we calculate the sentiment turns, in addition to the work of [25]. A sentiment turn occurs whenever the change of the sentiment score S_l of the cluster C_n+1 differs from the cluster C_n by more than a user-defined threshold δ (by default this is set to $\delta=0.5$). In Fig. 5 we use a horizon chart to show the sentiment change for #6nations, which was trending for the annual Northern hemisphere rugby union championship in the available dataset. For a hashtag to be trending, it is to be among the most popular topics discussed on Twitter at a given time. One indication for this is the number of Tweets that are mentioning a particular hashtag. In Fig. 5, the green-coloured box on the far left shows a slightly increasing positive sentiment with the beginning of the game (e.g., Tweet: "*...First weekend I have not worked in 2014, just in time for the start of the #6nations*"), and the gradual red colour (green-coloured box on the right) shows negative sentiments from England fans towards the first point for France (e.g., Tweet: "*...31 seconds and France score #WTF #6nations #englandrugby #fail*").

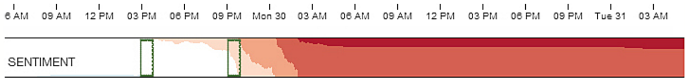

Fig. 5. Sentiment horizon chart for #6nations rugby tournament.

3) Certainty Variance: Words of Estimative Probability (WEP), as coined by Sherman Kent in [26], indicate the certainty in people-to-people dialogues, and was used for military intelligence analysis reports to assess the probability of events occurring. We use this probability estimation of words to *signal* the certainty of conversations on Twitter. As such, we adapt their WEP to indicate an overall certainty score for the conversation trajectories, and to indicate whenever the certainty changes. By following the work of [27], we first classify signal words into five categories (first word in each row in Table II. Several signal words are then assigned under each category with a certainty score (subsequent words in each row in Table II). Tweets that contain any given signal word will be assigned its corresponding certainty score. If a given Tweet does not contain any of the signal words, we handle it as 'certain' and assign a 1.0 certainty score. The resulting average of the certainty scores characterises the clusters, and the variance values indicate the changes of certainty. Fig. 6 shows how the certainty is visualised for selected hashtags in a heatmap visualisation, and how the hashtags are sorted according to the certainty value.

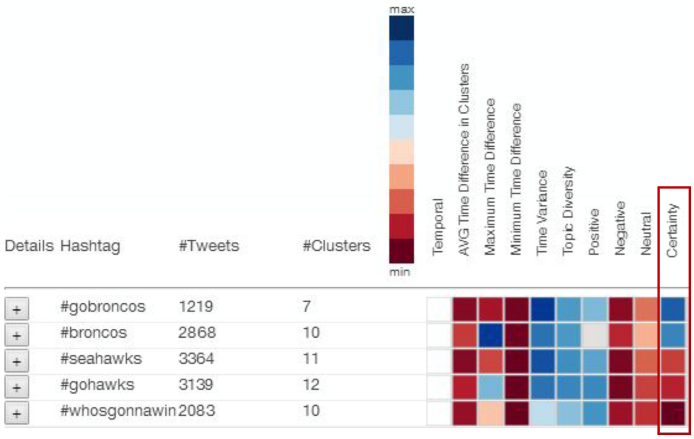

Fig. 6. The *certainty value* used for sorting the Superbowl-related hashtags. #whosgonnawin has the lowest certainty value.

TABLE III
CREDIBILITY FEATURES AND WEIGHTINGS.

Type	Features	Credibility Impact
Message	contains question mark (?)	3.5
	contains exclamation mark (!)	3.5
	contains emoticon smile (:-), ;-), ...)	2.71
	contains emoticon frown (:-(, ;-(, ...)	2.71
	contains URL	4.9
	contains user (@cnnbrk)	3.5
	contains hashtag (#melfast)	3.5
	contains stock symbol ($APPL)	3.5
	is retweet (contains "RT")	5.12
User	registration age (days passed since registration)	5.46
	status count (no. of tweets user has posted)	5.18
	count followers (no. of people following this author)	5.13
	count friends (no. of people author is following)	5.13
	has description (a non-empty "bio" 1)	5.0

