Finding true and credible information on Twitter

S. Sikdar[1], S. Adalı[1], M. Amin[2], T. Abdelzaher[2], K. Chan[3], J.-H. Cho[3], B. Kang[4]& J. O'Donovan[4]

[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180
[2]Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801
[3]Army Research Laboratories, Adelphi, MD 20783
[4]Department of Computer Science, University of California, Santa Barbara, CA 93106

***Abstract*—In this paper, we present a unique study of two successful methods for computing message reliability. The first method is based on machine learning and attempts to find a predictive model based on network features. This method is generally geared towards assessing credibility of messages and is able to generate high recall results. The second method is based on a maximum likelihood formulation and attempts to find messages that are corroborated by independent and reliable sources. This method is geared towards finding facts in which humans are treated as binary sensors and is expected to generate high accuracy results but only for those facts that have higher level of corroboration. We show that these two methods can point to similar or quite different predictions depending on the underlying data set. We then illustrate how they can be fused to capture the trade off between favoring true versus credible messages which can either be opinions or not necessarily verifiable.**

I. Introduction

Microblogs like Twitter are one of the first sources of information in times of unrest, uncertainty and other extreme conditions. As a result, developing methods to find true information in microblogs is of crucial importance. In this paper, we investigate an important fusion problem: whether to find credible or true information? Sometimes these two are the same, but sometimes they are not. Understanding the difference and being able to account for it is crucial in developing effective methods to process information. In this paper, we introduce a systematic and thorough study that shows how truth and credibility based methods can be combined.

There are two complementary approaches to finding true information in microblogs. The first approach tries to understand how different network based signals can be used to assess which information is likely to be true [3], [4], [12]. We will call this method the machine learning or `ML` method. The main emphasis in machine learning methods is to find indicators of truth, trust and credibility like the network location of message senders and propagators, textual content of the message, etc. The success of these methods depend heavily on the availability of sufficient training data that is representative of the current problem. This can be challenging in some situations where no network feature is really a good signal or no prior network activity is a good predictor of the future. For example, a model trained on the importance of signals such as the network location and textual content of the message can fail when such signals can be manipulated by others and can vary in different populations. An important distinction is that often machine learning methods are based not just on finding truth, but also credible information as features concentrate on isolating which sources and what types of messages are generally believable.

TABLE I. The two methods used to find true and credible information.

	Machine Learning (`ML`)	Expectation Maximization (`EM`)
Prediction	Per message	Per cluster of messages
Assumption	Past behavior is predictive of future behavior	Messages corroborated are more likely to be true; sources who author corroborated messages are reliable
Method	Trained on separate ground truth	No training data, based expectation maximization using a statistical model
Results	Binary (True or False)	Ranking (score based on credibility)

The second approach for finding true information ignores the message content and network features completely and integrates the information about the structure of the network explicitly in it's model. It looks at humans as sensors and through an expectation maximization algorithm arrives at an estimate of the reliability of the sensors and the claims made by them. We call this method the expectation maximization or `EM` algorithm. While machine learning methods are trained on a given ground truth, estimation based approaches rely on an underlying model of how claims can be considered reliable. This way, they are free of bias that could be present in the training sets. As a result, estimation methods typically try to find correct information, but not necessarily credible information. Such methods are dependent on having the correct model assumptions and good amount of data that fit these assumptions. For example, such a method would consider information as correct if it is corroborated by independent sources [16]. For a message to be considered true in this method, it must be sufficiently endorsed within the network or the source of the message should have other messages corroborated by others. In absence of these, such a model may not be able to say with any confidence whether information is correct or not.

The main problem we address in this paper is this: how can we combine these two types of methods? As we can see from the summary of these methods in Table I, they are not directly comparable. They solve related but slightly different problems. Estimation methods try to eliminate facts with little support and group the rest. Machine learning methods rank information but do not have a notion of support equivalent to the estimation methods. There are instances when both methods agree, but the more interesting problem is to understand when, why and how they disagree. A feature considered as a positive signal for a

machine learning method may turn out to be a negative signal for estimation. The correct way to combine these two methods is not straightforward without understanding the expected correct result: the ground truth. It turns out defining the ground truth is a significant challenge in itself. While some methods try to assess truth, others assess credibility [12]. These two are not the same thing. True information may not always be the most credible. The best fusion method should also take into account when messages are both credible and true, but also highlight when credibility and truth point to different results. Keeping this in mind, this paper makes the following unique contributions:

- We first collect three separate data sets from Twitter centered around different events, one during Hurricane Sandy and one on the relief effort after the hurricane, and the last data set on the Egypt elections. We then compute predictions based on two state of the art methods based on machine learning (ML) [13] and expectation maximization (EM) [16] on all the data sets.
- We then compute different types of ground truth on these datasets. Here the term ground truth refers to an external method to assess the truth or credibility of messages. Two of the methods are based on manual labelling for truth or credibility. The last method looks at network activity as a proxy for credibility. As we have shown in prior work [13], this type of ground truth tends to be noisy, but can be improved by considering multiple types of ground truth in conjunction.
- Next, we perform a comparative analysis of the predictions on these datasets. We look at the set of features from the machine learning approach and see how these features correlate with each prediction. We find that the EM method is signalled by similar features as ML in two of our datasets *despite the fact that* EM does not consider these features in training. In this case, truth and credibility point to similar results in both methods. However, we show that the predictions point to almost opposite results on the Egypt elections dataset. True and credible messages can be very different depending on the dataset.
- Finally, we introduce a number of fusion methods to combine these two methods based on various assumptions. We show how we can address the trade off between concentrating only on true messages which leads to the filtering out many credible ones, versus combining the two to introduce credible but not easily verifiable information.

