Constraint-Aware Dynamic Truth Discovery in Big Data Social Media Sensing

Daniel (Yue) Zhang, Dong Wang, Yang Zhang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
yzhang40@nd.edu, dwang5@nd.edu, yzhang42@nd.edu

***Abstract*—Social media sensing has emerged as a new big data application paradigm to collect observations and claims about the measured variables in physical environment from common citizens. A fundamental problem in social media sensing applications lies in estimating the evolving truth of claims and the reliability of data sources without knowing either of them *a priori*, which is referred to as *dynamic truth discovery*. We identified two critical challenges that are not fully addressed by solutions from current literature. The first challenge is "physical constraint-awareness" where the transition of truth is constrained by some physical rules that must be followed to ensure correct estimation of the evolving truth. The second one is "noisy and incomplete data" where the social media sensing data is sparse in nature and contains a lot of rumors and misinformation, making it difficult to capture the constantly evolving truth of measured variables. In this paper, we developed a new Constraint-Aware Dynamic Truth Discovery (CA-DTD) scheme to address the above challenges. To address the physical constraint-awareness challenge, CA-DTD develops a new constraint-aware Hidden Markov Model to effectively infer the evolving truth of measured variables by incorporating physical constraints. To address the noisy and incomplete data challenge, CA-DTD fuses sensing observations from online social media with information from traditional news media using a principled approach. We evaluate CA-DTD scheme using two real-world social media sensing data traces and the results show that CA-DTD significantly outperforms the state-of-the-art baselines.**

***Keywords*-Dynamic Truth Discovery, Big Data, Social Media Sensing, Constraint-Aware Hidden Markov Model, Data Fusion**

I. Introduction

Online social media has become as a new big data application paradigm to collect the observations (often called claims) about the physical world from the reports (e.g., tweets, pictures) shared by common citizens on the social media. This new sensing paradigm is motivated by the wide adoption of portable devices, ubiquitous wireless connection and the proliferation of online social media [1], [2]. An important problem in social media sensing is to accurately discover the truthful information from the massive noisy and potentially conflicting claims contributed by unvetted sources on social media [3]. In this paper, we refer to this problem as *dynamic truth discovery*. For example, in a terrorist attack scenario, real time reports about the current situation of the attack (e.g., the number of casualties, the escape path of suspects, and the safety alerts to the public) are available on social media (e.g., Twitter). Due to the unvetted nature of data sources (e.g., Twitter users) and dynamic nature of the variables of interest (often called measured variables), claims in social media sensing often contradict with each other and change over time [1], [4], [5]. The goal of dynamic truth discovery is to identify which claims are truthful and which data sources are trustworthy in real time.

Recent efforts have been made to solve the dynamic truth discovery problem in data mining, machine learning, and networked sensing communities [6]–[8]. These solutions include Markov models [6], Bayesian networks [8], maximum likelihood estimation (MLE) methods [7]. However, two critical challenges have not been fully addressed by the current solutions: *physical constraint awareness* and *noisy and incomplete data*.

Physical constraint awareness. Physical constraints have clear impacts on the truth discovery results in social media sensing applications [9], [10]. For example, the number of casualties cannot decrease during a terrorist attack event, and it is unlikely that a suspect can escape far away from the spot shortly after the attack. These physical constraints enforce restrictions on the transition of true values of the measured variables. Only a small number of truth discovery models start to explore the constraints on measured variables in their models [9]. However, they either assume the values of the dependent measured variables do not change over time (i.e., static truth discovery) [9] or assume measured variables are independent when their values change over time [10]. Therefore, there is a lack of a principled analytical framework that systematically considers the physical constraints on measured variables in solving the dynamic truth discovery problem.

Noisy and incomplete data. Social media sensing data is sparse in nature and contains a large amount of noise (e.g., rumors, misinformation), which makes the dynamic truth discovery problem more challenging [3]. For example, in the Boston Bombing event in 2013, CNN claimed that a bomber was arrested two days after the event. This original message was retweeted more than 3,000 times until half an hour later it was debunked by Boston police department claiming no arrest has been made. A few hours later the real arrest was finally made. Such misinformation can easily lead to the incorrect detection of the truth transition of the measured variables. Furthermore, social media data is observed to be sparse and incomplete due to various reasons such as the spontaneous nature of data sources, lack of incentives and privacy concerns. For example, in the Twitter data traces we collected for evaluation, more than 86% of the users only post one tweet and more than 91% post at most two tweets during the entire event. Such incomplete data often provides inadequate evidence to solve the dynamic truth discovery problem in social media sensing [11].

This paper develops a Constraint-Aware Dynamic Truth Discovery (CA-DTD) scheme to address the above challenges. In particular, to address the physical constraint awareness challenge, we develop a Constraint-Aware Hidden Markov Model (CA-HMM) to explicitly incorporate physical constraints into the dynamic state estimation of measured variables. To address the noisy and incomplete data challenge, the CA-DTD scheme fuses data from both online social media and traditional news media and seamlessly integrates the data fusion process with the CA-HMM. In the evaluation, we compare CA-DTD with the state-of-the-art baselines using two real-world social media sensing data traces collected from Twitter. The results show that the CA-DTD scheme significantly outperforms all baselines: it identifies the evolving truth of the measured variables more accurately and improves the computational efficiency.

In summary, our contributions are as follows:

- This paper addresses two important challenges of solving the dynamic truth discovery problem in online social media sensing applications: *physical constraint awareness* and *noisy and incomplete data*.
- We develop the CA-DTD scheme that incorporates a Constraint-Aware HMM to explicitly model the dynamic states of the measured variables and incorporate physical constraints to regulate the state transitions in the model.
- The CA-DTD scheme integrates a data fusion component that effectively fuses data from both online social media and traditional news media to address the noisy and incomplete data challenge.
- We evaluate the performance of the CA-DTD scheme and the state-of-the-art truth discovery solutions using two real-world datasets in social media sensing. The evaluation results demonstrate significant performance gains achieved by the CA-DTD scheme.

