Factuality Drift Assessment by Lexical Markers in Resolved Rumors

Piroska Lendvai
Computational Linguistics
Saarland University
Saarbrücken, Germany
piroska.r@gmail.com

Uwe D. Reichel
Research Institute for
Linguistics, HAS
Budapest, Hungary
uwe.reichel@nytud.mta.hu

Thierry Declerck
Computational Linguistics
Saarland University
Saarbrücken, Germany
declerck@dfki.de

ABSTRACT

Our study presents a social media content representation, visualization and assessment method for modeling the emergence and resolution of rumorous claims during crisis events. Interpreting the factuality with which a claim is expressed is typically context-dependent and can be subject to semantic drift. We identify and quantify temporally anchored distributional and polarity patterns of lexical cues present in microposts in order to track factuality trends in rumor timelines. The findings on our English dataset are discussed with respect to provenance modeling for the social media domain; we additionally show how to port the method to content in German.

CCS Concepts

•Mathematics of computing → Time series analysis; •Computing methodologies → Information extraction; Discourse, dialogue and pragmatics; Lexical semantics; •Information systems → *Social networks;*

Keywords

Claim verification; Factuality; Provenance; Social media; Trend analysis

1. INTRODUCTION

1.1 Factuality in claim verification

The automatic verification of claims, regardless whether they appear in social or mainstream media platforms, is positioned high on the wish list of not only the news industry but also of the research field of language technology. Linguistic vehicles and their mapping to speaker uncertainty, speculation, controversy on the one end of the spectrum and source attribution, evidence providing and confidence on the other receive growing attention in several applied academic fields such as polarity and stance detection, factuality and claim checking, argumentation research, and computational journalism.

A substantial body of computational linguistics studies exist in which extra-propositional aspects of meaning such as speculation and negation are analyzed, including benchmark corpora that encode factuality [10, 4] and systems built for factuality detection (e.g. [1, 13]). However, these resources target the processing of texts from the literary, biomedical, encyclopedia and newswire genres, whereas in social media posts the syntax and lexicon is known to be different from that of standardized language.

Recently, as a result of social media technologies gaining both popularity and practical exploitation, large quantities of freely accessible textual data became available for a variety of unexplored domains. A new segment of life where these platforms are being utilised is crisis situations; while both private persons and media outlets have been observed to post messages on e.g. Twitter and Facebook, the level of trustworthiness and authoritativeness of information sources in such user-generated content emerging varies on a wide range. The quality of scarce pieces of information in crisis situations can often be of crucial importance, whereas different claims emerge in large quantities in real time, therefore it is imperative that automatic claim verification be available to decide which ones to endorse or debunk.

A growing amount of studies assessed the spreading of rumors on social media [8, 7], as well as that of memes that keep cycling in the media [5]. In order to yield reliable trends and trustworthy conclusions, such analyses must aggregate information obtained from various layers pertaining to the content to be verified. Next to network analysis for information propagation in terms of hubs, authorities, spatial and temporal factors, procedures that are put to use for verification purposes process platform-specific metadata such as topic tags, sender details, identifiable entities, topically or structurally contextual posts, as well as formal features of the content such as the use of punctuation; cf. [11]. These approaches can already deliver powerful means, but equally important in most of the cases is what the semantics of the content conveys; such information is to be extracted via analyzing the linguistic properties of a post.

Factuality can be regarded as a speech act composed of the commitment, confidence and modality with which a speaker makes a statement (in our case: a tweeter posts a message), emerging in some context, but possibly already interpreted in another context. Context is dependent on information access, which is individually set and for out domain heavily anchored in time. Typically, pieces of context such as the resolution value of a claim emerge as time progresses, but may not yet be available at the time of posting a claim, so that the closer the interpretation takes place to the time of posting, the less local information may be available to readers for assessing factuality. In retrospective analysis however, annotators can act as oracles and characterize the factuality framing of a post in its global context, e.g. in terms of the resolution value of the rumor it pertains to, as well as its context in a window including both past and future information.