In the remainder of this paper, we introduce both methods briefly. We then discuss our experimental setup and present our results to make the case that fusion of these two methods can sometimes improve on both truth and credibility of the results, especially when they both point to the same type of messages. In other cases, their fusion provides a trade-off between the two if they point to different results.

Related Work

Information credibility has been studied extensively on Twitter. A common approach is to find which features are good indicators of credibility [4], [6], [10]. The features can be simple meta-data about the message or the sender [14], or go deeper into the processing of network structure [3], [8], textual tokens [11], behavior of individuals towards each other [1] and complex network based methods that incorporate a combination of these. These features are typically a combination of signals aimed at measuring the believability of the message text and the reliability and competence of the sender in a given topic. These features can be combined to give a good indication of what type of information will be found believable and trusted by others in the network [7]. This network level trust is then used as a proxy for how likely the information is likely to be true. Note that information that is trusted and found believable in the network can in fact be false. Herd behavior and the impact of influence in social networks is well-documented [2].

An important problem in feature based methods is identifying the ground truth to be used for training models to make future predictions. In many cases, an objective ground truth does not exist. While crowd-sourcing of surveys has become quite popular in this type of work [5], our prior work shows that the results of surveys can be quite noisy and vary from population to population greatly [12], [13]. Often credibility is hard to assess as the survey subjects are unlikely to know about the topic of the message. We have shown that such methods can be combined with network behavior based measures that are noisy but informed about the topic and source to obtain more stable measures. Hence, fusing ground truth from independent sources with possible uncorrelated noise is a way to obtain better ground truth after removing data points where the credibility is uncertain. We have shown in prior work that we can achieve the best prediction rates reported in the literature by choosing a representative set of features and choosing a stable ground truth to train these on [12], [13].

A different approach to credibility analysis comes from sensing literature [16], [17], [15]. These methods model humans as unreliable binary sensors generating observations. This model is focused on the observations pertaining to the state of the physical world, having unique and verifiable ground truths, independently observable by different sources. Majority voting is a trivial approach to asses such observations, when the sensors are all calibrated and have equal rates of error. However, human sensors differ largely in their level of activity and the likelihood of making wrong judgments. The literature dealing with this problem can be rooted back to classic algorithms like Hubs and Authorities [9]. Wang et al. [17] propose a maximum likelihood framework to jointly estimate the credibility of the sources and the claims, maximally consistent with the reported observations. Their method assumes the sources as independent. Subsequent works consider the human social network as the propagation channel of the observations, and accounts for the source dependencies directly in the maximum likelihood formulation [16]. [15] considers the case where the claims can have conflicts.

II. Machine learning method (ML)

We first approach the task of predicting the credibility of a single message. We frame this as a supervised learning problem. As in our past work [13], we define a notion of ground truth for credibility as a combination of two methods: the credibility assessments from a post-hoc human crowd-sourced survey, and the retweet count of the message which we consider as an in-network measure of credibility. We call this ground truth TRT, the message text was found credible

(T) and it was retweeted (RT). A message is *credible* with respect to TRT if the message is found credible by human annotators as well as by Twitter users who use retweet behavior as a noisy proxy for voting on the credibility of a message. A message is considered *not credible* if the two sets of annotators find the message to be not credible. To achieve good class separation we throw away instances where the two sets of annotators could not agree on the credibility of the message, of one annotator was uncertain (only one retweet or survey score is in the midrange). This method yields a ground truth that helps us find messages that were found to be credible by the crowd and were propagated by retweets and had a chance to make an impact on Twitter. We acknowledge that each of these methods on its own is a noisy source of ground truth, subjective and may be biased by the influence of the sources. By combining the two methods we tune our learning task to find credible messages that made a larger impact while also arriving at a ground truth where our features have better predictive ability. For our learning task we use features used in [13] that is representative of features used in other previous work as well. Broadly we have two kinds of features: Content based features (Table II) based on textual content and Behavioral user based features (Table III) computed from Twitter metadata and statistical measures of user behavior in pairs within our larger dataset.

Note that, the success of this method is measured on a model trained on each dataset separately. In all our experiments, we have collected ground truth for the specific dataset by first sampling a training set from the data, running a survey on this set, finding when survey subjects agree on a message and throwing away uncertain cases, combining this with the retweet behavior and then training a linear regression model. Then, this regression model over the given features is applied to the test data to determine prediction accuracy. As such, it is still a big challenge to assess when a model trained on one data set can be applied to another data set. Without an already trained model, this type of method is applicable only for posthoc analysis. However, understanding common features across different datasets is a first step towards developing general machine learning models.