II. Related Work

Truth discovery problem has received a significant amount of attention from machine learning, data mining, and networked sensing communities [3], [4], [6], [12]–[14]. For example, Yin *et al.* developed a Bayesian-based algorithm, *Truth Finder*, to utilize the inter-dependency between website trustworthiness and fact confidence to find trustable websites and true facts [4]. Dong *et al.* explicitly modeled the source dependency in their truth discovery solutions and studied its impacts [12]. A semi-supervised graph learning scheme is proposed to model the credibility propagation from the known ground truths [15]. Wang *et al.* developed an unsupervised truth discovery solution that offers rigorous accuracy bounds on the analysis results using estimation theoretical approaches [16]. However, the above solutions either assume that the truthfulness of claims are time-invariant or data are of reasonable quantity and/or quality. Such assumptions barely hold true in real world social media sensing applications [6].

There exist a few previous studies on the dynamic truth that share some similarity with our work. Pal *et al.* proposed a method that takes into account the evolving true information of objects and estimates the truth of variables in current time interval based on sources' historical claims [17]. Zhang *et al.* proposed a scalable and dynamic truth discovery framework that addressed the evolving truth problem using a simple Hidden Markov Model [6]. However, these above solutions do not explicitly consider the physical constraints in their models, which may lead to incorrect estimation of unlikely truth transition.

Our work is also related to the Hidden Markov Models that consider the constraints in the state transition, For example, Kim *et al.* incorporated global path constraints in the Viterbi algorithm that limits the duration of states in word recognition application [18]. Roweis proposed a constrained Hidden Markov Model by modeling each state as a spatial region in a topology space that only allows state transitions to happen between spatial neighbors [19]. Weintraub *et al.* developed a speech recognition system that integrated speech and linguistic knowledge into a HMM framework [20]. While these works share the similar idea of constraining state transitions, the solutions only focus on the hard constraints that prohibit certain state transitions. In contrast, our CA-DTD scheme provides a more general constraint-aware state transition model that leverages both hard and soft constraints to regulate the evolving truth transitions in social media sensing applications.

Finally, one should note that the CA-DTD scheme is significantly different from the author's previous works on truth discovery problem [3], [6], [21]. In particular, we developed a robust truth discovery scheme (RTD) to discover truthful information in online social media [3]. However, the RTD solution is static and did not consider evolving truth of measured variables. We also developed a scalable streaming truth discovery (SSTD) scheme to address the dynamic truth discovery problem [6]. However, the SSTD scheme did not consider the physical constraints on measured variables in the truth discovery model and is limited in its capability to handle incomplete and noisy data from social media sensing applications. In sharp contrast to the above works, the CA-DTD explicitly address physical constraint awareness and incomplete and noisy data challenges in dynamic truth discovery problem. We compare the performance of CA-DTD, RTD and SSTD in Section V.

III. Problem Statement

In this section, we present the dynamic truth discovery problem in social media sensing application. In particular, let us consider a social media sensing application where a group of M *primary sources* $S = (S_1, S_2, ..., S_M)$ report the true values of a set of N *measured variables* $MV = (MV_1, MV_2, ..., MV_N)$. Let C represent the set of all claims made by sources and $C_i^{u,t}$ represent the claim made by source S_i about the true value of the measured variable MV_u at time t. We refer the *primary sources* to the social sensors (i.e., social media users) who are the primary data contributors in a social media sensing application. A *measured variable* represents an entity, topic or event that is of interest to the application and a *claim* is a statement on the truth of a measured variable.

Consider a Twitter-based social media sensing application where observations from common citizens are used to obtain the real-time situation awareness of a disaster event (e.g., terrorist attacks, hurricanes). The primary sources are the Twitter users. A measured variable can be the number of casualties in the disaster and a tweet such as "3 people died" is considered as a claim that reports the possible true value of the measured variable.

In this paper, we assume the true value of a measured variable changes over time (i.e., evolving truth). We define $V^{(u)} = (V_1^{(u)}, V_2^{(u)}, ..., V_K^{(u)})$ as the set of all possible true values of the u-th measured variable where K denotes the size of $V^{(u)}$. We assume there only exists one true value of a measured variable at a particular time instant. We define the estimated truth of the u-th measured variable at time t as $\hat{x}^{u,t}$ where $\hat{x}^{u,t} \in V^{(u)}$. We denote the ground truth label of measured variable MV_u at time t as $x^{u,t}$.

In order to address the physical-constraint awareness, we explicitly consider the physical constraints that are associated with each measured variable. In particular, we define two types of constraints in our model:

DEFINITION 1: Hard Constraints: a set of constraints that strictly enforce or prohibit some true value transitions at a certain time instant.

DEFINITION 2: Soft Constraints: a set of constraints that do not enforce or prohibit state transitions but describe how "likely" the true value of a measured variable can evolve at a certain time instant.

For example, a hard constraint can be "the number of casualties cannot decrease at any time" and a soft constraint can be "it is hard for suspects to travel 200 miles within half an hour". We use $\Omega^{(u)} = (\Omega_1^{(u)}, \Omega_2^{(u)}, ..., \Omega_Z^{(u)})$ to denote all constraints on MV_u

To address the incomplete and noisy data challenge, we introduce a group of Y *complementary sources* (e.g., traditional news media), $S_E = (S_{E1}, S_{E2}, ..., S_{EY})$, who provide a set of additional claims F on the measured variables. A claim from the complementary source is denoted as $F_{Ey,k}^{u,t}$, which indicates whether the complementary source S_{Ey} claims $V_k^{(u)}$ to be the true value of MV_u or not at time t. In the disaster event example, the complementary sources can be the traditional news media, first responders, or third-party agencies.