1.2 Task definition

A previous study on the same data collection that we analyze revealed that, contrary to intuition, tweeters post messages that sound equally confident before and after a rumor is resolved, irrespective of whether the rumor has later been resolved as true or false, where the confidence level of tweeters was manually labeled as high/medium/low, using crowdsourcing [15]. However, no linguistic analysis was conducted on the data yet about how such confidence is expressed on the level of linguistic cues. We have therefore implemented a method to identify and quantify if temporal distribution and polarity of lexical markers used for factuality framing are affected by a set of contextual aspects that are specific for the domain of social media, such as threaded conversations, and for claim verification, such as rumor resolution position and value. We structure and model the presence and polarity of lexical factuality cues in a visually interpretable way, and provide statistical trend analyses on a large set of microposts relating to 39 resolved rumors.

1.3 Factuality in provenance modeling

Our study is carried out in the context of the Pheme project[1] that focuses on the detection and the classification of rumors that emerge in social networks and online media. One task in the project is the building of new and the extension of existing ontologies that enable the modeling and reasoning about veracity. Entities in such models describe notions including disputed claims, information propagation, conversation structure and temporal features. Pheme relies on focused ontologies that model two use case domains – journalism and medicine –, reusing existing social media ontologies such as SIOC[2] and DLPO[3]. The PHEME model of rumors is grounded in the PROTON top-level ontology[4]. While SIOC and DLPO contain information about sources and authorship of posts, supporting the connections between content objects and other community-specific objects, they lack the explicit modeling of provenance information.

Provenance encodes the chronology of information sources pertaining to an object, used in order to assign the object a status of some authority level. It may be considered as a form of event annotation, from which the history of a data object may be reconstructed. An aggregated representation of contextual information, for which lexical cue utility is going to be examined in the current study, is going to lead to improved representation of the chronology and authoritativeness of information sources. [12] recently proposed an annotation scheme that focuses on speaker perspectives in terms of four layers (events, attribution, factuality and opinion), while the model PROV-O, published as a W3C recommendation[5] offers an ontological framework for our application-specific view on provenance. The relevant classes, properties and vocabulary in these resources will need to be examined as potential extensions to the Pheme ontology for the purpose of tracking factuality drift within rumorous claim timelines.

[1]https://www.pheme.eu/
[2]http://rdfs.org/sioc/spec/
[3]http://www.semanticdesktop.org/ontologies/2011/10/05/dlpo/
[4]See http://ontotext.com/products/proton/ and [2] for integration details.

2. DATA REPRESENTATION

We utilized a project-internal, annotated social media corpus collected from the Twitter platform[6], a subset of which is freely available[7]. We focused on microposts pertaining to three crisis events: Ottawa shooting[8], Sydney Siege[9], and Germanwings crash[10]; the latter both in English and German. Each of the three events gave rise to several rumors – plausible but at the time of their emergence unconfirmed statements. E.g. "Shots fired on Parliament Hill" and "There were three separate shooting incidents" are two of the 13 manually formulated claims annotated as resolved for the Ottawa shooting event. For each of the rumorous claims[11], tweets discussing the claim were manually identified, were organized into threaded conversations of thread-initiating tweets and replying tweets and several task-specific aspects in the conversation were marked up on the tweet- and the conversation level, such as rumor resolving value; for details we refer to [14] and [15]. A resolving tweet for e.g. the claim "One or more shots/live ammunition have gone off at the cafe" is: *POLICE MOVE IN: Police confirm live ammunition used in Martin Place #sydneysiege*; this rumor was manually annotated as having been resolved True. Altogether we had 26 rumors resolved as True discussed in 8,918 posts, and 13 rumors resolved as False discussed in 2,291 posts.

2.1 Lexical cues as factuality markers

The domain of our study is user-generated content. Since the data collection method kept only posts that passed a retweet count threshold, a portion of the tweets originate from authoritative sources and feature well-formed language. Since discussions contain replying tweets, the material features plenty of instances of the ill-famed non-standard social platform language use.

Based on the factuality literature, we devised four factuality groups and populated them with about 20-40 cues:

- **knowledge cues**, e.g. 'admit', 'confirm', 'correct', 'discover', 'identify', 'learn', 'name', 'reveal'
- **report cues**, e.g. 'claim', 'footage', 'observe', 'report', 'say', 'show', 'tell', 'video'
- **belief cues**, e.g. 'assume', 'believe', 'perhaps', 'predict', 'suggest', 'seem', 'think'

[5]https://www.w3.org/TR/prov-o/
[6]twitter.com
[7]https://figshare.com/articles/PHEME_rumor_scheme_dataset_journalism_use_ case/2068650
[8]https://en.wikipedia.org/wiki/2014_shootings_at_Parliament_Hill,_Ottawa
[9]https://en.wikipedia.org/wiki/2014_Sydney_hostage_crisis
[10]https://en.wikipedia.org/wiki/Germanwings_Flight_9525
[11]Throughout the paper, we are going to use *claim* and *rumor* interchangeably to refer to the same concept.