4) Credibility: In assessing the credibility of Twitter data, the source of information plays a crucial role. However, this is not straight forward. Due to the non-authoritative nature of Twitter data, the source maybe unavailable, concealed, or missing (this is avoided by gatekeepers in authoritative data). [28] defined credibility as the *believability of a source or message, which comprises primarily two dimensions- the trustworthiness, and expertise*. Expertise contains objective characteristics such as accuracy, authority, competence, or source credentials [29]. Metadata about the origin of Twitter data can provide a foundation for the source credibility of Twitter data [30]. In this paper, we use message- and user- based credibility features derived by [31] in a supervised classification. Using the weightings given for these credibility features in [31], and the credibility impact for features found in the study of [32], we derive the credibility impact on a scale between 0 - 7 for 9 message- based features, and 5 user-based features (Table III), for each Tweet. The average credibility is calculated for each cluster in the different conversation trajectories.

C. Feature-based Ranking of Conversation Trajectories

Ranking of conversation trajectories means to sort the resulting trajectories based on an interestingness measure, that is relevant to the use-case at hand. Our ranking algorithm takes into consideration an interestingness measure based on the

TABLE IV
THE TOP FIVE HASHTAGS AFTER RANKING PARAMETERS.

Rank	Hashtag	No. of Clusters	No. of Tweets
1	#excited	2	886
2	#gobroncos	7	1219
3	#topgear	4	10271
4	#superbowlsunday	9	2741
5	#6nations	11	2274

following characteristics to sort the resulting hashtag-based conversation trajectories. *The topic diversity* as it aids to observe trajectories, *the sentiment variance* that gives us insights to the polarisation in discussions, *high structural linearity* to indicate movement, and *speed variance* to determine the virality of the topics. These characteristics are considered as individual feature dimensions in the following interestingness measure calculations.

In a first step we define the following method to calculate the distance for conversation trajectories.

$$D_n = (1 + \frac{((D_n - min(D_n)) * (100-1))}{(max(D_n)) - min(D_n))}) \quad (4)$$

Each feature dimension D_n is re-scaled to a value between 0 and 100 to make sure that there are no dominant dimensions. Then we define an interestingness value I_n for each of the dimensions between 0-100. For each conversation trajectory, the distance is then calculated by the Euclidean distance of each feature dimension to its interestingness value. Then the average difference is calculated to get the overall proximity according to the user defined interestingness values. The resulting formula, where D is defined as the number of feature dimensions is shown below.

$$d = \sqrt{\frac{\sum_{n=1}^{D} (D_n - I_n)^2}{D}} \quad (5)$$

We calculate the distance considering the following feature dimensions: topic diversity, sentiment variance, structure linearity, and speed variance. However, they can be extended to other trajectory characteristics.

Table IV is an example of conversation trajectory ranking, where our requirements were to see trajectories of content that reflect a more frequent discussion, with diverging topics. For this we set our interestingness measures as follows: low speed variance (set to 10), high topic diversity (set to 80), low structure linearity and high sentiment variance (set to 90).

In Table IV with #excited being an exception (very generic hashtag represented only in two clusters), the top ranked four conversation trajectories represent sports related discussions. A proof-of-concept of the developed methods is presented in the next section.

IV. PROOF-OF-CONCEPT: CHARACTERISATION AND RANKING OF CONVERSATION TRAJECTORIES TO DETECT SPORT-RELATED EVENTS

For this proof-of-concept we used the same dataset described in Section III and Table I. Our aim was to explore sports-related events by characterising and ranking the conversation trajectories found in this dataset. For the first data filtration stage, the DBSCAN clustering parameters are defined. For this we assume that sports-related discussions mostly originate from more populated cities, and we're interested in clusters that have a reasonable amount of data points. Therefore, we set the radius to 15km (eps = 15km) and the neighborhood minimum to 50 Tweets (MinTws = 50). As many sport events have a length of over one hour, we begin with a maximum time frame of 45 minutes (tf = 45min). As a result, we found 163 hashtag-based conversation trajectories within the dataset. To rank and sort these trajectories, we specify the characteristics parameters: as sports-related events often have polarising opinions, we set the *topic diversity* and *sentiment variance* to a maximum 100. However, we are not sure about the expected speed and linearity of the conversation trajectories, so we set these parameters to 50. This filtration returns us a sorted list of the hashtag-based conversation trajectories. However, hashtags that are not related to sports events, such as #traffic, #love, #nofilter, #weather (Fig. 7) are also included in this list with higher average cluster distance, compared to the rest of the hashtags. Therefore, by further filtering the results with a lower average cluster distance threshold, these non-relevant hashtags disappear.