TABLE II. THE SET OF CONTENT-BASED TWITTER FEATURES ANALYZED IN OUR EVALUATION.

feature name	*description*
char/word	# chars/# words
question	# question marks
excl	exclamation marks
uppercase	# uppercases in text
pronoun	# pronouns (count by corpus)
smile	# smile emoticons
frown	# frown emoticons
url	# urls
sentiment_pos / sentiment_neg	positive/negative word count based on lexicon sourced from NLTK[1]
sentiment	polarity (sentiment_pos - sentiment_neg)
num_hashtag	from entity metadata
num_mention	from entity metadata
tweet_type	0:none 1:RT 2:Mention 3:RT+Mention
ellipsis	counting ellipsis sign(. . .)
news	occurrence frequency of news sources
lex_diversity	proportion of unique words per tweet
dialog_act_type	category: statement, system, greet, emotion, ynquestion, whquestion, accept, bye, emphasis, continuer, reject, yanswer, nanswer, clarify, other
news_words	NLTK corpus of news article terms (sourced from Reuters), count of occurrence
chat_words	AOL messenger corpus, count of occurrence

TABLE III. THE SET OF USER-BASED TWITTER FEATURES ANALYZED IN OUR EVALUATION.

feature name	*description*
u-friend/u-follower	# friends/followers (log)
u-age	# years on Twitter
u-bal_soc	ratio of follower to friends
u-url / u-mention	mean # urls /mentions in tweets
u-hashtag	mean # hashtags in tweets
u-length	mean text length in tweets
u-balance	mean balance of number of followers
u-conv-balance	mean balance of conversations
u-tweets / u-favorite	# of tweets / tweets favorited
u-rt_out	total # tweets of the user that are retweets
u-time	mean time between tweets
u-directed-ratio	# directed tweets/#broadcast tweets
u-retweet-ratio	# retweets/#tweets
u-prop-from	# users the user propagates from
u-prop-to	# users that propagate the user
u-response	mean response time to tweets
u-propagated-tweets	# tweets propagated by other users
u-propagation-energy	amount of propagation energy spent on this user by others
u-worthiness	proportion of user's tweets found worthy of propagation by others
u-conv	mean # conversations
u-conv-tweets	mean # tweets per conversations (also min,max)
u-tau	mean time between messages between pairs

III. EXPECTATION MAXIMIZATION ALGORITHM (EM)

Unlike the above method that trains classifiers to recognize content features that signal credibility, the statistical method is an unsupervised approach that treats human sources as sensors who offer binary observations (e.g., tweets). The observations are binary because they can be either true or false. Clustering methods are used to group similar observations into the same cluster. Each such cluster is called a *claim*. A source-claim network is thus formed, where sources are connected to the claims they made, and the claims are connected to the sources who made them. The topology of this network is analyzed to extract the likelihood of correctness of claims.

If the reliability of individual sources were known, the problem of computing the likelihood of correctness of individual claims would be trivial. One would simply consider the reliability of sources that made a claim in determining the probability that the claim is correct. Bayesian networks offer a solution to this problem when more than one source makes the same claim. The challenge, however, lies in the fact that the reliability of the sources is not known. To address this challenge, recent research casts the problem as one of joint estimation of the reliability of sources and the correctness of claims, given only the source-claim network [17]. A likelihood function was developed that expresses the probability that the given sources would make the given claims, as a function of the (unknown) reliability of sources and the (unknown) correctness of claims. This likelihood function is then iteratively maximized, ultimately computing the values of the unknowns that lead to the maximum-likelihood solution. This is called the *expectation maximization* algorithm, or EM. Recent work described EM solutions that jointly determine the reliability of sources and the correctness of claims, applied to a source-claim

TABLE IV. DATASETS USED IN OUR STUDY; *set1* IS THE SET OF TWEETS WITH CREDIBILITY LABELS OBTAINED FROM A CROWDSOURCED SURVEY; *set2* IS THE SET OF TWEETS MANUALLY LABELLED BY OUR GROUP FOR TRUTH, AND *set3* IS THE REMAINING TEST DATA. ALL OF SET1, SET2 AND SET3 ARE UNIFORMLY SAMPLED FROM THE MAIN DATA SET AND ARE DISJOINT.

Dataset	#Tweets total	*set1*	*set2*	*set3*
Egypt (E)	1M	10K	200	190K
Frankenstorm (FR)	3.8M	9K	200	191K
Occupy Sandy (OS)	60K	6.5K	200	53.5K

network derived from Twitter feeds [17]. These solutions were further extended to the case of non-independent sources [16] and non-independent claims [15].

Non-independent sources refer to sources where one influences another. An example is the case where a source retweets (many of) the other's messages. If the original tweet relays some observation, its retweet is not considered to be an independent corroboration of this observation. Hence, if the observation is incorrect, retweets may give rise to correlated errors. The EM algorithm takes the probability of such correlated errors into account in computing the maximum likelihood estimates of the reliability of individual sources and the correctness of claims. Similarly, the EM algorithm formulation can account for non-independent claims; for example, pairs of claims that cannot be true at the same time.

The statistical approach described above has several limitations. Most importantly, it does not consider the content of the claims and textual features are not used. Supervised learning approaches nicely complement this deficiency by learning features correlated with high credibility. Second, it relies heavily on the degree of (independent) corroboration. Hence, tweets that are mentioned only once by an unknown source cannot be reasoned about. On the other hand, the approach is good at making use of corroboration patterns, when they exist. For example, it can suppress rumours by observing that their sources are not independent, but rather follow a retweet chain.

