iFACT: An Interactive Framework to Assess Claims from Tweets

Wee Yong Lim
National University of Singapore
a0109697@u.nus.edu

Mong Li Lee
National University of Singapore
leeml@comp.nus.edu.sg

Wynne Hsu
National University of Singapore
whsu@comp.nus.edu.sg

ABSTRACT

Posts by users on microblogs such as Twitter provide diverse real-time updates to major events. Unfortunately, not all the information are credible. Previous works that assess the credibility of information in Twitter have focused on extracting features from the Tweets. In this work, we present an interactive framework called iFACT for assessing the credibility of claims from tweets. The proposed framework collects independent evidence from web search results (WSR) and identify the dependencies between claims. It utilizes features from the search results to determine the probabilities that a claim is credible, not credible or inconclusive. Finally, the dependencies between claims are used to adjust the likelihood estimates of a claim being credible, not credible or inconclusive. iFACT allows users to be engaged in the credibility assessment process by providing feedback as to whether the web search results are relevant, support or contradict a claim. Experiment results on multiple real world datasets demonstrate the effectiveness of WSR features and its ability to generalize to claims of new events. Case studies show the usefulness of claim dependencies and how the proposed approach can give explanation to the credibility assessment process.

1 INTRODUCTION

The propagation of misinformation on social media is an increasing concern especially if such misinformation is able to negatively or unfairly affect the public response during an event. Different from mainstream media, the lack of editorial checks on user posts in social media platforms such as Twitter allows for malicious users to exploit the platform to post misinformation and unsuspecting users to propagate the misinformation.

Existing work to assess the credibility of information on social media has focused on discriminating the veracity of individual tweets [9] or tweet clusters [4, 14, 15, 19, 21]. Various features such as tweet content, hashtags, user characteristics and propagation patterns have been utilized to train classifiers that assign a credible or not-credible label to each tweet or tweet cluster.

The work in [12] observes that a tweet may contain multiple pieces of information, each of which may have a different level of credibility and assessing the credibility of information at the tweet level is too coarse-grained to be effective. They introduce the notion of claims based on subject-predicate tuples, and propose a framework to identify claims from a corpus of tweets related to some major event. For example, the tweet *"Data from EgyptAir Flight MS804 indicates in-flight fire; wreckage and remains found in Mediterranean ..."* has two claims, one on in-flight fire and the other on the location of wreckage.

In this work, we focus on assessing the credibility of information at the claim level. We observe that a major event has many claims that are *related* in some way. Consider the following two claims

"Leaked data indicate Egyptair flight MS804 fire crash"

"Egyptair flight MS804 set ablaze minute crash".

If we can verify that the first claim is credible, then we can increase the likelihood that the second claim is credible. Conversely, if we can establish that the claim *"Egyptair plane crash into sea"* is true, then it follows that the claim *"MS804 lands in Cairo as scheduled"* is likely to be false. Clearly, capturing such dependencies among claims will facilitate the verification process.

Further, existing techniques that predict the credibility of information does little to explain *why* they are classified as such and users often find it difficult to accept the predictions from this "black box" approach. Research that examines how users perceive the credibility and accuracy of online information show that users are motivated to be engaged in the evaluation process to be certain of information crediblity [17]. Hence, efforts are needed to present to users independent evidence that supports or contradicts the credibility assessment.

To this end, we present iFACT, an interactive Framework for Assessing Claims from Tweets. This framework will extract claims from tweets pertaining to a major event or crisis, and find dependencies among the claims. For each claim, iFACT will collect evidence from web search results and estimate the likelihood of a claim being credible, not credible or inconclusive.

The proposed framework is adaptable to new evidence as events unfold. This is important as the credibility of information may evolve over time with new evidence, e.g., initial speculations on the cause of a newsbreaking event. Users can provide feedback on the web search results of a claim and iFACT will automatically re-assess the credibility of the claim.

The contributions of this work are:

(1) Propose the use of web search results as independent evidence to authenticate the credibility of claims in tweets and address the issue that credibility assessment evolves over time as new evidence surfaced.
(2) Capture the dependencies among related claims and utilize these dependencies as additional evidence to substantiate the credibility of claims.
(3) Design an interactive framework that involves the user in the verification of the evidence, thus assuring user of the accuracy of claim status.

Experiment results on multiple tweet datasets from real world events demonstrate the effectiveness of the proposed approach.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132995