- **doubt cues**, e.g. '?', 'ask', 'contrary', 'deny', 'incorrect', 'misstate', 'not', 'why'.

Starting from these basic lexical items, each of the four lists were automatically further populated from available semantic resources: we extracted the top-3 most similar items from the pretrained Google News word embeddings vector[12], as well as lemmas from the top-3 synsets from the English WordNet via NLTK[13]. Only single-token items were harvested; each cue token was subsequently extended by its derivationally related forms via the corresponding NLTK function. Using the extended cue lists, we obtained counts for each tweet via matching each cue list to a tweet's content, applying the NLTK Snowball Stemmer[14] prior to lookup. E.g. in the tweet "*Rideau Centre general manager tell @globeandmail there was no shooting inside the mall. But people can't leave or enter. #OttawaShooting*" we match one report cue (*tell*) and two doubt cues (*no*, *can't*).

Since determining the scope of negation is known to be difficult to determine [6], as e.g. the function of negation in dependable on discourse context, we experimented with using dependency parsing to increase the precision on extracting doubt cues that originate from negated constituents, but abandoned it for no proven impact on the current setup and content.

From the raw counts obtained from the extended cue set lookup we derived a metric that we call Factuality Cue Ratio (FCR), and define it as

$$\frac{\#\text{knowledge\_cues} + \#\text{report\_cues} + \#\text{belief\_cues}}{\#\text{all\_cues}}$$

The metric expresses the proportion of matched cues signaling affirmative factuality polarity – on different certainty levels – over all matched cues, motivated by polarity opposition as in our setup doubt cues are designed to signal negative factuality polarity. The FCR for the above example tweet is .33.

Based on temporally aggregated FCR of individual tweets, we created color- and size-coded timeline plots for claims; see Figure 2 for the Germanwings event. For visualizing these data as time series, the posting time of the rumor-resolving tweet was central. For each rumor, there is always only a single rumor resolving tweet, which is marked by a triangle. The posting time of each tweet was related to its rumor's resolution time and discretized in terms of 10 minute intervals. The amount of evidence is encoded by size, e.g. large bubbles at time point 0 show increased activity around the rumor resolution point. From the visualization, temporal patterns and factuality patterns of tweeting activity can be straightforwardly inferred. For example, in our English Germanwings corpus tweeting intensity seems to be greater before a claim is resolved than after it both for rumours resolved as True and False, or that in the minutes immediately following claim resolution there is a burst in activity which is larger for rumors resolved as False than for rumors resolved as True. Yet another visually identifiable effect concerns the impact of using the basic vs extended cue set: when we use the extended cue set (shown on True rumors, middle and bottom plot), triangles are not blue, i.e. there are always some factuality cues present in the resolving tweets, both in True and False rumor cases. For this particular event, posts resolving True claims feature a higher FCR ratio than posts resolving False claims.

In Section 3 we are going to derive variables from this data representation, and statistically assess them in order to characterize trends that describe tweet- and rumor-level factuality drift.

3. FACTUALITY DRIFT ASSESSMENT

We aimed to characterize surface-level lexical cues as a dependence of their context in terms of the following properties: temporal (the timeline of the emergence and resolution of rumors), conversational (microposts appear in threaded conversations and pertain to rumors, rumors relate to real-life events), and claim-value related (i.e., a rumor is resolved either as true or false). To achieve this, we had to translate these properties into corresponding features. Our approach was to devise features that quantify local patterns of factuality, in terms of (i) tweet-intrinsic features and (ii) cross-tweet delta features, as well as global patterns of factuality in terms of (iii) trend analysis. The factuality cue ratio (*FCR*) derived from lexical cues was a central variable in the analysis.

3.1 Method

The relationship between factuality cues and rumor resolution was examined within analysis windows of three different sizes for:

- (i) tweet-intrinsic properties
- (ii) local discontinuity across subsequent tweets, and
- (iii) global discontinuity in linear cue trends.