The above analysis leads to simple heuristics for deconflicting tweet rankings when the unsupervised statistical approach disagrees with the learning-based approach. Namely, when the rankings are in conflict, if the tweet comes from a large cluster, we favor the judgement of the statistical approach, since we know that the accuracy of this approach improves with increasing cluster size. Otherwise, the judgement of the machine learning approach is favored. We investigate such de-conflicting heuristics to determine how best to combine the two approaches.

IV. EXPERIMENTAL SETUP

To study how the two methods can be merged, we collect three separate datasets. The data sets are summarized in Table IV. Datasets FR and OS were generated from tweets related to the Hurricane Sandy of 2012. Dataset FR was generated from tweets in the lead up to and in the immediate aftermath of the hurricane using keywords *Sandy* and *Frankenstorm*, while the dataset OS has tweets related to the Occupysandy movement which comprised of volunteers arranging relief measures for the victims of the hurricane

TABLE V. PREDICTION FOR THE GROUND TRUTH TRT OVER THE DATASETS FR, OS, E. FOR ML

Dataset	Baseline	Accuracy	Kappa	ROC Area
FR	56.94	89.43	0.78	0.95
OS	79.50	92.08	0.75	0.96
E	72.28	92.39	0.80	0.96

shortly after the disaster, collected using *OccupySandy* as the keyword. The dataset E on the recent political events in Egypt was generated from tweets related to the deposing of then President Morsi. These datasets also capture a variety of contexts from natural disasters to political events. We have separated each data set into three disjoint groups: *set1, set2, set3* where $set1 \cap set2 = \emptyset, set2 \cap set3 = \emptyset, set1 \cap set3 = \emptyset$. The FR and E datasets are too large for either method. As a result, we sample a set of at most 200K tweets uniformly at random from these. Then, we sample *set1* from this data set randomly. From the remainder, we sample *set2* randomly. The remaining data is put in *set3*.

a) Predictions by the ML *method.:* We use *set1* as the training data for the ML method. We run the Amazon Turk survey on this data set and ask participants the credibility of messages by showing them nothing but the text. We choose only those messages in which the majority of the survey takers agree on the label (either strongly credible or strongly not credible, we throw away the middle range). This gives us the ground truth T. We also find messages not retweeted (class 0), and retweeted more than once (class 1) to compute the RT ground truth. We combine T and RT into the credible and not credible classes as described in Section II to compute the ground truth: TRT on *set1*. This dataset is then used to train and obtain a linear model over the features using logistic regression and 10-fold cross-validation. We choose a linear model to prevent fitting noise and having a greater hope of generalization. We then use this model to obtain predictions for the remainder of the data, namely $set2 \cup set3$. We present the computed prediction accuracy, Kappa statistic and ROC Area on the validation data in Table V. We find that the models show good promise of generalization on test data.

b) Predictions by the EM *algorithm.:* The estimation method uses the full data set to make predictions. We compute clusters and make predictions at the cluster level on the whole data set by this method. In all our tests that compare ML and EM methods or their fusion, we report only on the predictions for $set2 \cup set3$ to make the results comparable.

To be able to analyse the results of the fusion methods, we have manually labelled 200 messages from each data set (Figure 1), given by *set2*. In this label, our main emphasis was to establish whether a message contains a true fact, a fiction (a false statement) or an opinion. Note that following the human as sensors analogy, we are only interested in factual statements in this context. The label opinion is used for cases when the statement is not a statement about the outside world, such as "I am sleepy". There are cases where the statement is factual but cannot be verified. We have included those statements as opinions as well. True and fiction labels are reserved only for those facts that can be verified by other news sources found on the network. The statistics for these manual labels are given

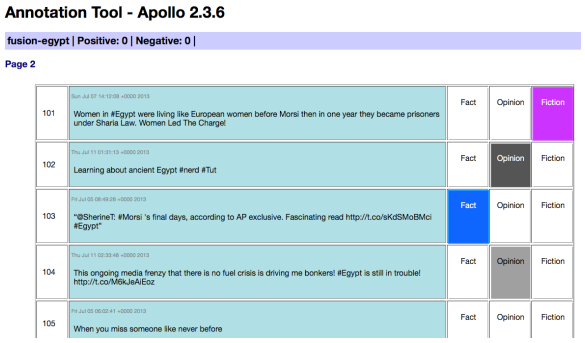

Fig. 1. Screen shot from the manual grading survey performed internally

TABLE VI. NUMBER OF MANUAL LABELS OBTAINED FOR *set2*

Dataset	Fact	Fiction	Opinion
E	89	14	97
FR	85	4	111
OS	102	7	91

in Table VI.

V. ANALYSIS

In this section, we analyse the results from the two different methods. We first take the EM and ML predictions on $set2 \cup set3$, and look at how they correlate with the features used in the ML method. Recall that these features are not used in the EM method. However, we would like to see whether the correlation between these features and a given method differ. In Figure 2, we show the correlation between a given prediction method and all the features in our study, divided into two groups: user (Table III) and content features (Table II). The correlations are reported for each message despite the fact that EM predictions are made for a given cluster. The mean cluster sizes for E, FR and OS are 1.5, 2 and 3.15 respectively.