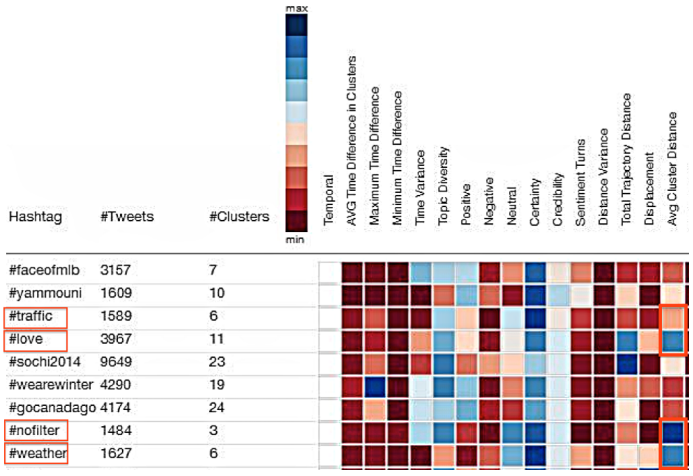

Fig. 7. Excerpt of the sorted list of hashtag-based trajectories.

These results, which now comprise of the top ranked 19 conversation trajectories are shown in Fig. 8 with a parallel coordinates plot to visualise the variance values of the trajectory characteristics derived in section III.

As prominently visible in the parallel coordinates plot in Fig. 8, #faceofmlb has a very high sentiment variance and low average credibility. As verified from ground-truth, #faceofmlb turns out to be a Twitter contest to select a major league baseball player to represent the playing season, based on votes from Tweeters.

When the #faceofmlb conversation trajectory is mapped on a geographic map as seen in Fig. 9, we can observe that at the beginning of the contest two specific players, Joey Votto and Felix Hernandez representing Cincinnati Reds and Seattle Mariners teams respectively, were being voted for. A sentiment and topic analysis of Tweets in these clusters indicate that Joey votto lost against Felix Hernandez.

However, upon analysing the average credibility of the clusters appearing close to Seattle and Cincinnati, we found a

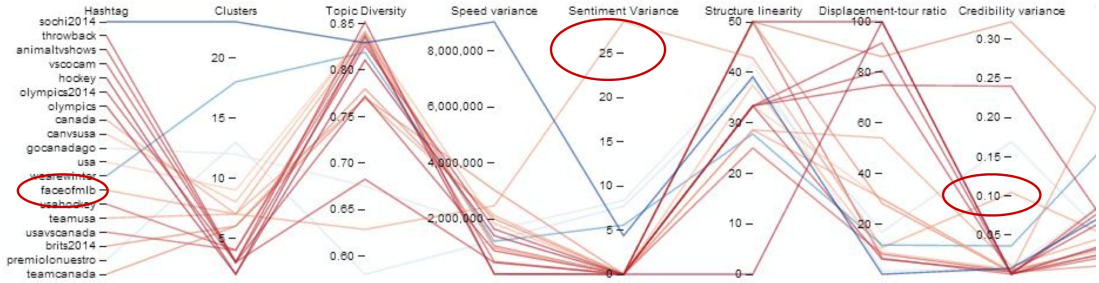

Fig. 8. The vertical axes of the parallel coordinates plot represents the derived characteristics of the top 19 conversation trajectories

Fig. 9. Content and sentiment analysis of the Tweets discussing the two players, Joey Votto and Felix Hernandez in #faceofmlb conversation trajectory.

smaller cluster in Seattle (representing the winning team player) with low credibility. This is shown Fig. 10 with the height of the bars. A Tweet in this cluster indicates that some voters cheated by voting many times using many fake accounts.

Fig. 10. Credibility of clusters mapped to the height of the bars.