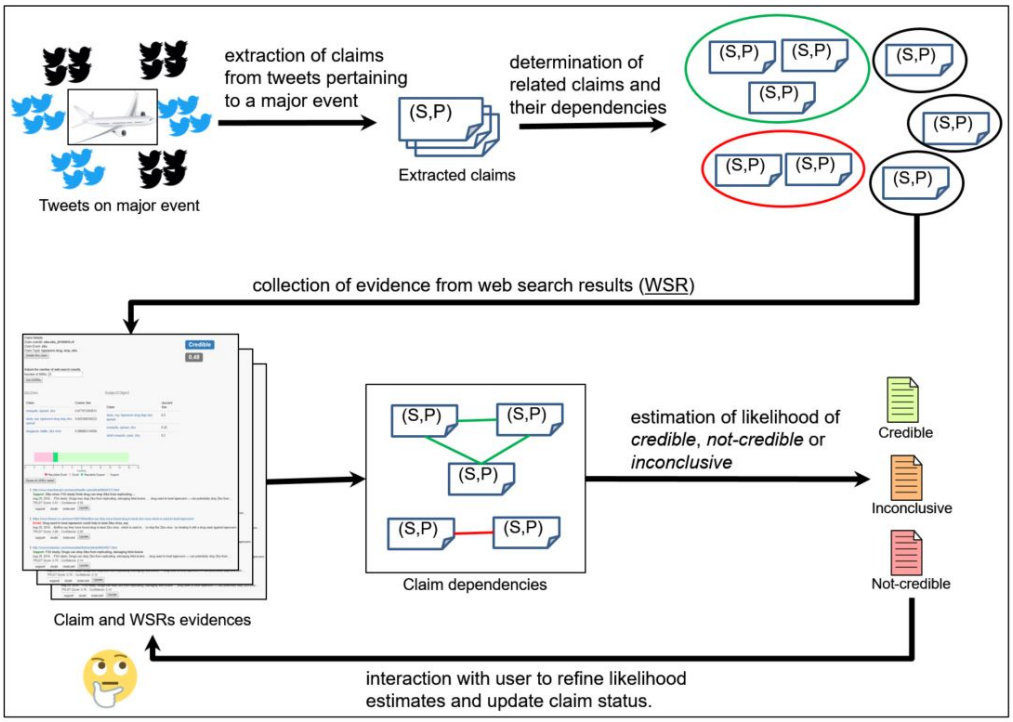


Figure 1: iFACT Framework

2 PROPOSED FRAMEWORK

Figure 1 gives an overview of the proposed credibility assessment framework. iFACT will extract claims from tweets pertaining to a major event and determine the related claims and their dependencies. These claims are represented in the form of Subject-Predicate (S, P) tuples.

Users can browse the claims and query the credibility of a claim. iFACT will perform a web search to collect evidence for the selected claim. These Web Search Results (WSR) are automatically classified as support or debunk the claim, and are ranked according to their source authority. Users can examine the WSRs and may disagree that certain search results support/debunk the claim, in which case, they can change the labels. The system will estimate the likelihood of the claim being credible, not credible or inconclusive based on the features of its WSR, as well as the claim dependencies. In addition, users can create a watchlist of claims and iFACT will periodically collect new web search results to determine if there is any change in the status, and alert the users.

The following subsections gives the details of the main steps.

2.1 Extract Claims and their Dependencies

The extraction of claims from tweets is similar to [12] which utilizes an Open Information Extraction tool called ClausIE [5] to extract a set of relation triples from tweet content. A relation triple is denoted by (E_1, R, E_2) where E_1 and E_2 are entities and R is the relationship between these entities. In order to facilitate subsequent clustering and identification of claims, each triple is mapped to a subject-predicate tuple (S, P) where $S = E_1 \cup E_2$ and $P = R$.

Table 1 shows several tweets and their corresponding Open IE triples and subject-predicate tuples. Table 2 gives the clustering of the subject-predicate tuples. Each cluster corresponds to a claim, and is represented by the union of the respective S and P terms of the subject-predicate tuples in the cluster.

Next, we determine the dependencies among the claims. There are two types of dependencies, namely, *direct* and *inverse*.

Direct dependency. A pair of claims (c_1, c_2) is considered to have a direct dependency, denoted by $direct(c_1, c_2)$, if c_1 is likely to be credible when c_2 is deemed credible, and vice versa.

Inverse dependency. A pair of claims (c_1, c_2) has an inverse dependency, denoted by $inverse(c_1, c_2)$, if c_1 is likely to be credible when c_2 is deemed not credible, and vice versa. In other words, it is unlikely for c_1 and c_2 to be both credible (or not credible) at the same time.

For example, claims 1 and 2 in Table 2 have a direct dependency while claims 3 and 4 have an inverse dependency.

We design an algorithm to suggest the dependencies between claims based on their subjects and predicates. If two claims refer to the same entity and their predicates are synonyms, then we say that they have a direct dependency. On the other hand, if their predicates are antonyms, then they have an inverse dependency.

The work in [12] uses Jaccard similarity to compare two word sets. However, this measure does not tell us if the word sets of two subjects refer to the same entity. For example, claims 3 and 4 in Table 2 have no common word in their subjects *flight MS804* and *plane*, i.e., their Jaccard similarity is 0, yet it is clear that they refer to the same entity.

Tweet Content	Open IE triples	SP tuples
EgyptAir Flight MS804 on fire before crash...	(EgyptAir flight MS804, on fire, crash)	({EgyptAir, flight, MS804, crash}, {on fire})
Flight MS804 on fire in mid-air...	(flight MS804, on fire, mid-air)	({flight MS804, mid-air}, {on fire})
MS804 set ablaze before crash...	(MS804, set ablaze, crash)	({MS804, crash}, {set ablaze})
Unconfirmed reports suggest flight MS804 landed. Not sure how true...	(flight MS804, land)	({flight MS804 }, land})
If the plane had crashed, wouldn't there be SOME debris floating in the water?	(plane, crash) (debris, float, water)	({plane}, {crash}) ({debris, water}, {float})

Table 1: Sample tweets, extracted Open IE triples and subject-predicate tuples.

	Cluster of SP tuples	Claim identified	Description
1	{ ({EgyptAir, flight, MS804, crash}, {on fire}), ({flight MS804, mid-air}, {on fire}) }	({EgyptAir, flight, MS804, crash, mid-air}, {on fire})	EgyptAir flight MS804 on fire
2	{ ({MS804, crash}, {set ablaze}) }	({MS804, crash}, {set ablaze})	EgyptAir flight MS804 set ablaze before crash
3	{ ({flight MS804}, {land}) }	({flight MS804}, {land})	flight MS804 land
4	{ ({plane}, {crash}) }	({plane}, {crash})	plane crash
5	{ ({debris, water}, {float}) }	({debris, water}, {float})	debris float in water

Table 2: Claims identified from Table 1

We address this issue by using the English news articles from Associated Press [10] to train a word embedding model and generate the vector representation of subjects/predicates in the claims. Then we compute the cosine similarity between the vector representations of word sets d_i, d_j as shown in Equation 1.

$$sim(d_i, d_j) = cosineSim(doc2vec(d_i), doc2vec(d_j)) \quad (1)$$

where *doc2vec* refers to the vector representation proposed in [11].

Next, looking at the predicates of the claims *"flight MS804 land"* and *"plane crash"*, we can have the dependency *inverse (("flight MS804 land", "plane crash")* since a plane that has landed implies that it has not crashed.