There seems to be no clear distinction between the feature correlations for EM and ML for the datasets, FR and OS. Despite the fact that EM is not trained specifically on these features, similar features tend to point to both credibility and truth. In both datasets, user based features are not significant for EM, and are negatively correlated for the ML approach. It is very interesting that these methods tend to track very similarly. In FR, there is a lot of random speculation about the storm as well as some announcements, but no true discussion between experts. No feature is strongly correlated with predictions in this dataset. In OS, a great deal of the messages tend to be requests for help. There are few features that are strong indicators, but these tend to be similar in both datasets.

The situation is very different for the Egypt dataset however. Quite a few of the features are strongly positive for one method and strongly negative for the other method. In this dataset, credible and true information are signalled by very different features. In essence, this shows us that EM and ML methods capture a different aspect of the messages. We summarize the significant content based features in all datasets in Table VII. In particular, we can see that features based on many words, question marks, parentheses, exclamation points and citation of news sources happen to be corroborated by reliable sources in this dataset (according to EM). However, features that are indicative of the type of words used are

TABLE VII. FEATURES THAT ARE STRONGLY CORRELATED FOR MACHINE LEARNING (ML) AND EXPECTATION MAXIMIZATION (EM) APPROACHES. FOR CASES WHERE THE CORRELATION IS STRONGLY INVERSELY CORRELATED, WE REPRESENT THIS WITH AN OUTLINE OF THE DATA SET LABEL (*e.g.*, ($\overline{\text{FR}}$)).

feature	**ML**	$\gtrless$	**EM**
char			
word		$<$	(E)($\overline{\text{FR}}$)
question		$<$	(E)($\overline{\text{FR}}$)
excl		$<$	(E)
uppercase			
pronoun			($\overline{\text{FR}}$)
smile		$<$	(E)($\overline{\text{FR}}$)
frown		$<$	(E)
url	(OS)	$=$	(OS)
retweet			
sentiment pos			
sentiment neg			
sentiment			
num hashtag			
num mention			
tweet type	(OS)	$=$	(OS)
ellipsis		$<$	(E)($\overline{\text{FR}}$)
news		$<$	(E)(FR)
lex diversity	(E)	$>$	
news words	(E)	$>$	
chat words	(E)	$>$	(FR)
dialogue act type	(E)	$>$	

significant for ML and message credibility; use of news words, number of unique words and chat words. Similarly, a majority of the user based features are strongly negatively correlated for ML, while they are strongly positively correlated for EM. In essence, there are users who are involved in discourse on this topic who are signalled by very specific word usage and social behavior that is not very obvious to an outsider who is ranking messages for credibility.

Given that the nature of these predictions are more similar in FR and OS based on the type of features that correlate with them, we expect that fusion will improve both truth and credibility. However, in the Egypt dataset, the fusion is combining two different types of predictions. As a result, it will mix two types of results.

VI. FUSION RESULTS

One of the main goals of this paper is to find simple fusion methods that predict messages of interest more accurately. An important question to consider is the choice of label to evaluate the effectiveness of the combined method. Depending on the scenario under which we are required to operate, we may want to exclusively report messages that are verifiable facts or also include messages that are credible. The manual labelling obtained for *set2* is an assessment of whether a given message is a verifiable fact, an opinion or fiction. We choose to assign the labels values 0 for Fiction, 0.5 for Opinion and 1 for Fact. The individual methods, the ML method and the EM algorithm are designed to find credible information and corroborated facts respectively.

The `ML` method uses linear models trained on annotations from a human survey to assess credibility. It is important to note that in the human survey users were also shown cues like the existence of a url in the message and it is reasonable to assume that messages that contain facts appear credible. This

