Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

Real-World Witness Detection in Social Media via Hybrid Crowdsensing

Stefano Cresci,[1] **Andrea Cimino,**[2] **Marco Avvenuti,**[3]
Maurizio Tesconi,[1] **Felice Dell'Orletta**[2]
[1] Institute of Informatics and Telematics, IIT-CNR, Italy {stefano.cresci, maurizio.tesconi}@iit.cnr.it
[2] Institute of Computational Linguistics, ILC-CNR, Italy {andrea.cimino, felice.dellorletta}@ilc.cnr.it
[3] Department of Information Engineering, University of Pisa, Italy marco.avvenuti@unipi.it

Abstract

The task of witness detection in social media is crucial for many practical applications, including rumor debunking, emergency management, and public opinion mining. Yet to date, it has been approached in an approximated way. We propose a method for addressing witness detection in a strict and realistic fashion. By employing hybrid crowdsensing over Twitter, we contact real-life witnesses and use their reactions to build a strong ground-truth, thus avoiding a manual, subjective annotation of the dataset. Using this dataset, we develop a witness detection system based on a machine learning classifier using a wide set of linguistic features and metadata associated with the tweets.

Introduction

Nowadays, social media have become one of the most important channels of information diffusion and the primary mean to understand the activities, the preferences, and the opinions of people. In all social media – and in particular in those with remarkable real-time features such as Twitter, Instagram, and Weibo – users increasingly report what is happening to and around them. For these reasons, social media have been recently used as information channels to broadcast live and unfolding events (Petrović, Osborne, and Lavrenko 2010). The feature of social media to be real-time information channels and the feature of social media users to be considered as distributed journalists or "social sensors", gave rise to the phenomenon of citizen and participative journalism (Allan 2009). In citizen journalism, it is important to distinguish between those few *witness* reports and all other social media messages that represent noise, unrelated content, and second-hand information. Identifying witnesses and witness messages is crucial for a number of reasons, such as to interview eyewitnesses of an event (Diakopoulos, De Choudhury, and Naaman 2012), in emergency management to know about people that are on-the-ground (Starbird, Muzny, and Palen 2012; Avvenuti et al. 2015), to estimate event attendance (de Lira et al. 2017), and in public opinion research to solicit feedback only from those users that experienced a given situation first-hand (Murphy et al. 2014).

Challengingly, witness detection can be regarded as the task of "finding a needle in a haystack", where a small number of witness messages is typically overwhelmed by the deluge of messages that is continuously shared in social media. Only recently Academia turned its attention to the task of witness detection with the goal of providing automatic, real-time systems capable of accurately distinguishing between *witness* and *non-witness* messages in social media (Fang et al. 2016). Yet to date, this task has been approached in an approximated, non-systematic way, thus limiting the real-world applicability of the proposed systems.

In this work we developed the first witness detection system for Italian tweets. We started by building a ground-truth corpus comprising 2,599 tweets, collected by using the hybrid crowdsensing technique (Avvenuti et al. 2017). The main novelty of this corpus is the annotation method, as here the separation into witness and non-witness messages is done using first-hand answers from contacted witnesses. This is in contrast with the typical, error-prone manual annotation process, aimed at inferring the correct classification from textual information. Then, we used the ground-truth to train, and to experiment with, a binary witness detection classifier. The main contributions of our witness detection system can be summarized as follows: (i) we employ a novel automatic data collection and annotation strategy that allows us to build a strong and realistic ground-truth; (ii) we propose a witness classifier based on a large number of linguistic features and metadata extracted from tweets.

Related work

Although relevant for a number of practical applications, automatic witness detection has to date received little attention from the scientific community. In WordNet and in (Diakopoulos, De Choudhury, and Naaman 2012), witnesses are respectively defined as "someone who sees an event and reports what happens" and "people who see, hear, or know by personal experience and perception". We ground our work on these definitions, as is also done in the system described in (Fang et al. 2016), which can be considered the state-of-the-art in automatic witness detection. This system, trained on a manually annotated corpus of emergency-related tweets, exploits linguistic and metadata features and achieves an average F-score of 89.7% on test data. Despite the promising results, the main flaws of (Fang et al. 2016)