One should note that primary and complementary sources complement each other in many ways (e.g., news freshness and coverage). However, neither of them are perfectly reliable. For example, during the Boston Marathon event, mainstream press repeatedly reported false news and premature conjectures. In particular, CNN, Associated Press, Fox and the Boston Herald all reported that an arrest had been made soon after the bombing event which turned out to be a false story [22]. On the other hand, online social media is known to suffer from widespread rumors, misinformation, and spams due to the open and unvetted nature of data contribution paradigm [3]. Therefore, we do not assume the data collected from either primary sources or complementary sources are completely reliable in our CA-DTD scheme. Instead, the CA-DTD scheme explores the heterogeneity of data sources in a fusion process to conquer the incomplete and noisy data challenge.

Finally, we formally define the problem of dynamic truth discovery as follows: given the claims contributed by both primary and complementary sources, the objective is to correctly estimate the true value of measured variables at each time instant. In particular, for each measured variable MV_u at each t, our goal is to derive the estimation $\hat{x}^{u,t}$ to be as close to the ground truth $x^{u,t}$ as possible, which is given by:

$$\underset{\hat{x}^{u,t}}{\arg\max}\, P(\hat{x}^{u,t} = x^{u,t} | S, C, MV, S_E, F, \Omega), \forall 1 \leq u \leq N$$

IV. Solution

In this section, we present the Constraint-Aware Dynamic Truth Discovery (CA-DTD) scheme to solve dynamic truth discovery problem formulated in the previous section. We first present an overview of the CA-DTD scheme and then explain its components in detail.

A. Overview of CA-DTD Scheme

The CA-DTD scheme consists of two key components: Constraint-Aware Hidden Markov Model (CA-HMM) and Complimentary Source Incorporation (CSI). First, the CA-HMM model is designed to capture the evolving truth on measured variables based on observations from the social sensors. It addresses the physical constraint awareness challenge by explicitly incorporating the physical constraints into state transition rules which regulate a Viterbi decoding process. Second, the CSI component is integrated with the

CA-HMM component to incorporate the claims from complementary data sources to address the incomplete and noisy data challenge. The complementary sources can provide additional information that may not be directly observed by the primary sources on social media. It provides an opportunity to further improve the truth discovery results when the data is incomplete and the source reliability is unknown *a priori*. An overview of the CA-DTD scheme is illustrated in Figure 1.

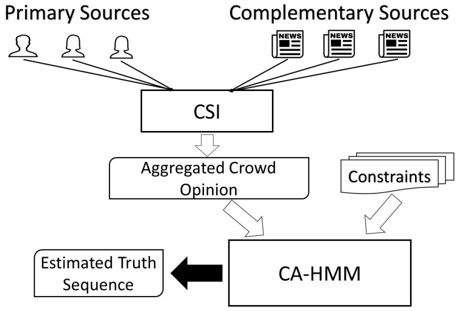


Figure 1: Overview of CA-DTD Scheme

B. Constraint-Aware Hidden Markov Model (CA-HMM)

In this subsection, we present the Constraint-Aware Hidden Markov Model (CA-HMM) component of CA-DTD scheme in detail. For the ease of reference, we first summarize all defined notations of CA-DTD in Table I. Some of the terms have been defined in Section III and the rest of them will be defined in this section. For ease of notation, we omit the index for measured variables (i.e. u) in this subsection. A Hidden Markov Model (HMM) is a stochastic state transition model that is commonly used to model systems with unobserved (hidden) states given a set of observation symbols [23]. The HMM is particularly suitable to handle streaming data and capture the dynamics of the state transition in real time. In particular, given a series of observations, it can decode the hidden states that generate those observations at each time instant. In CA-HMM, we define our observation sequence and hidden states as follows.

1) Aggregated Crowd Opinion and Hidden States of Evolving Truth: In CA-HMM, we take the evolving truth of a measured variable as the hidden states. Formally, we define the hidden states of truth as:

DEFINITION 3: Hidden State of Truth: the true value of the measured variable at a given time instant that is not directly observable.

The inference of hidden states (i.e. the evolving truth) requires a visible *observation sequence* that we can directly observe from data sources. In CA-HMM, we define our observations as Aggregated Crowd Opinion (ACO):

DEFINITION 4: Aggregated Crowd Opinion (ACO): the aggregated sources' opinion (from both primary and complementary sources) on the true values of the measured variables.

Table I: Definition and Notation

Symbol	Description
$x^{u,t}$	hidden truth as time t for measured variable MV_u
$\hat{x}^{u,t}$	estimated truth as time t for MV_u
$X^{(u)}$	state sequence of true values for MV_u, i.e. $(x^{u,1}$, $x^{u,2}, ..., x^{u,T})$
$V_k^{(u)}$	the k^{th} true value of MV_u
K	total number of true values for a measured variable
$ACO^{u,t}$	aggregated crowd opinion at time t for MV_u
$Obs^{(u)}$	observation sequence of crowd opinions, i.e. $(ACO^{u,1}$, $ACO^{u,2}, ..., ACO^{u,T})$
$\Omega_z^{(u)}$	the z^{th} physical constraint for MV_u
$F_{Ey,k}^{u,t}$	a claim from the complementary source S_{Ey} on whether $V_k^{(u)}$ the true value at time t
$w_i^{u,t}$	the trust score of the claim $C_i^{u,t}$ from S_i at time t
$CS_{i,k}^{u,t}$	the contribution score of S_i to estimate the true value of MV_u to be $V_k^{(u)}$ at time t

The key idea of the CA-HMM is to infer the hidden evolving truth based on the observation sequence (i.e., ACO). For each measured variable, we use the observations related with the variable (i.e., Obs = $(ACO^1, ACO^2, ..., ACO^T)$) as the input for the CA-HMM model and the output of the model is the corresponding sequence of estimated truth ($\hat{x}^1$, $\hat{x}^2, ..., \hat{x}^T$). The computation of the ACO from both primary and complementary sources are discussed in detail in the CSI component in Subsection IV-C.

2) Physical Constraints and State Transition: The CA-HMM model is defined by a set of key parameters that model the transition of truth and physical constraints.