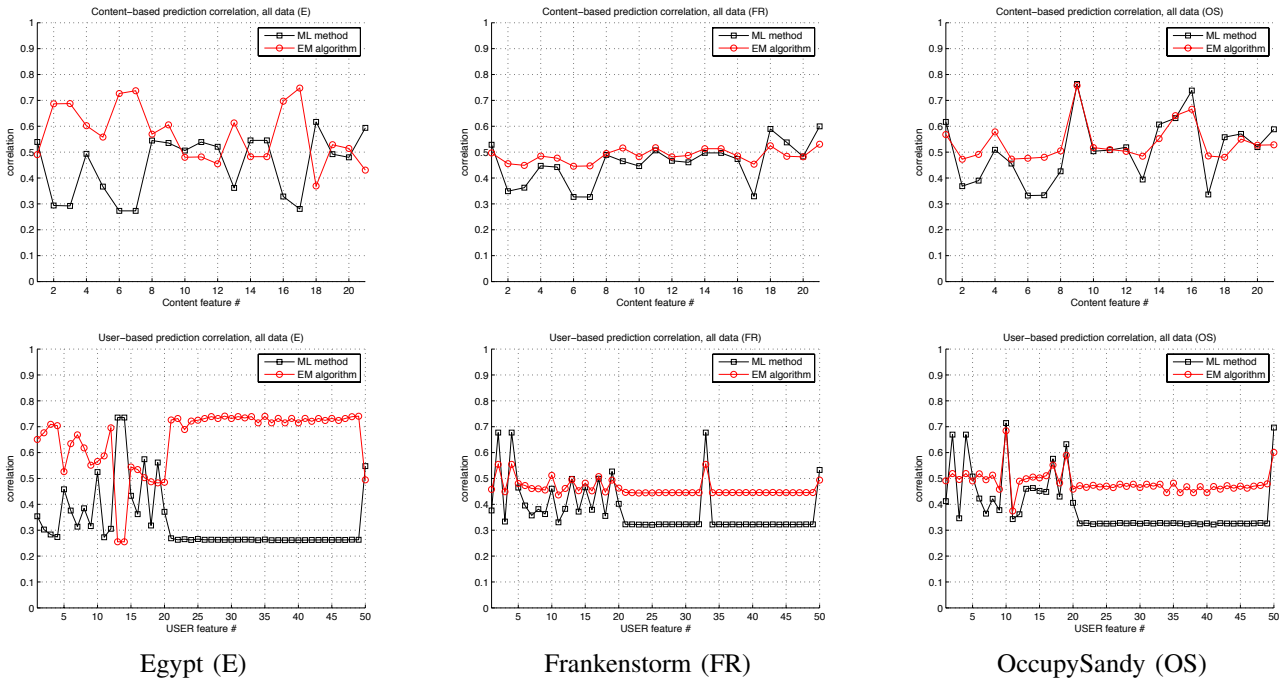

Fig. 2. Correlation of ML features with both ML and EM based predictions for content based features (top row) and user features (bottom row). Rescaled for convenience. A correlation of 0.5 is no correlation, a higher value shows a positive correlation and a lower value shows a negative correlation.

assumption is supported by previous work [7], [8], [13]. On the other hand it is also reasonable to assume that messages that are opinions are hard to make assessments of credibility on. Therefore unless an opinion conveys something that is in violation of raters beliefs, it is likely to be rated credible at best and as having uncertain credibility at worst. Since these assessments are made post-hoc, the survey respondents may have prior knowledge of the event or otherwise be able to judge a message as not credible if the claim made contradicts prior beliefs about the topic and related events or the claim made is otherwise incredible. The combination of these human crowdsourced survey assessments with the in-network retweet propagation behavior means that messages found to be highly credible by multiple sources are labelled credible and messages that were found to be not credible and were rejected by the users in the network are labelled as not credible. We also observe that there is significant class imbalance in our training data with a majority of instances found to be credible. This leads us to believe that a majority of messages on Twitter are credible. We note however that the models achieved low false positive and high true negative rates.

The EM algorithm meanwhile is designed to predict as more likely to be true messages that were corroborated by users with high reliability than messages that were promoted by many unreliable sources. With a good starting point of prior beliefs about the claims for a small subset and a dense network, the EM algorithm is therefore likely to correctly rate messages containing verifiable facts that were corroborated by reliable human sensors higher than messages containing fiction. By construction, sensors believed to be more reliable are more likely to make claims that are verifiable and believed to be true by the current state of the algorithm. The algorithm improves it's belief about the reliability of a sensor if it makes true claims more frequently than otherwise. Since opinions cannot be verified as true fact or as veritably false, if at the starting point we take a sensor to make, in expectation, a true or false claim with equal probability and hold the belief that a given claim is an opinion, then we learn nothing about the sensors or the claim.

We try some simple fusion methods towards developing robust methods that use the strengths in each individual method and that can be tuned to suit different operational requirements. To evaluate each of the fusion methods we compute the Root Mean-Square Error achieved at the task of predicting the manual grades obtained for *set2* by each of the available prediction methods: the EM algorithm, the ML method and our combined fusion methods. The evaluation is performed over claims rather than at the individual tweet level. We first find the average prediction score by each method for each of the claims for which we have a manual grades. If a claim has multiple manual grades, we assign the claim cluster the mean score obtained over all the graded messages belonging to the cluster. We then compute the RMSE of the perdictions made by each method against the manual grades assigned to the cluster over all such clusters. We present these errors in Table VIII. Lower error rates are more desirable in this case.

We first try a fusion method based on cluster size. This approach is motivated by the observation that the EM algorithm cannot make predictions with high confidence on claim clusters that are very small in size. A vast majority of claims correspond to small clusters. The ML method can use features computed on the messages making up the claim to make a prediction on the credibility of the claim with high accuracy regardless of the size of the claim due to the promise of good generalization from our observations on the performance of

TABLE VIII. RMSE OF THE PREDICTIONS BY ML METHOD, EM ALGORITHM AND VARIOUS FUSION METHODS ON MANUALLY GROUND TRUTHED TEST SET FOR EACH DATASET.