are related to the manually annotated dataset used to train the system. In fact, given that human annotators did not know whether a tweet was posted by a witness or not, the separation into the *witness* or *non-witness* class was inferred from textual information, thus reflecting only an approximation of the real situation. Moreover, such dataset is highly unbalanced, with only 401 *witness* tweets, corresponding to 0.3% of the whole dataset, thus raising concerns on the generalizability of results. The system described in (Morstatter et al. 2014) exploits a set of linguistic features and aims at detecting those tweets that originate from within a crisis region, in the aftermath of an emergency. Authors make the assumption that tweets from within the crisis region are also witness tweets. As also done by (Fang et al. 2016), we consider this assumption to be an approximation of the problem. Nonetheless, in our work we also included features to consider the geographic distance of a tweet from an event, since this might be an indicator of a *witness* tweet. (Starbird, Muzny, and Palen 2012) tackle the similar problem of identifying on-the-ground users during mass emergencies. They propose a machine-learning system that exploits collaborative filtering techniques and achieves an average accuracy of 77.2%. Lastly, (Truelove, Vasardani, and Winter 2015) carried out a preliminary study aimed at highlighting the characteristics that help distinguishing between *witness* and *non-witness* messages. Although they do not provide any algorithm nor a system to automatically perform the classification, the results of their studies provide useful insights that we leveraged while designing our machine-learning features. This brief survey of recent literature in witness detections highlights that few efforts have been done in this task. Moreover, until now witness detection has been tackled in an approximated fashion.

Ground-truth

Data collection and annotation method. In all previous works data was collected via an opportunistic approach, e.g. by listening to the hashtag of an event and collecting all tweets containing that hashtag. This unlabeled data was then manually annotated by a number of human operators that inferred the label (*witness* or *non-witness*) of every tweet from its textual information. In order to create a more realistic ground-truth, thus overcoming the limitations of previous manually annotated datasets, in this work we employed a novel data collection strategy called *hybrid crowdsensing* (Avvenuti et al. 2017). Hybrid crowdsensing combines opportunistic and participatory sensing phases in order to obtain more high-quality data during social media crowdsensing campaigns, compared to traditional approaches. In particular, our hybrid crowdsensing system first collects data in an opportunistic fashion. Then, in a participatory phase, the system automatically contacts authors of relevant tweets and asks them whether they were physically attending to the event (*witness*) or just commenting about it from a different location (*non-witness*). Questions are sent to users in reply to their original tweets, and answers, if any, are collected by a Twitter crawler. Parsing the received answers allows us to automatically label the users (and their tweets) as witness or non-witness. This way, instead of performing a time consuming and error-prone manual annotation, we effectively make the users themselves label their own tweets. Two examples of such automatic "conversations" (translated from Italian) are shown in Table 1.

Ground-truth composition. A total of 6 months long data collection campaigns were carried out from June to November 2016. Over this time span, we collected and analyzed a total of $72,305$ tweets from 13 different real-world events that took place in Europe. For generalizability purposes, we focused on a broad set of events, including arts, folk, fashion, technology, music, and culture. All data was collected from Twitter by exploiting public APIs. As the implementation of a hybrid crowdsensing system is language-dependent, and given the timeliness required by the realtime nature of events, we limited data collection and analysis to tweets written in the Italian language.

Table 2 shows detailed statistics about the events covered by our study, together with the number of *witness* and *non-witness* tweets included in our ground-truth. Even though, as expected, not all users answered our questions, our data collection and annotation process resulted in a notable dataset size of $2,599$ tweets, of which $1,111$ (42.7%) are *witness* and $1,488$ (57.3%) are *non-witness* tweets. Original tweets for which we did not receive any answer were discarded and not used in the remainder of this work.

The advantages of our witness detection strategy over previous ones can be summarized as follows. Firstly, our ground-truth is likely to be much more realistic than previous ones, since we made users themselves tell us whether they were witnesses or not, thus avoiding arbitrary decisions of human annotators based on tweet content. Secondly, the number of *witness* tweets in our ground-truth is larger than in previous works. Also, it is well balanced with the number of *non-witness* tweets, thus addressing known limitations of past works in training machine-learning algorithms.

Witness detection system

We address witness detection as a binary classification task. Our classifier operates on morpho-syntactically tagged texts and is based on LIBSVM's implementation of quadratic Support Vector Machines (SVM). Since our approach relies on multi-level linguistic analysis, the documents belonging to our ground-truth were automatically morpho-syntactically tagged by the state-of-the-art POS tagger for Italian tweets described in (Cimino and Dell'Orletta 2016).

Witness detection features

In this study, we focused on a wide set of features ranging from different levels of linguistic description to information extracted from the tweets metadata. The linguistic features are organized in two main categories.