Algorithm 1 gives the details to find direct and inverse dependencies between claims. It takes as input a set of claims C, a selected claim c and output a set of related claims $\widetilde{C}$ and the pairwise relations $\mathcal{R}$ between c and the related claims. We find the set of claims M whose subjects are similar to that of the input claim c (Line 3). For each such claim c' whose subject is similar to c, we compute the cosine similarity of their predicates. If the similarity of their predicates exceeds the threshold α, then we consider c and c' to have a direct dependency. We add the relation $direct(c, c')$ to $\mathcal{R}$ and c' to $\widetilde{C}$ (Lines 5-8). On the other hand, if the predicates of c and c' are antonyms, we compute the cosine similarity of the predicate of c and the antonyms of the predicate of c', denoted by $antonym(pred(c'))$. If the similarity exceeds the threshold α, then we consider c and c' to have an inverse dependency. We add the relation $inverse(c, c')$ to $\mathcal{R}$ and c' to $\widetilde{C}$ (Lines 9-12).

Algorithm 1 *GetRelatedClaims*

Input: selected claim c
Input: set of claims from the same event $C = \{c_1, c_2, \cdots\}$
Input: similarity threshold α
Output: set of related claims $\widetilde{C}$ and pairwise relations $\mathcal{R}$

1: $\widetilde{C} = \emptyset$
2: $\mathcal{R} = \emptyset$
3: $M = \{c' \mid c' \in C \wedge sim(subj(c), subj(c')) \geq \alpha\}$
4: **for** each claim $c' \in M$ **do**
5: **if** $sim(pred(c), pred(c')) \geq \alpha$ **then**
6: $\mathcal{R} \leftarrow \mathcal{R} \cup \{direct(c, c')\}$
7: $\widetilde{C} \leftarrow \widetilde{C} \cup \{c'\}$
8: **else**
9: **if** $sim(pred(c), antonym(pred(c'))) \geq \alpha$ **then**
10: $\mathcal{R} \leftarrow \mathcal{R} \cup \{inverse(c, c')\}$
11: $\widetilde{C} \leftarrow \widetilde{C} \cup \{c'\}$
12: **end if**
13: **end if**
14: **end for**

2.2 Collect Evidence from Web Search

After extracting the claims for each event and determining the related claims, iFACT allows the user to browse and select the claim s/he is interested in. Then we carry out a web search for the selected claim as well as all its related claims. This web search is performed using keywords from the event and the wordsets of the claim. For example, the claim ({*body, part, sea*}, {*retrieve*}) regarding the retrieval of body parts from the sea for the EgyptAir MS804 crash event would have the search query ``MS804 body part retrieve sea''.

Each web search returns a list of results where each Web Search Result (WSR) comprises of a source URL, text title and snippet. We view each WSR as evidence expressing its source's belief regarding the credibility of the selected claim. We evaluate if a WSR's title and snippet content is relevant to the claim, whether it supports or doubts the claim, and whether the WSR is from a reputable or trustworthy source.

allege	apprehension	baffle	bluff	bogus	bullshit
conspiracy	contrary	controversy	cryptic	debate	debunk
deceit	deceive	deception	defraud	deny	disbelief
dishonest	dispute	doubt	dubious	erroneous	extraordinary
fabricate	fake	fallacy	false	fiction	fishy
gossip	hearsay	hoax	illusion	imaginary	inaccurate
incorrect	inexplicable	intrigue	lies	misguided	misinform
mislead	mistaken	mysterious	mystical	myth	ostensibly
perplex	phony	purport	puzzle	question	rumour
scam	seemingly	sham	smear	speculate	spurious
supposed	suspicion	theory	trick	unbelievable	wrong

Table 3: Set of doubt words used to determine if a WSR debunks a claim.

Feature	Description
Results	Number of results for the search query as estimated by the web search engine.
Doubt	Fraction of top-k WSRs that contains some doubt word or a question mark '?'.
Reputable	Fraction of top-k WSRs from reputable sources.
Reputable doubt	Fraction of top-k WSRs from reputable sources that express doubt.
Trust score	Average trust score of the top-k WSR sources.
Doubt score	Average trust score of the top-k WSRs sources that express doubt.

Table 4: Features extracted from WSRs pertaining to a claim.

In order to determine whether a WSR expresses *doubt* on the claim, we collect a set of rumours from fact-checking websites such as www.snopes.com and www.truthorfiction.com. We manually identify words that express doubts on the credibility of the rumours in this set. Table 3 shows the list of doubt words obtained together with their synonyms. We say that a WSR does not support the claim if it contains some doubt words in Table 3. Otherwise, we assume that the WSR provides evidence to support the claim.

The *source reputation* of each WSR can indicate if the evidence is reliable. A source is reputable if its domain is in a pre-defined list of reputable news sites such as bbc.co.uk, reuters.com, etc. We also compute a *trust score* for each source using information from the Web of Trust[1]. The Web of Trust provides the *trustworthiness* and *confidence* scores of website through crowdsource reviews. Our *trust score* is given by *trustworthiness* * *confidence*.

2.3 Estimate Likelihood of Claim Credibility

After collecting evidence from web search results, the next step is to estimate the likelihood of the credibility of claims.

We pre-train a classifier called $ClaimClassifier$ using the WSRs of claims in a training set. The features used by the classifier are listed in Table 4. The classifier outputs the probabilities of a claim being in the class `credible`, `not-credible` or `inconclusive`, denoted as P_{CR}, P_{NC}, P_{IC} respectively.

Given a user's selected claim and its related claims, we first call $ClaimClassifier$ to obtain their probabilities of being in the various classes. Then we use Probabilistic Soft Logic (PSL) [1] to take into account the dependencies among the claims and adjust the classifer probabilities. This is in contrast to prior work that consider each claim independently.