For (i), we examined the relationship between factuality cues and rumor resolution in terms of FCR in resolving vs non-resolving tweets. For (ii), we captured the degree of local FCR discontinuity for each rumor at each tweet position by two variables:

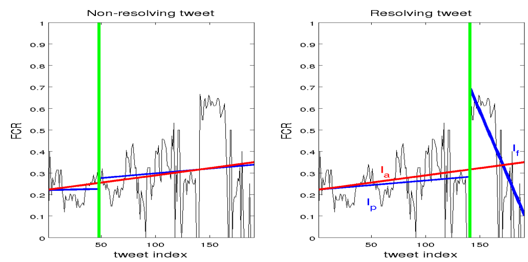

Figure 1: FCR trend analysis. For each tweet (position marked in green), three regression lines are fitted to the FCR sequence (plotted black): to its preceding and following sequences (l_p, l_f; blue), and to the entire rumor (l_a; red), the latter representing the general trend. For resolving tweets (right) l_p and l_f deviate more from the common trend, shown by a larger reset and larger root mean squared deviation values from the overall trend line l_a.

[12]http://code.google.com/p/word2vec/

[13]http://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html

[14]http://www.nltk.org/api/nltk.stem.html

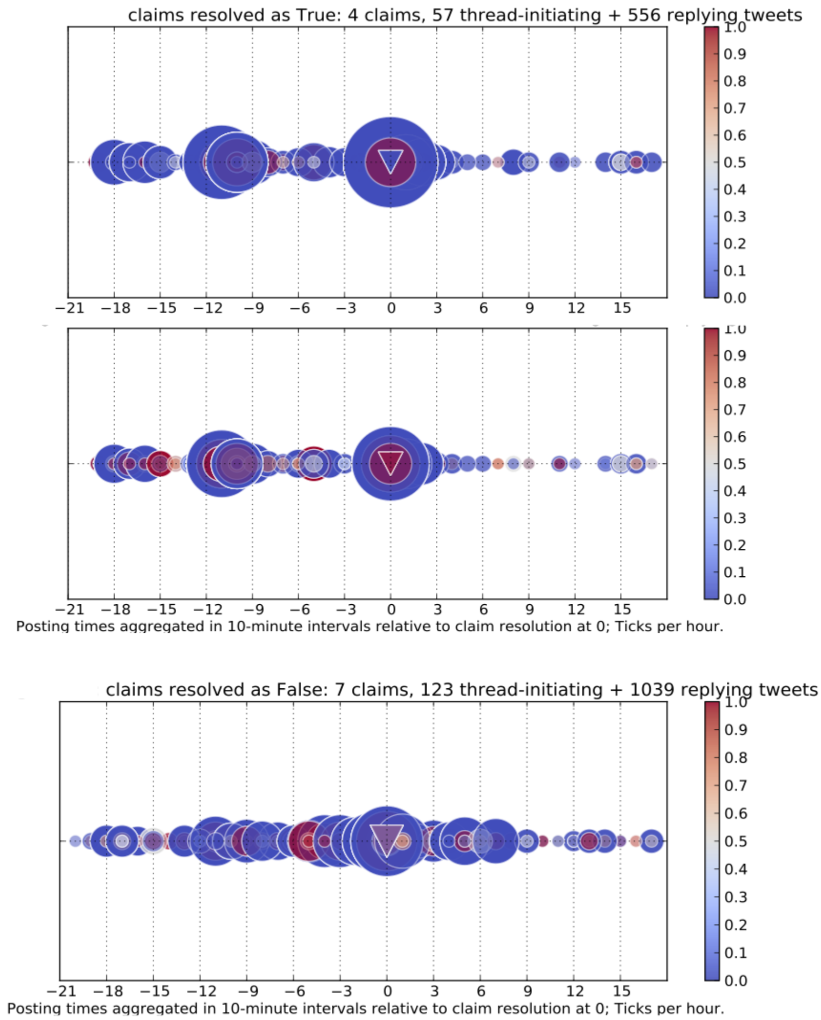

Figure 2: Partial timeline window of English tweets discussing Germanwings rumors resolved as True (top and middle: basic cue set vs extended cue set) vs False (bottom, extended cue set). X axis: tweeting times aggregated to 10 minute intervals, relative to rumor resolving tweets (as triangles) positioned at value 0. Color-coded factuality in terms of lexical cue presence ratio, size-coded amount of evidence.