Raw and lexical text features: ***Number of tokens***. ***Character, Word, and Lemma*** n***-grams***. ***Repetition of*** n***-grams chars***: this feature checks the presence or absence of contiguous repetition of characters in the analyzed tweet. ***Mentions Number***. ***Hashtags number***. ***Punctuation***: checks

	witness conversation	*non-witness* conversation
user's original tweet	Dinner on the beach with @*anonymized* and a few roommates in #Venezia73	Musical, aliens, young popes, tv series, virtual reality: the Venice Film Festival #Venezia73 is very interesting. The ... http://fb.me/140v9U7xl
our automatic question	Hello @*anonymized*, we are monitoring the #biennale, are you there too?	Hi @*anonymized*, are you in Venice for the #Biennale?
user's answer	@*anonymized* Yes I am here from day one and I will stay till the end! What do you think so far?	@*anonymized* Not really. I follow it for passion and because I'm a frequent guest on tv by #Marzullo :-)

Table 1: Examples of hybrid crowdsensing for collecting and automatically labeling our witness ground-truth.

event	place	type	total tweets	answered *witness*	answered *non-witness*
Venice Biennial	Venice, Italy	Arts	49,349	340	708
Lucca Comics and Games	Lucca, Italy	Culture	4,987	260	220
Romics	Rome, Italy	Culture	6,971	243	210
Oktoberfest	Munich, Germany	Folk	4,351	42	188
Vogue Fashion's Night Out	Milan, Italy	Fashion	1,463	42	45
Vasco Live Kom	Rome, Italy	Music	741	41	21
Home Festival	Treviso, Italy	Music	893	41	14
Internet Festival	Pisa, Italy	Technology	1,051	35	10
Internationale Funkausstellung Berlin	Berlin, Germany	Technology	1,435	30	47
Sziget Festival	Budapest, Hungary	Music	665	21	25
Settembre Prato	Prato, Italy	Folk	237	11	0
Metarock	Pisa, Italy	Music	154	4	0
Prato Comics	Prato, Italy	Culture	8	1	0
Total			72,305	1,111	1,488

Table 2: Statistics about our Twitter dataset. Events are ordered in decreasing number of *witness* tweets.

whether the tweet finishes with one of the following punctuation characters: "?", "!".

Morpho-syntactic features: ***Coarse grained POS*** n***-grams***: presence or absence of contiguous sequences of coarse-grained POS, corresponding to the main grammatical categories. ***Fine grained POS*** n***-grams***: presence or absence of contiguous sequences of fine-grained POS, which represent subdivisions of the coarse-grained tags (e.g. the class of nouns is subdivided into proper vs. common nouns, etc.). ***Coarse grained POS distribution***: the distribution of nouns, adjectives, adverbs, numbers in the tweet.

The twitter metadata features are based on metadata information associated with tweets, available via Twitter APIs. ***Client***: the client used to post the tweet (e.g. Web client, Twitter for Android, etc.). ***Is reply***: `true` if the tweet is a reply to another tweet, `false` otherwise. ***Is geolocalized***: `true` if the tweet is geolocalized (tagged with GPS coordinates), `false` otherwise. ***Geographic distance***: tweets that are posted from the vicinity of the event are more likely to be posted by witnesses. Thus, if the tweet is geolocalized, this measures the geographical distance between the place from where the tweet has been posted and the location of the event. ***Temporal distance***: tweets that are posted during the event are more likely to be posted by witnesses. This feature measures the (absolute value of the) time passed between the start of the event and the timestamp of the tweet.

System evaluation

In order to evaluate our system, we performed a 5-fold cross validation with the goal of testing different feature configurations and of assessing the contribution of linguistic and metadata features in the witness detection task.

For each configuration, we computed the average F1 and the global Accuracy. In addition, we computed the Precision, Recall and F1 for each class. Results of the evaluation are reported in Table 3. The first row reports the results achieved by a baseline classifier, which always outputs the most probable (i.e., the majority) class, according to the class distribution of the dataset (the *non-witness* class). The second row reports the results achieved by the *Metadata* configuration which exploits only the metadata features and does not resort to any textual information. Even though in this configuration we considered only five features, we measure a general improvement with respect to the results achieved by the *baseline* classifier ($\simeq 14\%$ average F1 improvement). This result demonstrates that these features significantly contribute to the witness detection task. It is reasonable to assume that the usage of a mobile client and a short distance from the location of the event are strong witness indicators. Unfortunately, only a small portion of the tweets are geolocalized (Cresci et al. 2015; Avvenuti et al. 2018) thus somewhat limiting the usefulness of the *geographic distance* feature. However, for the future one can envision the possibility to employ geoparsing al-

			witness			*non-witness*		
configuration	*Accuracy*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
baseline	0.572	0.362	0.000	0.000	0.000	0.572	1.000	0.727
Metadata	0.609	0.535	0.669	0.276	0.357	0.617	**0.865**	0.713
Linguistic	0.697	0.681	0.673	0.566	0.615	0.710	0.795	0.750
Linguistic + Metadata	**0.701**	**0.687**	**0.679**	**0.572**	**0.621**	**0.714**	0.798	**0.753**