PSL utilizes first order logic rules for probabilistic reasoning with relational structures, in this case, the direct and inverse dependencies among claims. We express these dependencies using first order logic rules as follows:

$$evidence(c1, class) \Rightarrow likely(c1, class)$$
$$direct(c1, c2) \wedge evidence(c1, class) \Rightarrow likely(c2, class)$$
$$inverse(c1, c2) \wedge evidence(c1, class) \Rightarrow \neg likely(c2, class)$$

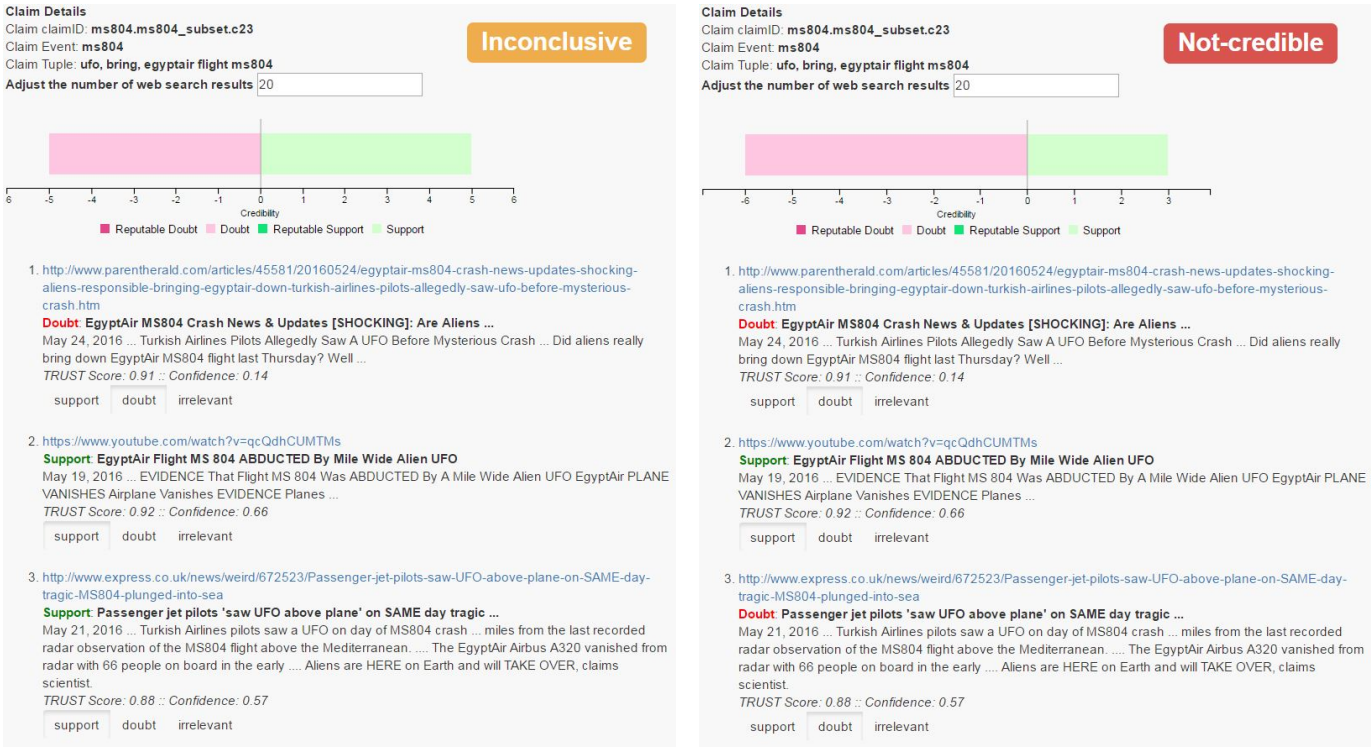
If the evidence shows that a claim $c1$ belongs to *class*, then $c1$ is labeled *likely* to belong to *class*. If $c1$ belongs to *class* and has a direct dependency with another claim $c2$, then it is likely for $c2$ to belong to *class* too. However, if $c1$ belongs to *class* and has an inverse dependency with $c2$, then it is unlikely for $c2$ to belong to *class*.

For a given claim, the logic rules are grounded by instantiating all its related claims and their claim status, yielding a set of ground rules. In the PSL framework, atoms in these ground rules take on soft truth values in the interval [0,1]. In order to handle these continuous truth values, the conjunction and negation logic operators are relaxed using *Łukasiewicz* t-norm operators as shown below:

$$a \hat{\wedge} b = max\{0, a + b - 1\}$$
$$\hat{\neg} a = 1 - a$$

A ground rule is satisfied if the truth value of the head of the rule is greater than or equal to the truth value of the body of the rule. We use the Most Probable Explanation (MPE) inference algorithm [1] to obtain the assignment of truth values for the claims that maximises the number of ground rules satisfied.

[1] https://www.mywot.com/



(a) Before user feedback **(b) After user feedback**

Figure 2: Change of claim status after user modifies a previously 'Support' WSR to 'Doubt'.

Algorithm 2 shows the details of the likelihood estimation of the credibility of claims. The input is a claim c, and the output is the status (CR, NC, IC) of the claim c. The algorithm first obtains the set of related claims $\widetilde{C}$ together with the set of pairwise relations R by calling Algorithm 1 (Line 1). For each claim c' in $\{c\} \bigcup \widetilde{C}$, we retrieve its WSRs and obtain the features listed in Table 4. Then we call $ClaimClassifier$ to get its probabilities $P_{\texttt{CR}}$, $P_{\texttt{NC}}$, $P_{\texttt{IC}}$ (Lines 2-4). For each $class \in \{\texttt{CR}, \texttt{NC}, \texttt{IC}\}$, we use the corresponding probability $P_{class}(c')$ as the soft truth values of $evidence(c', class)$ for a claim c'. PSL is then applied to infer the likelihood of the selected claim c (Lines 7-14). We return the status of c which is given by $argmax_{class}\ likely(c, class)$ (Lines 15-16).