- Truth Transition Probability Matrix A - each element $a_{i,j}$ is the probability that the true value transits from value V_i to value V_j.
- Emission Probability Matrix B - each element $b_{i,t}$ denotes the probability of observing ACO^t when the true value is V_i.
- Initial State Distribution π - each element π_i denote the probability the initial true value is V_i.
- Truth Transition Constraints Ω - a set of physical constraints that govern the state transitions based on physical laws or prior knowledge.
- Truth Transition Hardness Matrix HA^z - each element $ha_{i,j}^z \in [0, 1]$ denotes the hardness of true value of a measured variable transits from value V_i to value V_j given a constraint Ω_z.

In the above parameter definitions, we explicitly incorporate the physical constraints into our CA-HMM model, i.e., Truth Transition Constraints Ω and the Truth Transition Hardness HA^z. In particular, we further define a set of physical constraints as follows:

DEFINITION 5: Order Constraints (Hard): the true values of measured variables can only be in a certain order.

DEFINITION 6: Global Path Constraints (Hard): some *states are not reachable at a certain time instant.*

DEFINITION 7: Frequency Constraints (Soft): the frequency of state transitions may not exceed a certain limit.

DEFINITION 8: Spatial-Temporal Constraints (Soft): the state transitions should follow constraints imposed by space and time.

Consider the disaster event (e.g., hurricane) we discussed in Section III. An example of the *Order Constraints* can be "the number of casualties" can only be in non-decreasing order. *Global Path Constraints* can be the "current weather condition during the hurricane cannot be snow or blizzard". *Frequency Constraints* can be the true value of "whether a university campus is flooded" is less likely to change more than three times within a couple of days. Finally, *Spatio-Temporal Constraints* can be "the affected area of the hurricane" should be within a reasonable physical range given a certain time period.

The above constraints are often defined based on common sense or prior knowledge of the event. We proposed a generic framework that incorporates these constraints as plug-in functions that can be easily included or excluded in different social media sensing applications. In particular, we use the State Transition Hardness matrix HA^z to define how difficult the transition between two values of a measured variable can happen under each constraint Ω_z. Each element of matrix $ha^z_{i,j}$ is a hardness function to describe the specific constraint. For example, the hardness function for *Spatial-Temporal Constraints* in the hurricane example can be:

$$ha^z_{i,j} = \frac{1}{1 + exp(-\frac{dist(i,j)-thres)}{\tau})} \tag{1}$$

where the likelihood of the affected area of the hurricane changes from location i to j decreases (controlled by a parameter τ) when the distance between the two places exceeds a certain threshold.

3) Decoding Hidden States Via Constrained Decoding Algorithm: We then introduce our Constrained Decoding Algorithm to decode the true value of a measured variable at each time instant. The goal of this decoding step is to find the sequence of true values that is most likely to generate the observed ACO sequence given physical constraints. Formally, it is given by:

$$(\hat{x}^1, \hat{x}^2, ..., \hat{x}^T) = \arg\max_X P((ACO^1, ACO^2, ..., ACO^T)|X, \Omega) \tag{2}$$

where $X = (x^1, x^2, ..., x^T)$ is the hidden sequence of the true values of a measured variable.

While the Viterbi Algorithm [24] is a standard approach to solve the decoding problem in HMM, it does not consider the physical constraints that regulate the truth transitions of measured variables. In our CA-HMM, we extend the original Viterbi algorithm to incorporate the physical constraints. In particular, we define a *accumulative hardness score* $AHC^{z,t}_j$ for each true value at each time instant. This score describes how difficult the transition is if the true value of a measured variable changes to V_j. Formally, we calculate:

$$AHC^{z,t}_j = \begin{cases} AHC^{z,t-1}_j + \max\limits_{1\le i\le K} ha^z_{i,j}, \sum\limits_{i=1}^{K} ha^z_{i,j} \neq 0 \\ 0, \sum\limits_{i=1}^{K} ha^z_{i,j} = 0 \end{cases} \tag{3}$$

The idea is to keep track of the hardness of each truth transition of a measured variable and check whether the accumulated hardness score breaks any of the physical constraints. The intuition is that consecutive "hard" transitions will eventually lead to an unreasonable accumulated hardness score that makes such transition impossible. We clean up the accumulated hardness score when no constraint exists (i.e., $\sum_{i=1}^{K} ha^z_{i,j} = 0$).

To decode true value sequence, we follow the dynamic programming procedure of Viterbi decoding. At each time instant t and for each true value V_i, we calculate:

$$\delta^t(i) = \max_{x^1,x^2,...,x^{t-1}} P(x^1, x^2, ..., x^{t-1}, ACO^1, ACO^2, ..., ACO^t, x^t = V_i|\lambda, \Omega) \tag{4}$$

where $\delta^t(i)$ represents the probability that the HMM's current true value is V_i after seeing the first t observations and passing through the most likely true value sequence $x^1, ..., x^{t-1}$. Given the estimated parameter set λ, it can be solved recursively as:

$$\delta^{t+1}(i) = b_{i,t} \times \max_{1\le i\le K} \delta^t(i) \times a_{i,j} \tag{5}$$

where $1 \le j \le K$ and $1 \le t \le T$-1. The initialization is: $\delta^{t+1}(i) = \pi_i \times b_{i,1}, 1 \le i \le K$. To impose physical constraints, we compare the ACH score associated with each constrain and check whether the transition is allowed. We prohibit the invalid transition by setting $\delta^t(i) = -\infty$. This constrained decoding procedure is shown in Algorithm 1.

The training phase of our CA-HMM model follows the standard unsupervised Baum Welch algorithm [25]. Specifically, we find the set of parameters that maximize the probability of the observed sequence of crowd opinions for each measured variable:

$$\lambda^* = \arg\max_\lambda P(\{ACO^1, ACO^2, ..., ACO^T\}|\lambda) \tag{6}$$

C. Complimentary Source Incorporation (CSI)

In this subsection, we describe the Complimentary Source Incorporation (CSI) component to address the incomplete and noisy data challenge. In particular, the CSI component fuses data from both primary sources and complementary sources and the fused results are integrated the CA-HMM component discussed above.