Table 3: Witness detection results for different configurations of the classifier. Highest values of evaluation metrics are in **bold**.

gorithms, such as those described in (Avvenuti et al. 2016; 2018), in order to increase the number of geolocalized tweets. The third row reports the results achieved by the *Linguistic* configuration that exploits only the linguistic features. As shown, we obtain a large improvement on the average F1 ($\simeq 13\%$) with respect to the *Metadata* configuration. More interestingly, the F1 obtained on the *witness* class improves even more: $\simeq 25\%$. This result shows that the information contained in the text is strongly relevant for the witness detection task. The fourth row reports the results achieved by the *Linguistic + Metadata* configuration that simultaneously exploits both the linguistic features and the metadata features. In our experiments this configuration obtained the best results on the most relevant metrics (average F1 and *witness* F1), showing that metadata and linguistic features are complementary.

Conclusions

We presented the first real-world witness detection study in a social media scenario. To achieve this goal, we adopted a novel strategy to create a realistic ground-truth of *witness* and *non-witness* users. Our ground-truth is based on actual answers by Twitter users, rather than on arbitrary manual annotation. This dataset was used to develop a machine learning classifier based on a wide set of features ranging across different levels of linguistic description, and on information extracted from the tweets metadata. The proposed data collection and annotation strategy, based on hybrid crowdsensing, can be easily employed in many other tasks in order to build better ground-truths.

Acknowledgements This research is supported in part by the EU H2020 Program under the schemes INFRAIA-1-2014-2015: Research Infrastructures grant agreement #654024 *SoBigData: Social Mining & Big Data Ecosystem* and from the MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca) and Regione Toscana (Tuscany, Italy) funding the *SmartNews: Social sensing for Breaking News* project: PAR-FAS 2007-2013.

References

Allan, S. 2009. *Citizen journalism: Global perspectives*.

Avvenuti, M.; Del Vigna, F.; Cresci, S.; Marchetti, A.; and Tesconi, M. 2015. Pulling information from social media in the aftermath of unpredictable disasters. In *ICT-DM'15*. IEEE.

Avvenuti, M.; Cresci, S.; Del Vigna, F.; and Tesconi, M. 2016. Impromptu crisis mapping to prioritize emergency response. *Computer* 49(5):28–37.

Avvenuti, M.; Bellomo, S.; Cresci, S.; La Polla, M. N.; and Tesconi, M. 2017. Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In *WWW'17 Companion*. ACM.

Avvenuti, M.; Cresci, S.; Nizzoli, L.; and Tesconi, M. 2018. GSP (Geo-Semantic-Parsing): Geoparsing and Geotagging with Machine Learning on top of Linked Data. In *ESWC'18*.

Cimino, A., and Dell'Orletta, F. 2016. Building the state-of-the-art in POS tagging of italian tweets. In *EVALITA'16*.

Cresci, S.; Cimino, A.; Dell'Orletta, F.; and Tesconi, M. 2015. Crisis mapping during natural disasters via text analysis of social media messages. In *WISE'15*. Springer.

de Lira, V. M.; Macdonald, C.; Ounis, I.; Perego, R.; Renso, C.; and Times, V. C. 2017. Exploring social media for event attendance. In *ASONAM'17*. ACM.

Diakopoulos, N.; De Choudhury, M.; and Naaman, M. 2012. Finding and assessing social media information sources in the context of journalism. In *SIGCHI'12*. ACM.

Fang, R.; Nourbakhsh, A.; Liu, X.; Shah, S.; and Li, Q. 2016. Witness identification in twitter. In *EMNLP'16*.

Morstatter, F.; Lubold, N.; Pon-Barry, H.; Pfeffer, J.; and Liu, H. 2014. Finding eyewitness tweets during crises. *ACL'14 Workshops*.

Murphy, J.; Link, M. W.; Childs, J. H.; Tesfaye, C. L.; Dean, E.; Stern, M.; Pasek, J.; Cohen, J.; Callegaro, M.; and Harwood, P. 2014. Social media in public opinion research executive summary of the AAPOR Task Force on emerging technologies in public opinion research. *Public Opinion Quarterly* 78(4):788–794.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *NAACL-HLT'10*.

Starbird, K.; Muzny, G.; and Palen, L. 2012. Learning from the crowd: collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In *ISCRAM'12*.

Truelove, M.; Vasardani, M.; and Winter, S. 2015. Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal* 80(3):339–359.