V. Discussion & Future Research Perspectives

Tools similar to the presented approach, such as [33] or [34] are used in situations where the user possesses to some extent prior knowledge of what they want to explore. In our approach, in addition to the keyword-based relevance method, we rely on a grouping strategy to query and thereby filter the Twitter data based on the geospatial and content structure to derive trajectories. These structural analyses further help to detect changes in evolving conversations through the spatial, temporal, and contextual modalities. The key advantage of the visual analytics approach presented here, is that it caters to wider analysis and explorative possibilities that do not necessarily require the user to possess prior knowledge of the events. Further, our approach aims at reducing the uncertainties that inherently come with such data, thereby allowing to take well informed analytical decisions. The user-defined interestingness measure relies on the geographic and content characteristics of trajectories. This considerably saves discovery time in comparison to parameter browsing and searching, as done in previous works. An additional advantage here is that the user can narrow down the search space specific to the analysis interests, thereby reducing the uncertainties of the outcome.

The main disadvantage of using our hashtags-based approach is that it limits the data search space. While this helps to narrow down the data, this also removes a lot of Tweets from the dataset that may not use hashtags. We would like to tackle this problem, and explore methods to meaningfully include more Tweets into our dataset in future work. Furthermore, the uncertainty measures and the parameterisation of the model (e.g., ranking by interestingness features) for our use-cases were done based on a heuristic nature. Such heuristics can adapt to the anecdotal use-cases at hand. A compilation of such anecdotal use-cases may serve as a body of knowledge to learn from in future work.

Although we evaluate our approach through the anecdotal findings that are clarified by ground-truth (as also shown in [35]), in future work a comprehensive evaluation of our approach can be carried out for example by following the evaluation framework presented by [36], together with the validation guidelines by [35].

The framework of [36] is specifically developed for evaluating social media monitoring tools, addressing three issues: (1) the main concepts related to social media monitoring such as, analysis, insights, engagements etc., (2) the technology used, and (3) user interface. The focus thus far has been the identification of characteristics to meaningfully explore the hashtag-based conversation trajectories. Therefore, extensive content analysis using Natural Language Processing (NLP) has not been a priority in this paper. In future research, it is also aimed to conduct more in-depth text analysis on the conversation trajectories using NLP methods.

VI. Conclusions

In this paper we have presented a visual-interactive approach to detect events from Twitter, by exploring the various geospatial and content characteristics of its conversation movement trajectories. Unlike the popular keyword-based techniques to extract trajectories, we use a grouping strategy that considers the geospatial and content structures as the characteristics to filter out the meaningful and interesting hashtags, and the conversation trajectories based on these hashtags. As geospatial characteristics we have derived *distance variance*, *trajectory linearity*, and the *speed variance*, and as content characteristics we have derived *topic diversity*, *sentiment linearity*, *certainty variance*, and *credibility variance* to identify interesting conversation trajectories. Relying on the exploratory capabilities of the tool that implements our methodology, this approach does not require us to possess prior knowledge of events on Twitter. The derived geospatial and content characteristics are further used as feature dimensions to rank and sort the conversation trajectories based on what we want to explore (called an interestingness measure). The usefulness of this approach is demonstrated as a proof-of-concept to detect sports-related events.

References

[1] Y. Zheng and X. Zhou, *Computing with spatial trajectories*. Springer, 2011.

[2] H. Senaratne, A. Bröring, T. Schreck, and D. Lehle, "Moving on twitter: using episodic hotspot and drift analysis to detect and characterise spatial trajectories," in *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 2014, pp. 23–30.

[3] J.-G. Lee, J. Han, and X. Li, "A unifying framework of mining trajectory patterns of various temporal tightness," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1478–1490, 2014.

[4] H.-R. Wu, M.-Y. Yeh, and M.-S. Chen, "Profiling moving objects by dividing and clustering trajectories spatiotemporally," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2615–2628, 2012.

[5] H. Senaratne, M. Mueller, M. Behrisch, F. Lalanne, J. Bustos-Jiménez, J. Schneidewind, D. Keim, and T. Schreck, "Urban mobility analysis with mobile network data: a visual analytics approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1537–1546, 2017.

[6] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom, "Thematic patterns in georeferenced tweets through space-time visual analytics," *Computing in Science & Engineering*, vol. 15, no. 3, pp. 72–82, 2013.