Prediction Method	E	FR	OS
ML prediction alone	**0.393**	**0.354**	**0.329**
EM prediction alone	0.664	0.533	0.532
Choose ML or EM based on cluster size	**0.391**	**0.334**	**0.327**
Weighted sum of ML or EM w.r.t. cluster size	**0.392**	**0.352**	0.329
Min of predictions by ML and EM	0.661	0.533	0.524
Max of predictions by ML and EM	0.397	**0.354**	0.341

each model during cross validation. The ML method predicts a credibility rating. We have already established above that there is a relationship or dependence between the credibility of a message and whether it is an opinion or a claim that is fact or fiction. We therefore use the prediction made by the ML method on claims with a smaller number of messages in our dataset and use the EM algorithm's predictions on claims for which we have larger clusters. We choose this approach since the EM algorithm makes predictions based on the corroboration of a claim by a larger number of independent sources and accounts for the reliability of each sensor. We learn the threshold on the size of clusters to decide between the ML method or the EM algorithm to predict a score for a claim. The behavior of this fusion method is shown as a function of this threshold in Figure3 for each of our datasets. It is important to note that although the error reported for the EM algorithm is consistently higher, this is because it is likely to perform poorly on smaller claim clusters and the majority of the claims correspond to small clusters. In each of the datasets we find that beyond a certain threshold the fusion method performs better than the best individual method. We also find that there is a critical threshold cluster size at which the fusion method is the most improved beyond which the error increases again. This is because the EM algorithm starts to provide an improvement over the ML method beyond a certain cluster size but as we increase the threshold we use fewer of the predictions from the EM algorithm that yielded improved performance.

We also tried a weighted sum approach to combining the two methods. This was again motivated by the same observation we made for the cluster size based approach. We first compute a weight to be the size of the cluster corresponding to the current claim normalized by the size of the largest cluster. We assign greater weight to the ML method on smaller clusters and favour the EM algorithm for larger clusters. Again, the combined fusion method has a lower error rate than the best individual method.

Lastly, we tried an optimistic approach of using the higher of the scores output by each of the two methods. We find that there is no significant improvement by using such an approach although the error is not as high as the approach of using the minimum of the two scores. This is can be attributed to the EM algorithm predicting a score of 0 on the smaller clusters since it has no information and a majority of claims in our dataset being credible or true.

VII. CONCLUSION

In this paper we have presented two alternate and successful approaches of predicting credibility and truthfulness of messages and the claims made by these messages on microblogging sites. We have established a case for the use of each individual method depending on the requirement of the user and shown similarities and differences between the two methods and the trade-offs inherent to each. The EM algorithm depends on the initial beliefs on the claims and suffers from the running time and convergence properties of expectation maximization algorithms. The EM algorithm is unsupervised and must run on each set of source-claim topology for which a prediction is required. It also ignores features of otherwise credible messages. However, while it does not have sufficient information to make good predictions on small clusters, on large clusters it makes very good predictions. Under the assumption that the reliability of human sensors can be estimated given dense network information, the EM algorithm can leverage information from a greater set of independent sources from the available dataset alone without requiring a lengthy post-hoc labelling task. The ML method on the other hand trains models on each dataset and requires computing features on each message. It predicts credibility but not truthfulness due to the lack of a suitable mechanism to collect ground truth. As such, it also suffers from the biases of individual raters and is susceptible to noise in the annotations. However the ML method once trained can quickly output predictions on test instances with good guarantees on performance on the task of predicting credibility. Since the ML method is trained on annotations obtained from humans it can also capture subtle aspects of credibility that are not necessarily machine-analyzable and could capture more detail than a purely statistical approach.

We have also shown that the simple fusion methods we proposed to combine the two methods provided improvements in the prediction error over the best available individual method consistently over multiple datasets representing a broad range of topics and scenarios. Given a dataset, we can learn the optimal threshold to achieve the lowest prediction error. We can also tune the fusion method to suit operational needs in practice. If only verifiable facts are desired at the expense of potentially useful credible information we can place a greater weight on the output of the EM algorithm. In all other cases, the ML method provides better accuracy and generalizes well to the test set and can be used to quickly provide predictions for large test datasets individually or in combination with the EM algorithm to improves on the predictions made by the ML method for large potentially significant claims.

This paper is a first attempt to understand how these two complementary approaches differ in their predictions and how they can be combined. However, the design space for fusing these two methods is large and provides many interesting avenues of research. Given that ML method can be very useful for cases in which there is little corroboration, it can be used as a prior on source reliability for the EM method. Similarly, EM method can serve as a secondary ground truth to bias ML models. As the EM method can provide much large ground truth sets, it can be very useful for learning the best features especially for data sets like the Egypt dataset in which there is a distinct difference between truth and credibility based indicators. However, many questions remain on how to define the appropriate ground truth in these cases and how to capture the trade-off concerning accuracy versus recall when considering facts and opinions. The investigation of these problems is future work.