Algorithm 2 *GetClaimStatus*

Input: selected claim c of event e; threshold α
Output: *status* of c

1: call Algorithm 1 to get $\widetilde{C}$ and R
2: **for** each $c' \in \{c\} \bigcup \widetilde{C}$ **do**
3: WSR = WebSearch(e, c')
4: call $ClaimClassifier(WSR)$ to get
5: $P_{\texttt{CR}}(c')$, $P_{\texttt{NC}}(c')$, $P_{\texttt{IC}}(c')$
6: **end for**
7: **for** each $class \in \{\texttt{CR}, \texttt{NC}, \texttt{IC}\}$ **do**
8: **for** each $c' \in \{c\} \bigcup \widetilde{C}$ **do**
9: set $evidence(c', class) \leftarrow P_{class}(c')$
10: let $R_{c'} \subset R$ be the set of relations involving c'
11: instantiate the rules in PSL with $R_{c'}$
12: **end for**
13: infer $likely(c, class)$
14: **end for**
15: $status = argmax_{class}\ likely(c, class)$
16: return *status* of c

2.4 User Interaction

iFACT provides an interactive platform to engage the user in the credibility evaluation process. Figure 2(a) shows the partial web search results of a user selected claim regarding the purported UFO downing of flight MS804 *"ufo bring EgyptAir flight MS804"*. This claim is determined to be `inconclusive` with equal number of WSRs supporting/doubting it, and the sources are not from reputable sites.

A user can easily review the WSR evidence of the selected claim, and toggle whether a WSR is relevant, and if relevant, whether the WSR supports or doubts the claim. The user can also review the claims that are related to the selected claim and the dependencies among them. He can change the dependency type from *direct* to *inverse* and vice versa, as well as specify additional dependencies among claims. Then iFACT will call Algorithm 2 to re-compute the likelihood estimate of a claim and update the claim status. Figure 2(b) shows that the claim status is updated to `not-credible` after the user changes a WSR from *support* to *doubt*.

Dataset	Event	Duration	#tweets	# credible	# not credible	# inconclusive
MH370	Disappearance of Malaysia Airline, likely over Indian Ocean, on 8 March 2014	8-22 Mar 2014	21,102	12	12	1
Panama Papers	Leaked documents from a Panama-based law firm to selected press (3 Apr 2016) and subsequent limited release to public (9 May 2016)	4 Apr - 17 May 2016	11,308	20	0	4
MS804	Crashing of EgyptAir plane over Mediterranean Sea on 19 May 2016	19 May - 4 Jun 2016	7,525	21	3	6
Jongnam	The assassination of Kim Jongnam, step brother of the N. Korean leader on 13 February 2017	14-24 Feb 2017	5,500	18	2	0
Brussels	Multiple bombings in Brussels on 22 March 2016	23 Mar - 4 Apr 2017	9,205	31	2	0

Table 5: Characteristics of tweet datasets.

Users can select a set of claims to monitor and iFACT will periodically initiate a web search to retrieve the most recent set of WSRs which may contain new evidence for a claim. Users are alerted when the status of a claim changes as a result of the new evidence.

Credibility assessment of claims in tweets is a challenging task. The proposed framework recognises the need to put humans in the loop and achieves this by suitably introducing points where users can review evidence and give feedback.

3 EXPERIMENT EVALUATION

We carry out experiments on several real world datasets to demonstrate the effectiveness of the proposed framework. We implement the proposed algorithms in Python, and carry out experiments on a 2.6 GHz CPU with 16 GB RAM running on Ubuntu 14.04.

3.1 Datasets

We curated 5 datasets of tweets related to major news events such as the MH370 and MS804 plane disasters, Panama Papers leaks, bombings in Brussels, and the assassination of Kim Jongnam. These tweets are crawled within days of these events using the respective keyword for the dataset name.

The first two events are major aviation disasters with no survivors, garnering widespread interest on the cause of the crashes. Perceived mishandling in the public dissemination of information where initial conflicts between information sources led to unfortunate public confusion on the veracity of information. For example, conflicting statements by different authorities made it initially unclear if *"flight MH370 deviated from its course"* or if *"flight MS804 swerved before it crashed"*.

The Panama Papers leak refers to the disclosure of stolen documents from a Panama based law firm involving private financial information on wealthy or powerful clients across the world. The prominent nature of the people involved and allegations of tax evasion unsurprisingly lead to tweets from users who wish to share information pertaining to the leaks. However, the veracity of some claims, e.g., claim concerning a popular Nigerian pastor *"TB Joshua deny owning company"*, can be difficult to ascertain. Such claims are thus annotated as inconclusive in this work.

The Brussels and Jongnam datasets are on highly visible, publicly traumatising events where there exist public witnesses or video recordings of the perpetrators of the terror act or assassination attempt. As such, these two datasets do not have inconclusive claims. Nonetheless, tweets from the aftermath of these events still contain some rumours such as *"Germany Chancellor Angela Merkel took a selfie with one of the terrorists involved in the Brussels bombings"*.

We use the ClaimFinder framework in [12] to identify the claims in each event. Non-newsworthy claims are discarded while newsworthy claims are manually annotated as either credible, not credible or inconclusive. Table 5 gives the characteristics of the datasets.

3.2 Effectiveness of WSR Features

In this section, we investigate the effectiveness of using WSR as features in iFACT for the credibility assessment of claims.

We first compare the proposed WSR features in Table 4 with the tweet-based features [4] in Table 6 on the first 3 datasets in Table 5. Table 7 shows that cross validation results using various ensemble classifiers: Adaboost [6], Gradient Boosting Tree [7], Random Forest [3] and Extremely Randomized Trees [8]. Overall, we observe that the WSR features leads to higher precision, recall and F1 measures. For the credible class, we see that using WSR features achieves the highest precision of 0.835 as compared to 0.728 achieved using tweet-based features. For the not-credible class, using WSR features outperform tweet-based feaures with a precision of 0.75 versus 0.59. For the inconclusive class, the precision is 0.73 using WSR features versus 0.67 using tweet-based features.

Next, we examine the effect of individual WSR feature on the precision, recall and F1 measure. We train a Random Forest classifier by omitting a WSR feature at a time from the training set. Table 8 shows the cross validation results. We observe that omitting either the feature *Doubt* (number of WSRs that contain doubt words) or *Reputable* (number of WSR from reputable sources) leads to the largest drop in the metrics for both credible and not-credible classes.