One key challenge in social media sensing applications lies in the fact that sources are often unvetted and they

Algorithm 1 Constrained Decoding Algorithm

```
Initialize: create path probability matrix $\delta[K+2,T]$, accumulated hardness array $accuha[K,Z,T]$, back pointer array $backpointer[K+2,T]$ and valid states array $NPS[K,T]$ for each $MV_u$
for all $i, 1 \leq i \leq K$ do                                   ▷ Initialization Step
    $\delta[i,1] \leftarrow \pi_i^{(u)} \times b_{i,1}^{(u)}$,
    $backpointer[i,1] \leftarrow 0$, $accuha[j,z,1] \leftarrow 0$
end for
for all $t, 2 \leq t \leq T$ do
    for all $z, 1 \leq z \leq Z$ do
        while $accuha[i,z,t-1] + ha_{i,j}^{z} \geq 1$ do
            prune transition $i->j$ at time t                      ▷ Imposing Constraints
        end while
        calculate $accuha[j,z,t]$ using Equation (3).
        $NPS[i,t] \leftarrow \{j|1 \leq j \leq K, i->j\ not\ pruned\}$
        if $NPS[i,t] = \emptyset$ then
            $\delta[i,t] \leftarrow -\infty$
        else
            calculate δ[i,t] based on Equation (4), $i \in NPS[i,t]$
            backpointer[i,t] ← $\arg\max_{i \in NPS[i,t]} \delta^{u,t}(i) \times a_{i,j}^{(u)}$
        end if
    end for
end for
$\delta[end,T] \leftarrow \max_{1 \leq i \leq K} \delta^{u,T}(i) \times a_{i,j}^{(u)}$       ▷ Termination Step
$backpointer[end,T] \leftarrow \arg\max_{1 \leq i \leq K} \delta^{u,t}(i) \times a_{i,j}^{(u)}$
return the backtrace path by following backpointers from backpointer[end,T]
```

may not always report truthful claims. Therefore, we need to explicitly model the quality of data sources. In our model, we define a *contribution score* for each source to represents how much a source contributes to the belief that indicates the true value of a measured variable. We first define the following terms that are related to the contribution score.

DEFINITION 9: Coherence Score($\rho_i^{u,t}$) : a score in the range of (0,1) that measures the relevance of a claim $C_i^{u,t}$ to its corresponding measured variable MV_u. A higher score is assigned to a claim that is more relevant.

DEFINITION 10: Uncertainty Score ($\kappa_i^{u,t}$): a score in the range of (0,1) that measures the uncertainty of a claim $C_i^{u,t}$. A higher score is assigned to a claim that expresses more uncertainty.

DEFINITION 11: Independent Score: ($\eta_i^{u,t}$): a score in the range of (0,1) that measures whether the claim $C_i^{u,t}$ is made independently or copied from others. A higher score is assigned to a claim that is more likely to be made independently.

We define the above scores to identify important features of both sources and claims that are important in the dynamic truth discovery solution. We then define *trust score* of a claim to represent how much we could "trust" the claim considering the above features (i.e., coherence, uncertainty and independence) of the claim. It is formally defined as:

$$w_i^{u,t} = \rho_i^{u,t} \times (1 - \kappa_i^{u,t}) \times \eta_i^{u,t} \tag{7}$$

Using the trust score of a claim, we define the *contribution score* of a source S_i on claim $C_i^{u,t}$ for a measured variable at time t as:

$$CS_{i,k}^{u,t} = w_i^{u,t} \times D_{i,k}^{u,t} \tag{8}$$

where $D_{i,k}^{u,t}$ is the *Source Attitude*, which is a binary variable that represents whether a source "agrees with" or "disagrees with" the statement $\hat{x}^{u,t} = V_k^{(u)}$ on claim $C_i^{u,t}$. It's formally defined as:

$$D_{i,k}^{u,t} = \begin{cases} 1, S_i \text{ aggrees with } \hat{x}^{u,t} = V_k^{(u)} \\ -1, S_i \text{ disaggrees with } \hat{x}^{u,t} = V_k^{(u)} \end{cases} \tag{9}$$

The trust score measures how trustworthy a claim is, and the source attitude denotes whether the source agrees with the assertion of the claim. Therefore, the *contribution score* represents how much source S_i contributes to the belief that the true value of a measured variable at time t is $V_k^{(u)}$.

Another important problem of fusing heterogeneous data sources is to identify the "weight" of each source. We define the authority weights of primary and complementary sources as follows.

DEFINITION 12: Authority Weight: the score that represents the extent of expertise of each source in reporting truthful information. A more expertized source has a higher authority weight.

Given the authority scores for primary and complementary sources, the ACO for a measured variable at time instant t can be calculated as:

$$\begin{aligned} ACO^{u,t} &= \arg\max_{k \in K} \Delta W_I^t + \sum_{1 \leq y \leq Y} W_{Ey}^t \times F_{Ey,k}^{u,t} \\ \Delta &= normalize(\sum_{i=1}^{M} CS_{i,k}^{u,t}) \end{aligned} \tag{10}$$

where $CS_{i,k}^{u,t}$ is the contribution score of source S_i on claim $C_i^{u,t}$ which is defined in Equation (8). W_I^t and W_{Ey}^t denote the authority weight of the primary sources and y-th complementary source respectively at time t. In this paper, we focus on a particular type of social media (e.g., Twitter) as the primary sources and use a unified authority score for the primary sources as a whole [1]. In contrast, the complimentary sources may consist of various entities (e.g., news websites, radio stations, third party agencies) and we define an authority score for each of them to respect their diversity. Δ normalizes the aggregated contributions scores of primary sources to a scale between -1 and 1.