- *FCR_Delta*, the deviation of a tweet i from the temporally preceding tweet $i-1$, measured by subtracting the FCR of tweet $i-1$ from the FCR of tweet i
- *FCR_Diff*, calculated for each tweet i as the FCR mean values in a symmetric window of 10 minutes length, where the FCR mean of the left (preceding) window half is subtracted from the mean of the (subsequent) right window half.

For (iii), three regression lines were fitted: line l_p through the FCR values of the preceding tweet sequence $1 \dots i-1$, line l_f through the FCR values of the subsequent tweet sequence $i+1 \dots n$ (n be the number of tweets in a claim), and line l_a through the entire tweet sequence $1 \dots n$.

The method is illustrated in Figure 1. In order to measure the amount of FCR discontinuity at each tweet, we calculated

- the reset, i.e. the absolute distance between the FCR offset of l_p and the FCR onset of l_f (*Reset*), as well as
- the deviation of each of l_p and l_f from l_a, in terms of root mean squared deviation (*RMSD_pre*, *RMSD_post*, respectively).

The method was adopted from intonation research [9], where it is used to quantify pitch discontinuities for prosodic boundary strength prediction. Applying the reasoning of [9], *Reset* quantifies the FCR disruption at each tweet position, and *RMSD_pre,post* quantify the deviation of preceding and following tweets' regression lines from a common trend.

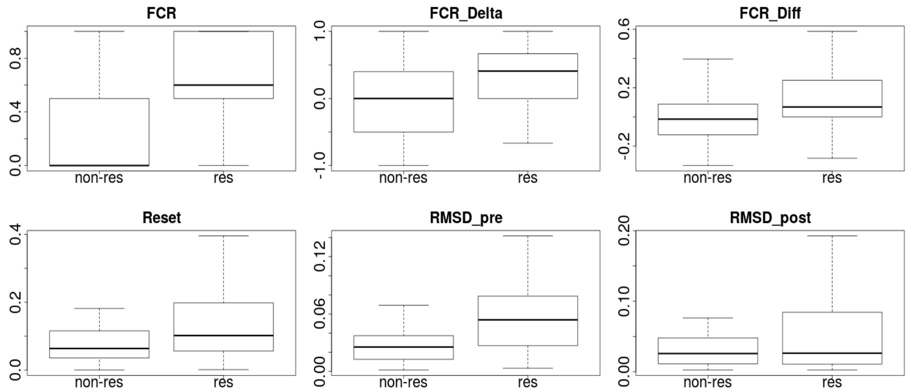

Figure 3: Comparison of FCR-derived variables between non-resolving (*non-res*) and resolving (*res*) tweets. The results indicate higher tweet-intrinsic FCR values for resolving tweets, as well as higher local FCR discontinuities (*FCR_Delta, FCR_Diff*) and a higher impact on global FCR trends (*Reset, RMSD_pre, RMSD_post*).

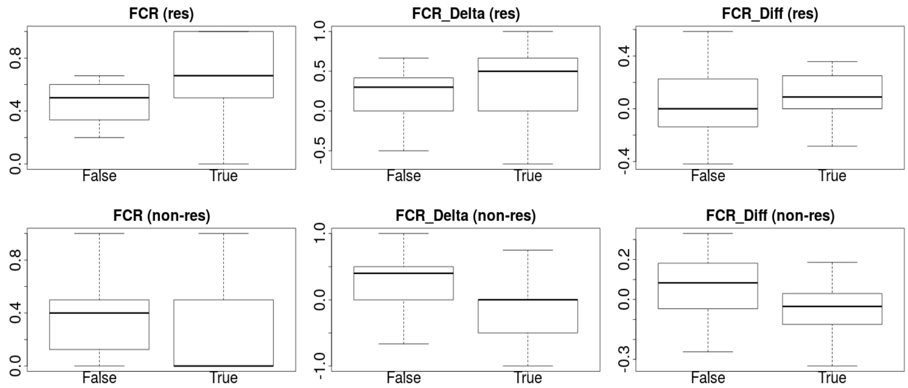

Figure 4: Comparison of FCR-derived variables in rumors resolved as *False* vs *True*. Upper row: resolving tweets, lower row: non-resolving tweets. When rumors get resolved as True, FCR in the resolving tweet is higher, and there is more increase of local FCR (*FCR_delta, FCR_Diff*) than when rumors get resolved as False. For non-resolving tweets, the opposite pattern is found: lower *FCR, FCR_Delta*, and *FCR_Diff* in True rumors vs in False rumors.