Finally, we demonstrate that the classifiers obtained using WSR features can be generalized to claims from new events. We use the first 3 datasets in Table 5, which are the older events, to train the classifiers and test them on the last 2 datasets. Table 9 shows the results. We observe that the precision, recall and F1 measures of the classifiers trained using WSR features outperform those trained using tweet features, indicating that WSR features can better generalize to new events.

Feature(s)	Description / Justification
Words/Characters	average number of words, characters and unique characters in tweets
Question mark	fraction of tweets containing '?'
Exclamation mark	fraction of tweets containing '!'
Hashtag	fraction of tweets containing hashtag(s)
Wh-words	fraction of tweets containing 'what', 'why', 'where', 'when' or 'how'
Swear word	fraction of tweets containing 1 or more pre-specified swear words
Pronoun	fraction of tweets containing pronoun(s)
Smiley	fraction of tweets containing smile/frown smiley(s)
Unique user mention	fraction of user mentions that are unique
Changes in number of tweets	maximum increase/decrease in number of tweets between days
Users	number of users normalized against number of tweets
Verified	fraction of verified users
Location	fraction of users with location information
Description	fraction of users with description information
Profile image	fraction of users with profile image
Geo enabled	fraction of users with geo enabled
Registration year	average registration year of user accounts
Followers/Friends	average followers and friends count

Table 6: Features from tweets and their user accounts.

		Credible			Not Credible			Inconclusive		
Feature	**Classifier**	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**
Tweet	Adaboost	0.723	0.691	0.706	0.417	0.333	0.370	0.167	0.278	0.208
	Gradient Boosting Tree	0.728	0.709	0.718	0.493	0.467	0.479	0.667	0.250	0.364
	Random Forest	0.712	0.808	0.757	0.328	0.333	0.331	0.500	0.167	0.250
	Extremely Randomized Trees	0.728	0.709	0.719	0.593	0.533	0.561	0.500	0.250	0.333
WSR	AdaBoost	**0.835**	0.866	0.850	0.476	0.533	0.503	0.556	0.278	0.370
	Gradient Boosting Tree	0.780	0.924	0.846	0.700	0.400	0.509	**0.733**	**0.278**	**0.403**
	Random Forest	0.785	**0.961**	**0.864**	**0.750**	**0.533**	**0.623**	0.333	0.083	0.133
	Extremely Randomized Trees	0.794	0.846	0.820	0.481	0.467	0.474	0.400	0.194	0.262

Table 7: WSR features versus tweet features.

	Credible			Not Credible		
Feature Omitted	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**
Results	0.807	0.942	0.870	0.550	0.467	0.505
Doubt	**0.719**	0.867	0.786	0.583	**0.333**	**0.424**
Reputable	0.742	**0.809**	**0.774**	**0.500**	0.400	0.444
Reputable doubt	0.771	0.865	0.815	0.514	0.533	0.523
Trust score	0.812	0.827	0.820	0.621	0.667	0.643
Doubt score	0.772	0.845	0.807	0.565	0.533	0.549

Table 8: Effect of omitted WSR features on the precision, recall and F1 of Random Forest classifier.

3.3 Effect of Claim Dependencies

In this section, we highlight two claims from the MH370 dataset whose status are changed as a result of analyzing the dependencies between claims. Consider claim c6 *"mh370 change course"* in Table 10. This claim has direct dependencies with claims c14 *"plane travel hour off course"* and c19 *"mh370 turn back"*. This makes sense because if the plane has travelled off-course implies that it is likely to have changed course or turned back. Thus, if any of these claims is determined to be credible, it will increase the likelihood of the other claims being credible.

Based on the WSR evidence, $ClaimClassifier$ gives the highest probability for c6 being not credible, while the highest probabilities are given to c14 and c19 being credible. Given the dependencies $direct(c6, c14)$ and $direct(c6, c19)$, the likelihood estimate of c6 being credible is adjusted higher and consequently the status of c6 is correctly determined as credible.

		Credible			Not Credible		
Feature	**Classifier**	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**
Tweet	AdaBoost	0.918	0.918	0.918	0.000	0.000	0.000
	Gradient Boosting Tree	0.932	0.837	0.882	0.000	0.000	0.000
	Random Forest	0.918	0.918	0.918	0.000	0.000	0.000
	Extremely Randomized Trees	0.922	0.959	0.940	0.000	0.000	0.000
WSR	AdaBoost	0.919	0.694	0.791	0.125	0.250	0.167
	Gradient Boosting Tree	0.936	0.898	0.917	1.000	0.250	0.400
	Random Forest	0.960	0.980	0.970	1.000	0.500	0.667
	Extremely Randomized Trees	**0.961**	**1.000**	**0.980**	**1.000**	**0.500**	**0.667**

Table 9: Generalization of classifiers to new events.

Claim	Dependency	ClaimClassifier	After PSL
c6: mh370 change course		$P_{\mathbf{NC}}$ highest	status becomes **CR**
c14: plane travel hour off course	direct(c6,c14)	P_{CR} highest	CR
c19: mh370 turn back	direct(c6,c14)	P_{CR} highest	CR

Table 10: Effect of direct dependency on the status of claim c6

Claim	Dependency	ClaimClassifier	After PSL
c2: mh370 land china		$P_{\mathbf{CR}}$ highest	status becomes **NC**
c7: plane crash water	inverse(c2,c7)	P_{CR} highest	CR
c26 mh370 land nanning	direct(c2,c26)	P_{NC} highest	NC

Table 11: Effect of inverse dependency on the status of claim c2

On the other hand, claim c2 *"mh370 land China"* in Table 11 has the highest probability of being credible. This claim has an inverse dependency with c7 *"plane crash water"* because if the plane lands in China, it cannot crash into the water. Further, suppose there is another claim c26 *"mh370 land nanning"* that share the same predicate "land" and has direct dependency with claim c2, as both claims imply the plane has landed. After PSL processes these dependencies, the status of c2 is correctly changed to not credible.

3.4 Explanability of iFACT

In this section, we demonstrate how the proposed framework helps provide explanation for the credibility assessment of claims.