We compute the W_I^t and W_{Ey}^t based on the proportion of claims that are consistent with the estimated truth of the measured variables that are obtained from the CA-HMM discussed in Section IV-B:

$$\begin{aligned} W_I^t &= \frac{\sum_{u \in MV(I)} \chi_I(u,t)}{|MV(I)|} \\ W_{Ey}^t &= \frac{\sum_{u \in MV(Ey)} \chi_E(u,t)}{|MV(Ey)|} \end{aligned} \tag{11}$$

[1] Each individual primary source still has it's unique contribution score.

$$\chi_I(u,t) = \begin{cases} 1,\ \hat{x}^{u,t} = V_k^{(u)}\ and\ k = \underset{k' \in K}{\arg\max} \sum_{i=1}^{M} CS_{i,k'}^{u,t} \\ 0,\ \hat{x}^{u,t} = V_k^{(u)}\ and\ k \neq \underset{k' \in K}{\arg\max} \sum_{i=1}^{M} CS_{i,k'}^{u,t} \end{cases}$$

$$\chi_E(u,t) = \begin{cases} 1,\ \hat{x}^{u,t} = V_k^{(u)}\ and\ F_{Ey,k}^{u,t} = 1 \\ 0,\ \hat{x}^{u,t} = V_k^{(u)}\ and\ F_{Ey,k}^{u,t} \neq 1 \end{cases}$$

where $MV(I)$ denotes all the measured variables that primary sources S contribute claims to and $MV(Ey)$ denotes all the measured variables that complementary source S_{Ey} contribute claims to. $\chi_I(u,t)$ and $\chi_E(u,t)$ indicate whether the aggregated opinions from primary sources and claims from complementary sources are consistent with the estimated truth, respectively. The intuition of Equation (11) is that a source should have higher authority score if its opinion agrees with the estimated truth.

Finally, the CA-DTD scheme is summarized in Algorithm 2. The convergence analysis of Algorithm 2 is shown in the evaluation section.

Algorithm 2 Constraint-Aware Dynamic Truth Discovery (CA-DTD) Scheme

```
Initialize $W_{Ey}$ for each complementary source and $W_I$ for primary sources.
while $\{W_{Ey}\}$ and $\{W_I\}$ do not converge do
    for all $u, 1 \leq u \leq N$ do
        for all $i, 1 \leq i \leq M$ do
            for all $k, 1 \leq k \leq K$ do
                compute $CS_{i,k}^{u,t}$ based on Equation (8).
            end for
        end for
        compute $ACO^{u,t}$ based on Equation (10).
    end for
    re-estimate $A^{(u)}, B^{(u)}, \lambda^{(u)}$ using the EM algorithm based on Equation (6).
    for all $u, 1 \leq u \leq N$ do
        estimate $\hat{x}^{u,t}$ using Constrained Decoding Algorithm.
    end for
    update $W_{Ey}$ and $W_I$ based on Equation (11).
end while
```

V. Evaluation

In this section, we evaluate the performance of the CA-DTD scheme and compare it with the state-of-the-art truth discovery baselines on two real-world datasets collected from social media sensing applications. The results demonstrate that the CA-DTD scheme significantly outperforms all compared baselines in terms of both truth discovery accuracy and computational efficiency.

A. Experimental Setups

1) Baseline Methods:

- **TruthFinder:** It uses a pseudo-probabilistic function to estimate source reliability and claim truthfulness using an iterative algorithm [4].
- **CATD, ETCIBoot:** These two methods provide confidence interval estimators for source reliability in a sparse dataset [11]. ETCIBoot improves CATD by incorporating a bootstrapping technique.
- **RTD:** A truth discovery algorithm designed for inferring truthful information on online social media. The algorithm is shown to be robust against rumors and misinformation spread [3].
- **SSTD:** A dynamic truth discovery scheme that uses a simple Hidden Markov Model to capture the evolving truth of the measured variables in social sensing applications [6].
- **Recursive EM:** It applies a recursive EM algorithm to capture the time-varying truth in streaming data [7].

We note that the first four baselines (i.e., Truth Finder, CATD, ETCIBoot, RTD) are batch-based algorithms that are designed to solve the static truth discovery problem. To run these baselines on dynamic data traces, we treat the same measured variable with different values as different variables and run the batch algorithms on the whole datasets. In contrast, CA-DTD, SSTD, and Recursive EM are designed to solve the dynamic truth discovery problem and we compare them with other baselines under the same experimental setting. *To make our comparison fair between baselines, the input data to all compared schemes are the same. In particular, we treat the complementary source (e.g., traditional news media) as an additional data source to all baselines.*

2) Evaluation Metrics: To evaluate the performance of all schemes, we use the following metrics: *Accuracy*, *Precision*, *Recall* and *F1-Score*. Their definitions are given in Table II. In the table, TP_j, TN_j, FP_j and FN_j represents True Positives, True Negatives, False Positives and False Negatives respectively for one possible true value j of a measured variable and L denotes the set of all possible true values. To evaluate the efficiency, we also report the execution time of all compared schemes in our evaluation.

Table II: Evaluation Metrics

Accuracy	$\frac{\sum_{j \in L} TP_j + TN_j}{\sum_{j \in L} TP_j + TN_j + FN_j + FP_j}$
Precision	$\frac{\sum_{j \in L} TP_j}{\sum_{j \in L} TP_j + FP_j}$
Recall	$\frac{\sum_{j \in L} TP_j}{\sum_{j \in L} TP_j + FN_j}$
F1-Score	$\frac{2 * Precision * Recall}{Pecision + Recall}$

3) Data Collection and Pre-Processing: We evaluate our proposed scheme on two real-world data traces collected from Twitter. We found that there exists a non-trivial amount of widely spread misinformation, spams, and noisy data on Twitter due to the open data collection environment and unvetted nature of data sources [26]. Moreover, few users on Twitter contribute many tweets towards a given topic, which leads to a social media sensing scenario with incomplete data. The above characteristics of Twitter provide us a good opportunity to investigate the performance of the CA-DTD scheme in a real world setting. We present the details of the two data traces[2] as follows (Table III).