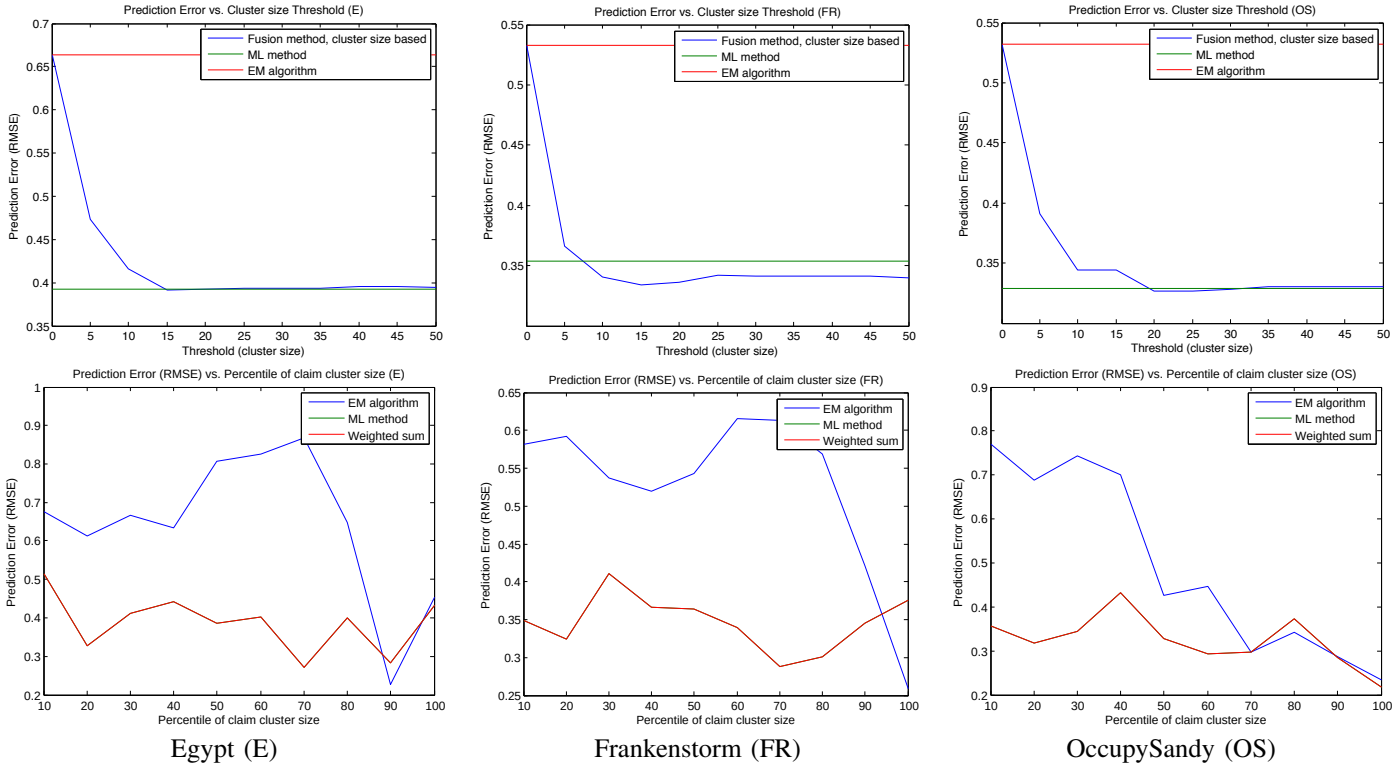

Fig. 3. Row 1. RMSE of predictions made by cluster size based fusion method against labels obtained from the manual survey (*set2*) as a function of the cluster size threshold used. Error made by the individual methods on the same test data are also shown. Row 2. Error rate obtained by individual methods and the weighted sum fusion method on subsets of the test set partitioned by the percentile of the size of the corresponding claim. Since the weighted sum fusion method matches the ML method in prediction error, only the weighted sum is shown in the bottom row.

Acknowledgements. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] S. Adalı, M. Magdon-Ismail, and F. Sisenda. Actions speak as loud as words: Predicting relationships from social behavior data. In *Proceedings of the WWW Conference*, 2012.

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence based contagion from homophily driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[3] K. Canini, B. Suh, and P. Pirolli. Finding credible information sources in social networks based on content and social structure. In *SocialCom*, 2011.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW*, pages 675–684, 2011.

[5] S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John. Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist*, 59(2):93–104, 2004.

[6] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *PSOSM*, pages 2:2–2:8, New York, NY, USA, 2012. ACM.

[7] B. Hilligoss and S. Y. Rieh. Developing a unifying framework of credibility assessment: Construct, heuristics and interaction in context. *Information Processing and Management*, 44:1467–1484, 2008.

[8] B. Kang, J. O'Donovan, and T. Hollerer. Modeling topic specific credibility on twitter. In *IUI*, pages 179–188, 2012.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, Sept. 1999.

[10] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *SOMA*, pages 71–79, 2010.

[11] J. O'Donovan, B. Kang, T. Hollerer, and S. Adalı. Credibility in context: An analysis of feature distributions in twitter. In *SocialCom*, 2012.

[12] S. K. Sikdar, B. Kang, J. O'Donovan, T. Hollerer, and S. Adalı. Cutting through the noise: Defining ground truth in information credibility on twitter. *ASE HUMAN Journal*, (3):151–167.

[13] S. K. Sikdar, B. Kang, J. O'Donovan, T. Hollerer, and S. Adalı. Understanding information credibility on twitter. In *Proceedings of SocialCom*, 2013.

[14] B. Suh, L. Hong, P. Pirolli, and E. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*, 2010.

[15] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Exploitation of physical constraints for reliable social sensing. In *Proceedings of the 2013 IEEE 34th Real-Time Systems Symposium*, RTSS '13, pages 212–223, 2013.

[16] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le. Humans as sensors: An estimation theoretic perspective. In *Proceedings of the ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN'14)*, 2014.

[17] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, IPSN '12, pages 233–244, New York, NY, USA, 2012. ACM.