The work in [4] reported multiple clusters of newsworthy tweets corresponding to events, each characterised by a set of keywords. Specifically, each cluster of tweets is annotated via crowd sourced effort to be either credible or not credible based on 10 random sample tweets in the cluster. On closer examination of these clusters, we found some clusters that have been annotated as not credible, which appears to be questionable.

We use iFACT to retrieve independent evidence from the web. Table 12 shows the top 5 WSRs for 3 clusters TM1293, TM1338 and TM1456. These WSRs are published by their sources in the same year 2010 as the tweet clusters. The sources are major news sites (e.g. nytimes.com, abcnews.go.com, telegraph.co.uk, etc.) with high *trust* scores and they consistently agree with the keywords in the tweet cluster. Based on the evidence, the user can be convinced that the earlier annotations may not be correct.

The credibility of a claim can evolve across time too. Consider the claim *"Favre not return Vikings"* extracted from tweet cluster TM2210 that has been labeled as not credible. This claim is about an American athlete Favre not returning to his football club Vikings. When we try to understand why the claim is labeled as such, we search the web for supporting evidence.

Figure 13 shows the top 4 web search results obtained over two time periods. WSRs from the earlier time period (3-9 Aug 2010) contain initial speculations that Favre would not be playing for the club. However, WSRs obtained one week later (17-23 Aug 2010) reveal that Favre has confirmed his return to Vikings after all. This demonstrates that iFACTcan seamlessly handle new evidence to allow users to track the credibility of a claim as the event unfolds.

4 RELATED WORK

Credibility assessment on user generated content or trustworthiness of users has been extensively studied for microblogs [4, 9, 15, 16, 21]. A supervised learning approach is typically adopted to learn features that can help in discriminating the veracity of individual tweets, tweet clusters or users.

An early work [16] presented a manual analysis of the veracity of tweets during the major Chilean earthquake of 2010. A total of 14 claims were extracted from tweets and a crowd sourcing platform was used to allow users to question, deny or support these claims. The results suggest that the community can act as a collaborative truth filter of information as users tend to question rumors more than news. This led to the work in [18] which introduced a crowd

Tweet Cluster from [4]	Top-5 WSR from iFACT
TM1293 *"1 trillion dollar worth untapped mineral find in Afghanista"* Crowdsource: `Not Credible` iFACT: `Credible`	http://www.nytimes.com... :: U.S. Identifies Vast Mineral Riches in Afghanistan - The New York ... :: Jun 13, 2010... The nearly $1 trillion in untapped deposits are enough to fundamentally ...
	http://www.telegraph.co.uk... :: Afghanistan claims mineral wealth is worth $3trillion - Telegraph :: Jun 17, 2010 ... The Afghan government claims its untapped mineral wealth could be worth Âč2 ... US geologists find $1trillion of mineral reserves in Afghanistan ...
	http://www.popsci.com... :: U.S. Geologists Uncover Staggering $1 Trillion Cache of Unmined ... :: Jun 14, 2010 ... estimated $1 trillion worth of untapped geological resources there, ...
	http://www.reuters.com... :: Afghan mineral wealth could top $1 trillion: Pentagon | Reuters :: Jun 14, 2010 ... an indication that even the trillion dollar figure underestimates what the ...
	http://www.independent.co.uk... :: Afghanistan's untapped minerals 'worth $3 trillion' | The Independent :: Jun 17, 2010 ... Afghanistan's untapped mineral wealth is worth at least $3 trillion - triple ...
TM1338 *"Lee Gardner execute early Friday"* Crowdsource: `Not Credible` iFACT: `Credible`	http://abcnews.go.com... :: Convicted Killer Ronnie Lee Gardner Is Executed in Utah - ABC News :: Jun 18, 2010 ... When a prison official opened a curtain to reveal the death chamber to witnesses early Friday, convicted killer Ronnie Lee Gardner was already ...
	http://www.ksl.com... :: Ronnie Lee Gardner executed by firing squad | KSL.com :: Jun 18, 2010 ... Condemned inmate Ronnie Lee Gardner was executed by firing squad early Friday morning ...
	http://www.cnn.com... :: Killer executed by Utah firing squad - CNN.com :: Jun 18, 2010 ... Draper, Utah (CNN) – Convicted killer Ronnie Lee Gardner was executed early Friday by firing squad ...
	http://www.nbcnews.com... :: Death-row inmate dies in barrage of bullets - US news - Crime ... :: Jun 18, 2010 ... A firing squad executed convicted killer Ronnie Lee Gardner early on Friday ...
	https://www.youtube.com... :: Ronnie Lee Gardner executed by firing squad - YouTube :: Jun 18, 2010 ... Condemned inmate Ronnie Lee Gardner was executed by firing squad early Friday morning ...
TM1456 *"Medvedev meet Steve Job"* Crowdsource: `Not Credible` iFACT: `Credible`	https://themoscowtimes.com... :: Apple CEO Advises Medvedev to Change Russian Mentality :: Jun 24, 2010 ... President Dmitry Medvedev and Apple CEO Steve Jobs checking out an iPhone 4 on Wednesday at the company's headquarters in Cupertino, ...
	http://www.geek.com... :: Russian President visits Apple and tweets about iPhone 4 before ... :: Jun 25, 2010 ... The Russian leader visits Apple and tweets an image of Steve Jobs ...
	http://www.rferl.org... :: In Silicon Valley, Medvedev Looks For Investment Possibilities :: Jun 24, 2010 ... Medvedev met with industry leaders including Apple CEO Steve Jobs ...
	http://phys.org... :: Cisco commits $1B in meeting with Russian leader :: Jun 23, 2010 ... where he was expected to meet with CEO Steve Jobs...
	http://en.kremlin.ru... :: Visit to Apple Inc President of Russia :: Jun 23, 2010 ... Dmitry Medvedev met with Apple Inc. CEO Steve Jobs.

Table 12: WSR evidence from iFACT shows why claims are credible.

sourcing platform called Verily. This platform leveraged on reputation points to incentivise participants to provide evidence and evaluate the credibility of claims.