Data Trace	Boston Bombing	Hurricane Matthew
Start Date	Apr. 15 2013	Oct. 6 2016
Time Duration	4 days	14 days
# of Sources	64,381	125,932
# of Claims (Tweets)	73,331	247,313
# of Claims per Source	1.14	1.96

Table III: Data Trace Statistics

We also observe the majority of the measured variables are non-static in the data traces (Figure 2). For example, the truth values of 72% measured variables in Boston Bombing dataset and 75% measured variables in Hurricane Matthew dataset evolve at least once during the data collection period, which provides good case studies to study the performance of CA-DTD and baselines in a dynamic setting.

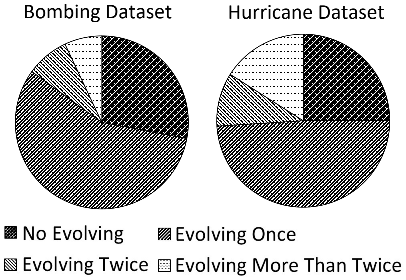


Figure 2: Frequency of Truth Evolving

Boston Bombing Trace: We collected Twitter data related to the 2013 Boston Bombing event through Twitter open search API (using the query terms "Boston", "Marathon", "Bombing") and specified geographic regions related to the event (using a circular region centered at Boston with a radius of 100 miles). We use the mainstream news media as the complementary sources in this case study. Since it's a historical event, we built a crawler to search for top news reports that are relevant to bombing event using the Google Search's customized time frame feature. We collected 78 reports from six major news medias (New York Times, CNN, Fox News, Washington Post, USNews, and BuzzFeed) during the event.

Hurricane Matthew Trace: We collected Twitter data related to the Hurricane Matthew event in 2016, which was a deadly and destructive hurricane from the Caribbean to the United States. We use the same API to collect the tweets with query terms "Hurricane", "Matthew". We set a geographic region that covers all United States cities. For complementary sources, we collected 150 reports from the same news medias.

Data Pre-processing: we first group tweets with similar contents into the same cluster using a variant of K-means clustering algorithm and a distance metric that is commonly used for Twitter data (i.e., Jaccard distance) [27]. We treat a topic directly related to the event of interest in each cluster as a measured variable (e.g., the location of the suspect, number of casualties during the hurricane) and tweets discussing the same topic are considered as claims associated with the corresponding measured variable.

In order to compute the *trust score* of claims, we first derive the *Coherence Score* by calculating the content similarity between a claim and its corresponding measured variable using the Jaccard distance. We then calculate the *Uncertainty Score* by implementing a simple text classifier using skit-learn and trained it with the training data provided by CoNLL-2010 Shared Task [28]. To compute the *Independent Score*, we classified the retweets or tweets that are significantly similar to the previous tweets as repeated claims and assign them relatively low independent scores. To label *Source Attitude*, we applied a method that classifies a tweet as "agree" or "disagree" based on its content (e.g., whether a tweet includes certain negative words such as "fake", "false", "not true", "debunked", "rumor"). Finally, we divide the data traces into *one-hour* intervals based on the timestamps of the tweets and get 131 time intervals for Boston Bombing and 335 for Hurricane Matthew respectively. The time interval is chosen as a trade-off between the frequency of the truth discovery updates and the amount of data in each time interval.

Labeling Ground Truth: Since both datasets are based on historical events and rumors and myths have been revealed over time, we obtain the ground truth of the measured variables based on credible post-event reports. In particular, we hired three individuals to manually look up facts about each claim from credible sources and reconcile our collected facts. We divided the claims at each time interval into fact-checkable and non-fact-checkable. For example, a claim "The Boston Bombing suspect was seen to escape from Stata Center, M.I.T" is considered as non-fact-checkable because we cannot identify where the suspects have been at that time. We excluded all "non-fact-checkable" claims from evaluation.

B. Experiment Results

We report the results of the Boston Bombing dataset in Table IV. Observe that the CA-DTD scheme performs the best among all compared truth discovery schemes. For example, the CA-DTD scheme outperforms the best-performed baseline by 8.5%, 7.8%, 2.8% and 5.6% on accuracy, precision, recall and F1-Score respectively. The performance gains of CA-DTD scheme are mainly achieved by (i) explicitly modeling the evolving truth of measured variables using CA-HMM, (ii) incorporating physical constraints to regulate the truth transitions of measured variables, and (iii) seamlessly integrating information from both online social media and traditional news media to handle incomplete and noisy

[2] http://apollo.cse.nd.edu/

social sensing data. The results of the Hurricane Matthew data trace are shown in Table V. We observe that CA-DTD continues to achieve the best performance among all compared schemes. For example, CA-DTD outperforms the best-performed baseline by 4.5%, 7.5%, 1.7% and 5.0% on accuracy, precision, recall and F1-Score respectively.

Table IV: Results on Boston Bombing Data Trace

Method	Accuracy	Precision	Recall	F1-Score
CA-DTD	**0.906**	**0.928**	**0.889**	**0.909**
SSTD	0.821	0.850	0.858	0.853
Recursive EM	0.754	0.822	0.793	0.807
TruthFinder	0.770	0.797	0.861	0.828
RTD	0.791	0.719	0.765	0.742
CATD	0.789	0.838	0.832	0.835
ETCIBoot	0.793	0.840	0.831	0.837

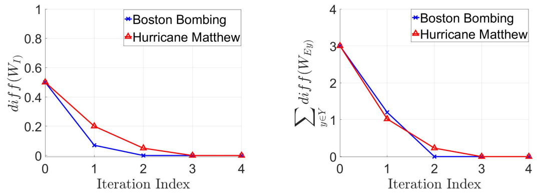
Table V: Results on Hurricane Matthew Data Trace