3.2 Results

For all our six measures, the difference between resolving and non-resolving tweets was statistically tested by linear mixed-effect models[15] with each of the six measures as the dependent variable, *+/− resolving tweet* as well as *Rumor resolved as True vs False* as the fixed effects, and *event* as random effect. Data were balanced with respect to the effect *+/− resolving tweet*.

The analysis results for the effects *+/− resolving tweet* and *rumor is Resolved as True vs False* are shown as boxplots in Figures 3 and 4, respectively. In Figure 4, claim resolution is shown separately for resolving and non-resolving tweets for those variables for which a (weakly) significant interaction between the two effects has been observed ($p < 0.1$ for *FCR_Delta*, $p < 0.05$ for *FCR, FCR_Diff*).

In the tweet-intrinsic analysis, one of the two local discontinuity (*FCR_Delta*) and all global discontinuity variables values turned out significantly higher in resolving tweets than in non-resolving tweets ($p < 0.01$ for *FCR*, $p < 0.05$ for the others). No significant difference was observed with respect to the value of the rumor resolution, however, three interactions were identified between the two effects for *FCR, FCR_Delta*, and *FCR_Diff*. These indicate that rumor resolution values affect the intrinsic FCR value and its local discontinuities in a different way in resolving and non-resolving tweets: in rumors resolved as True all three FCR values measurements are higher in the resolving tweet, that is, in True rumors factuality appears to be more emphasized in the resolving tweet, and there is a higher amount of factuality support increase relative to temporally preceding tweets. For non-resolving tweets the opposite tendencies were found: in rumors that were resolved as True, we observed lower intrinsic factuality and more local factuality decrease, as opposed to False rumors.

[15] as implemented in the `lme4` package in R

4. ANALYSIS OF GERMAN TWEETS

The German-language tweet corpus includes four resolved rumors in terms of 45 thread-initiating and 278 replying tweets that discuss the Germanwings event. All the four rumors are verified as True. An example of a resolved rumor is Only one pilot was in the cockpit at the time of the crash / the other pilot was (deliberately) locked out. The corresponding resolving tweet text is: *Staatsanwaltschaft Marseille: Co-Pilot hat Abwesenheit des Kapitäns ausgenutzt. Ob es geplant war, wissen wir nicht. #Germanwings #4U9525* ('Public Prosecutor's Office Marseille: Co-pilot exploited captain's absence. We don't know if it was planned.')

To examine the cross-language portability of our factuality drift assessment method, we created German cue word lists, equivalent to the English basic cues exemplified in Section 2.1. Examples of German tweets that express doubt are given below.

- *@zeitonline @Tristan8002 das ist niemals ein Gleitflug, der würde viel länger dauern.* ('it is never a gliding flight, that would take much longer')
- *@SPIEGELONLINE ich frage mich warum Hollande dazu schon was sagt wenn es noch nicht von @germanwings bestätigt wurde...* ('I wonder why Hollande comments on it already when it has not yet been confirmed by @germanwings...')

In the examples, doubt cues include negation words such as *niemals* ('never') or *nicht* ('not'), the verb *fragen* ('ask') and the adverb *warum* ('why'). Tokens that pertain to affirmative cues for the report and knowledge factuality types are also present, i.e. *sagt* ('says') and *bestätigt* ('confirms') .

For detecting lexical cues in the German data, we make use of an in-house lexical resource and processing tools suite, the prototype for which is described in [3]. The lexical base contains currently 21.5k lemmas expanding to 3.3 million full-forms, also including frequent typos occurring on social media, e.g. *bekommnn* next to *bekommen* ('to get'). High level of morphological variation in the German lexicon particularly holds for verbs and adjectives, while the complexity of nouns origins in highly productive compounding. In the current study, we refrained from complex morphosyntactic analyses – in our current cue matching procedure, we use word list lookup. The most frequently matched cues in our German corpus are *kein* ('not a'), *kann* ('can'), *Frage* ('question'), *wissen* ('to know'). Since the dataset is very small, and False rumors are absent, in order to conduct full comparison with the English data we need to collect more German microposts. The visual analysis chart corresponding to our current German data is shown in Figure 5.