The work in [4] investigated an extensive list of content, user and propagation based features to determine if a given cluster of tweets is *{newsworthy,chat}* and for the newsworthy clusters of tweets, whether the tweets are *{credible,not-credible,unsure}*. The evaluation was carried out using crowd-sourced ground truth labels. Additional features such as hashtags and URLs [19], event location and client program [21], negative/positive sentiments [4] were used to improve the accuracy of the credibility assessment task. A recurrent neural network was trained using tf-idf features from tweets at discrete time windows [15]. However, the trained models cannot be easily generalized to tweets from another event [2].

TM2210: *"Favre not return to Viking"*	
Top WSRs from 3 Aug - 9 Aug 2010	**Top WSRs from 17 Aug - 23 Aug 2010**
http://www.espn.com... :: Reports: Brett Favre tells Minnesota Vikings he will retire - ESPN.com :: Brett Favre has informed his Vikings teammates that he **will not return** for another season in Minnesota ...	www.espn.com... :: Brett Favre: 'I owe it' to Minnesota Vikings to **return again** - ESPN.com :: The Minnesota Vikings say Brett Favre is about to practice with his teammates ...
http://www.nj.com... :: Brett Favre plans to play for Vikings if he is healthy | NJ.com :: Vikings offensive coordinator Darrell Bevell and Favre's agent, Bus Cook, ... ankle was not healing and that he **wasn't going to return** for a second season in Minnesota ...	www.nj.com... :: Lure of another Super Bowl brings Brett Favre **back to Vikings** | NJ.com :: Favre underwent left ankle surgery on May 21 and just a few weeks ago texted several teammates and Vikings officials that he would not return because the ...
http://www.mprnews.org :: Vikings player: Favre tells teammates he's **retiring** | Minnesota Public :: Minnesota Vikings tight end Visanthe Shiancoe says Brett Favre has texted his ...	content.usatoday.com :: Brett Favre arrives in Minnesota; **return to Vikings imminent** - USA Today :: The Vikings have not confirmed the news. Favre's agent, Bus Cook, has not returned messages from USA TODAY ...
http://www.dailymail.co.uk... :: NFL legend Brett Favre stuns Minnesota Vikings by retiring with a ... :: Brett Favre has stunned his Minnesota Vikings team-mates by **retiring** - and ... has been recovering from ankle surgery but was expected to return for a 20th NFL ...	www.nydailynews.com :: Brett Favre returns to Minnesota in private jet; no ... - NY Daily News :: Brett Favre is back in the building. The 40-year-old quarterback **returned to** Minnesota on Tuesday, arriving in a private jet trimmed in the **Vikings'** purple and ...

Table 13: WSR evidence over two time periods.

Both tweet level [9] and cluster level [13, 14] features have been explored for real time rumor detection. Tweets or users are classified as being *supportive*, *negating*, *doubting* or *neutral* towards the event based on heuristics using specified lists of positive and negative words. Using pre-specified positive words are difficult to generalize across different events as these words tend to be diverse and context dependent. Our work differs from these approaches as we seek independent evidence from web search results. Our experiments show that features from WSR can be generalized to new events.

The work in [20] proposed a framework called ClaimEval to determine the credibility of subjective claims such as *"Turkey meat is healthy"* based on relevant webpages and their domains. The claims are assumed to be independent. In contrast, iFACT utilizes the dependencies between claims to obtain a more consistent credibility assessment of claims.

5 CONCLUSION

In this work, we have made a step towards putting humans in the loop for credibility assessment of information. We have presented an interactive framework called iFACT to evaluate the credibility of claims from tweets pertaining to major events. In contrast to prior work that focus on features extracted from tweets for credibility classification in a "black box" environment, our proposed approach is based on independent evidence from web search results and claim dependencies. We have provided a platform for the user to examine if a WSR is irrelevant, support or doubt a claim, and modify the type of dependencies between claims.

Experiment results on multiple datasets demonstrate the effectiveness of WSR features and its ability to generalize to new events. Case studies further confirm the importance of taking claim dependencies into consideration, and how iFACT can provide explanations for users to verify the credibility of claims.

REFERENCES

[1] S. H. Bach, M. Broecheler, B. Huang, and L. Getoor. Hinge-loss markov random fields and probabilistic soft logic. *CoRR*, abs/1505.04406, 2015.
[2] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of computational verification in social multimedia. In *WWW*, 2014.
[3] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.
[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, 2011.
[5] L. Corro and R. Gemulla. Clausie: Clause-based open information extraction. In *WWW*, 2013.
[6] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1), 1997.
[7] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.
[8] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42, 2006.
[9] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*, 2014.
[10] J. H. Lau and T. Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *ACL*, 2016.
[11] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.
[12] W. Y. Lim, M. L. Lee, and W. Hsu. Claimfinder: A framework for identifying claims in microblogs. In *WWW Workshop on Making Sense of Microposts*, 2016.
[13] X. Liu, Q. Li, and A. Nourbakhsh et al. Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. In *CIKM*, 2016.
[14] X. Liu, A. Nourbakhsh, and Q. Li et al. Real-time rumor debunking on twitter. In *CIKM*, 2015.
[15] J. Ma, W. Gao, and P. Mitra et al. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*, 2016.
[16] M. Mendoza, B. Pobletey, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *1st Workshop on Social Media Analytics,*, 2010.
[17] M. J. Metzger and A. J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59, 2013.
[18] A. Popoola, D. Krasnoshtan, A.-P. Toth, V. Naroditskiy, C. Castillo, P. Meier, and I. Rahwan. Information verification during natural disasters. In *WWW*, 2013.
[19] V. Qazvinian, E. Rosengren, and D. Radev et al. Rumor has it: Identifying misinformation in microblogs. In *International Conference on Empirical Methods in Natural Language Processing*, 2011.
[20] M. Samadi, P. Talukda, M. M. Veloso, and M. Blum. Claimeval: Integrated and flexible framework for claim evaluation using credibility of sources. In *AAAI*, 2016.
[21] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic detection of rumor on sina weibo. In *ACM SIGKDD Workshop on Mining Data Semantics*, 2012.