Method	Accuracy	Precision	Recall	F1-Score
CA-DTD	**0.903**	**0.918**	**0.833**	**0.873**
SSTD	0.858	0.831	0.816	0.823
Recursive EM	0.819	0.837	0.684	0.752
TruthFinder	0.746	0.702	0.642	0.671
RTD	0.818	0.843	0.674	0.749
CATD	0.781	0.709	0.774	0.741
ETCIBoot	0.793	0.822	0.620	0.707

We also evaluate the execution time of all compared schemes. In particular, we run our experiments on a desktop with an 8-core Intel i7 Processor and 16G of RAM. The running time results are reported in Table VI. We observe that the CA-DTD scheme outperforms all baselines including the dynamic schemes that handle the streaming data (i.e. SSTD, Recursive EM). The computational efficiency is mainly achieved by the CA-HMM where the physical constraints greatly reduce the dimensions of solution space. We also attribute the performance gain to the fast convergence of our CA-DTD scheme as shown in Figure 3. The y-axis is the difference of sources' authority scores W_{Ey}, W_I and the x-axis denotes the current iteration index while the 0-th iteration represents the initial assignments of the authority scores. We observe that CA-DTD converges after a couple of iterations.

Table VI: Evaluation on Running Time (Seconds)

Data Trace	Bombing	Hurricane
CA-DTD	**22.443**	**53.711**
SSTD	28.714	66.252
Recursive EM	37.913	78.918
TruthFinder	159.474	202.594
RTD	92.831	177.785
CATD	42.782	102.343
ETCIBoot	45.094	133.515



(a) Convergence of W_I

(b) Convergence of W_{Ey}

Figure 3: Convergence Analysis of CA-DTD

VI. Conclusion

This paper develops the CA-DTD scheme to solve the dynamic truth discovery problem in big data social media sensing applications. The CA-DTD scheme explicitly addresses two important challenges: physical constraint awareness and noisy and incomplete data. In particular, we develop a CA-HMM to model the dynamic states of the measured variables and incorporate physical constraints to regulate state transitions in the model. The CA-DTD also integrates a data fusion component into the CA-HMM to explore the data heterogeneity to improve the truth discovery accuracy. The evaluation results on two real world social media sensing data traces demonstrate that CA-DTD achieves significant performance gains compared to the state-of-the-art baselines in dynamic truth discovery. The results of this paper are important because they lay out a solid analytical foundation to address dynamic truth discovery by explicitly exploring physical constraints in social media sensing applications.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-17-1-0409. 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 Office 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] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, Apr. 2012, pp. 233–244.

[2] D. Wang, T. Abdelzaher, and L. Kaplan, *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[3] D. Y. Zhang, R. Han, D. Wang, and C. Huang, "On robust truth discovery in sparse social media sensing," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1076–1081.

[4] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, Jun. 2008.

[5] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, and Y. Zhang, "Large-scale point-of-interest category prediction using natural language processing models," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017.

[6] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[7] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 530–539.

[8] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, "A bayesian approach to discovering truth from conflicting sources for data integration," in *Proceedings of the VLDB Endowment*, vol. 5, no. 6, 2012, pp. 550–561.

[9] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu, "Exploitation of physical constraints for reliable social sensing," in *Real-Time Systems Symposium (RTSS), 2013 IEEE 34th*. IEEE, 2013, pp. 212–223.

[10] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu, "Reliable social sensing with physical constraints: analytic bounds and performance evaluation," *Real-Time Systems*, vol. 51, no. 6, pp. 724–762, 2015.

[11] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng, and A. Zhang, "Towards confidence in the truth: A bootstrapping based truth discovery approach," in *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.

[12] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: the role of source dependence," in *Proceedings of the VLDB Endowment*, 2009, pp. 550–561.

[13] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le, "Using humans as sensors: An estimation-theoretic perspective," in *Proc. 13th Int Information Processing in Sensor Networks Symp. IPSN-14*, Apr. 2014, pp. 35–46.

[14] D. Wang, T. Abdelzaher, H. Ahmadi, J. Pasternack, D. Roth, M. Gupta, J. Han, O. Fatemieh, H. Le, and C. C. Aggarwal, "On bayesian interpretation of fact-finding in information networks," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011, pp. 1–8.

[15] X. Yin and W. Tan, "Semi-supervised truth discovery," in *Proceedings of the 20th international conference on World wide web*, no. 217-226. ACM, 2011.

[16] D. Wang, L. Kaplan, and T. F. Abdelzaher, "Maximum likelihood analysis of conflicting observations in social sensing," *ACM Transactions on Sensor Networks*, vol. 10, no. 2, pp. 1–27, Jan. 2014.

[17] A. Pal, V. Rastogi, A. Machanavajjhala, and P. Bohannon, "Information integration over time in unreliable and uncertain environments," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 789–798.

[18] W.-G. Kim, J.-Y. Choi, and D. H. Youn, "Hmm with global path constraint in viterbi decoding for isolated word recognition," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–605.

[19] S. T. Roweis, "Constrained hidden markov models," in *Advances in Neural Information Processing Systems*, 2000, pp. 782–788.

[20] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic constraints in hidden markov model based speech recognition," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 699–702.

[21] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, "On credibility estimation tradeoffs in assured social sensing," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1026–1037, 2013.

[22] E. F. Davis, A. A. Alves, D. A. Sklansky *et al.*, "Social media and police leadership: Lessons from boston," *Australasian Policing*, vol. 6, no. 1, p. 10, 2014.

[23] S. R. Eddy, "Profile hidden markov models." *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[25] L. E. Baum, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[26] C. C. Aggarwal and T. Abdelzaher, "In managing and mining sensor data," *Springer Science & Business Media*, pp. 237–297, 2013.

[27] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference, Christchurch, New Zealand*, 2008, pp. 49–56.

[28] R. Farkas, V. Vincze, G.Mora, J. Csirik, and G.Szarvas, "The conll-2010 shared task: Learning to detect hedges and their scope in natural language text," in *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning.*, 2010.