5. CONCLUSIONS AND DISCUSSION

We have shown a method to utilize lexical cues for encoding and tracking of factuality drift in timelines of rumorous microposts. For the end task of supporting journalists to verify claims appearing on social media portals, our general goal is to mark up claims in terms of claim origin and factuality type, for which we aimed to find patterns that show how factuality levels in microposts relate to the resolution of the discussed claim. Such mark up supplies important information for a claim's provenance, especially in the verification scenario, as it enables to contextualize claims in an

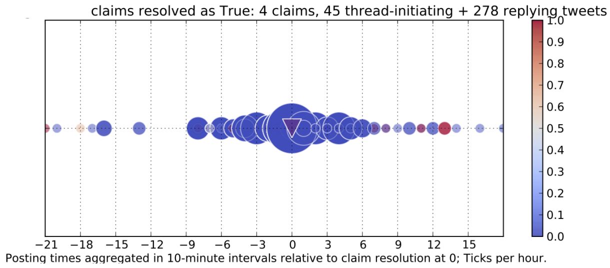

Figure 5: Partial time window representing factuality drift in rumors verified as True in German microposts discussing the Germanwings event.

objective way, as factuality judgments can be associated to a set of views and information sources.

The quantitative observations gained from our analysis of the impact of resolving tweets on a factuality trend confirm the qualitative trend suggested by the aggregated and normalized FCR metric as visualized on claim timelines. For the data at hand, factuality drift is characterized by the following properties:

- the presence of a rumor-resolving tweet induces a significant change in the rumor-level factuality cue ratio (FCR) trend (*Reset, RMSD_pre, RMSD_post*),
- the presence of the resolving tweet induces prominent increase in local FCR (*FCR_Delta*)
- resolving tweets contain higher FCR than non-resolving tweets.

Patterns of the factuality drift were found to be influenced by claim resolution values: for claims resolved as True, the resolving tweet tends to contain more factuality cues and forms the origin of local FCR increase. However, we did not find evidence that tweets discussing (but not resolving) True rumors would contain more affirmative factuality cues than negative cues, which however holds for tweets discussing (but not resolving) False rumors.

In ongoing work we are extending the current method with more complex linguistic analysis as well as with features encoding tweeter certainty as an additional attribute of factuality. The insights gained from the current study are used to design tweet- and rumor-level classification experiments, for tagging claims with factuality judgments and populating verification-focused ontologies for enhanced contextual retrieval and reasoning.

6. ACKNOWLEDGMENTS

This study has been in part supported by the PHEME FP7 project (Grant No. 611233); the work of the second author is supported by a grant of the Alexander von Humboldt Society.

7. REFERENCES

[1] M.-C. de Marneffe, C. D. Manning, and C. Potts. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333, 2012.

[2] T. Declerck, P. Osenova, G. Georgiev, and P. Lendvai. Ontological modelling of rumors. In *Linguistic Linked Open Data*, pages 3–17. Springer, 4 2016.

[3] T. Declerck and M. Vela. Linguistic dependencies as a basis for the extraction of semantic relations. In C. Wroe, R. Gaizauskas, and C. Blaschke, editors, *ECCB'05 Workshop on Biomedical Ontologies and Text Processing*, 2005.

[4] R. Farkas, V. Vincze, G. Móra, J. Csirik, and G. Szarvas. The CoNLL-2010 Shared Task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the 14th Conference on Natural Language Learning*, 2010.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD-09*, 2009.

[6] R. Morante and E. Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 2012.

[7] R. Procter, F. Vis, and A. Voss. Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214, 2013.

[8] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1589–1599, 2011.

[9] U. D. Reichel and K. Mády. Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian. In *Proc. Interspeech 2014*, pages 111–115, 2014.

[10] R. Saurí and J. Pustejovsky. Factbank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3), 2009.

[11] L. Tolosi, A. Tagarev, and G. Georgiev. An Analysis of Event-Agnostic Features for Rumour Classification in Twitter. In *Proc. of Social Media in the Newsroom Workshop*, 2016.

[12] C. van Son, T. Caselli, A. Fokkens, I. Maks, R. Morante, L. Aroyo, and P. Vossen. GRaSP: A Multilayered Annotation Scheme for Perspectives. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

[13] E. Velldal and J. Read. Factuality detection on the cheap: Inferring factuality for increased precision in detecting negated events. In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, 2012.

[14] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Towards Detecting Rumours in Social Media. *CoRR*, abs/1504.04712, 2015.

[15] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3), 2016.