[7] N. Adrienko and G. Adrienko, "Spatial generalization and aggregation of massive movement data," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 2, pp. 205–219, 2011.

[8] T. von Landesberger, S. Bremm, T. Schreck, and D. Fellner, "Feature-based automatic identification of interesting data segments in group movement data," *Sage Information Visualization*, vol. 13, no. 3, pp. 190–212, 2014, peer-reviewed article.

[9] J. Chae, Y. Cui, Y. Jang, G. Wang, A. Malik, and D. S. Ebert, "Trajectory-based visual analytics for anomalous human movement analysis using social media." in *EuroVA@ EuroVis*, 2015, pp. 43–47.

[10] M. A. Beber, C. A. Ferrero, R. Fileto, and V. Bogorny, "Towards activity recognition in moving object trajectories from twitter data." in *GeoInfo*, 2016, pp. 68–79.

[11] R. Krueger, G. Sun, F. Beck, R. Liang, and T. Ertl, "Traveldiff: Visual comparison analytics for massive movement patterns derived from twitter," in *2016 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2016, pp. 176–183.

[12] F. Luo, G. Cao, K. Mulligan, and X. Li, "Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of chicago," *Applied Geography*, vol. 70, pp. 11–25, 2016.

[13] X. Sui, Z. Chen, L. Guo, K. Wu, J. Ma, and G. Wang, "Social media as sensor in real world: movement trajectory detection with microblog," *Soft Computing*, vol. 21, no. 3, pp. 765–779, 2017.

[14] Q. Huanga and X. Liua, "Semantic trajectory inference from geo-tagged tweets," *Abstracts of the ICA*, vol. 1, 2019.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[16] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[17] A. Weiler, M. Grossniklaus, and M. H. Scholl, "Situation monitoring of urban areas using social media data streams," *Information Systems*, vol. 57, pp. 129–141, 2016.

[18] C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[19] R. M. Reese and A. Bhatia, *Natural Language Processing with Java: Techniques for building machine learning and neural network models for NLP*. Packt Publishing Ltd, 2018.

[20] L. Xu and H. Zhong, "Detecting inconsistent thrown exceptions," in *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)*. IEEE, 2021, pp. 391–395.

[21] M. Srinivasan, M. P. Shahri, I. Kahanda, and U. Kanewala, "Quality assurance of bioinformatics software: a case study of testing a biomedical text processing tool using metamorphic testing," in *Proceedings of the 3rd International Workshop on Metamorphic Testing*, 2018, pp. 26–33.

[22] V. V. Bochkarev, A. V. Shevlyakova, and V. D. Solovyev, "Average word length dynamics as indicator of cultural changes in society," *arXiv preprint arXiv:1208.6109*, 2012.

[23] E. H. Simpson, "Measurement of diversity." *Nature*, 1949.

[24] A.-M. Popescu and M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1873–1876.

[25] M. Tsytsarau, T. Palpanas, and K. Denecke, "Scalable discovery of contradictions on the web," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1195–1196.

[26] S. Kent, "Words of estimative probability," *Studies in Intelligence*, vol. 8, no. 4, pp. 49–65, 1964.

[27] P. Campbell, "Understanding the receivers and the reception of science's uncertain messages," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1956, pp. 4891–4912, 2011.

[28] C. Hovland, I. Janis, and H. Kelley, "Communication and persuasion; psychological studies of opinion change." 1953.

[29] A. Flanagin and M. Metzger, "The credibility of volunteered geographic information," *GeoJournal*, vol. 72, no. 3, pp. 137–148, 2008.

[30] J. Frew, "Provenance and volunteered geographic information," *Retrieved March*, vol. 10, p. 2008, 2007.

[31] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 675–684.

[32] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing?: understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 441–450.

[33] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 181–190.

[34] H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, and T. Ertl, "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2022–2031, 2013.

[35] T. Munzner, "A nested model for visualization design and validation," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 921–928, 2009.

[36] I. Stavrakantonakis, A.-E. Gagiu, H. Kasper, I. Toma, and A. Thalhammer, "An approach for evaluation of social media monitoring tools," *Common Value Management*, vol. 52, no. 1, pp. 52–64, 2012.