12-2015

Intent Classification of Short-Text on Social Media

Hemant Purohit
Wright State University - Main Campus, purohit.5@wright.edu

Guozhu Dong
Wright State University - Main Campus, guozhu.dong@wright.edu

Valerie L. Shalin
Wright State University - Main Campus, valerie.shalin@wright.edu

Krishnaprasad Thirunarayan
Wright State University - Main Campus, t.k.prasad@wright.edu

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Intent Classification of Short-Text on Social Media

Hemant Purohit
Dept. of Information Sciences & Technology
VSE, George Mason University
Fairfax, USA
hpurohit@gmu.edu

Guozhu Dong, Valerie Shalin*, Krishnaprasad Thirunarayan, Amit Sheth
Dept. of CSE and Dept. of Psychology*
Kno.e.sis Center, Wright State University
Dayton, USA
{guozhu.dong,valerie.shalin,t.k.prasad,amit.sheth}@wright.edu

***Abstract*—Social media platforms facilitate the emergence of citizen communities that discuss real-world events. Their content reflects a variety of intent ranging from social good (e.g., volunteering to help) to commercial interest (e.g., criticizing product features). Hence, mining intent from social data can aid in filtering social media to support organizations, such as an emergency management unit for resource planning. However, effective intent mining is inherently challenging due to *ambiguity* in interpretation, and *sparsity* of relevant behaviors in social data. In this paper, we address the problem of multiclass classification of intent with a use-case of social data generated during crisis events. Our novel method exploits a hybrid feature representation created by combining top-down processing using knowledge-guided patterns with bottom-up processing using a bag-of-tokens model. We employ pattern-set creation from a variety of knowledge sources including psycholinguistics to tackle the *ambiguity* challenge, social behavior about conversations to enrich context, and contrast patterns to tackle the *sparsity* challenge. Our results show a significant absolute gain up to 7% in the F1 score relative to a baseline using bottom-up processing alone, within the popular multiclass frameworks of One-vs-One and One-vs-All. Intent mining can help design efficient cooperative information systems between citizens and organizations for serving organizational information needs.**

***Keywords*-Intent Mining; Social Media; Psycholinguistics; Declarative Knowledge; Contrast Mining; Crisis Informatics**

I. Introduction

Web 2.0 provides a natural platform for citizen communities to discuss real-world events. Using social media, citizens can readily generate massive amounts of data, share information and express opinions in discussions. As a result, organizations with limited resources are trying to incorporate information nuggets from citizen-generated data to enrich their decision making. For example, a crisis response organization requires relevant data to distribute resources effectively. Our premise is that intent mining provides insights that are not explicitly available in citizen-generated data.

Intent is defined as a purposeful action. We attribute intent to behaviors every day, from a user querying a search engine to buy a laptop to a user participating in a conversation to inform. Intent can help identify actionable information. Here we assess intent in social media regarding cooperation between citizens and organizations during real-world events.

Much prior work in intent mining addresses the challenge of understanding queries obtained from search engine logs [1], [2]. Search intent can be navigational, informational or transactional rather than for social communication. Our objective is to model intent related to cooperation in user-generated content. Therefore, our research question is “How can we mine relevant social intent from an ambiguous, unconstrained natural language short-text document?” By definition, the relevant intent classes meet actionable information needs of an organization in a given context, e.g., resource seeking during crisis response coordination [3], [4]. Table I shows examples of short-text documents and associated potential intent.

Short-text Document	Potential Intent
Text redcross to 90999 to donate $10 to help those people that were effected by hurricane sandy please donate #SandyHelp	Seeking help
Anyone know where the nearest #RedCross is? I wanna give blood today to help the victims of hurricane Sandy	Offering help
@Zuora wants to help @Network4Good with Hurricane Relief. Text SANDY to 80888 & donate $10 to @redcross @AmeriCares & @SalvationArmyUS #help	Seeking help
Would like to urge all citizens to make the proper preparations for Hurricane #Sandy - prep is key - http://t.co/LyCSprbk has valuable info!	Advising
Thx to all in Kettering who brought supplies for those affected by Hurricane Sandy. Visit http://t.co/IWSCVity to help. ...	Acknowledging

TABLE I
EXAMPLES OF SHORT-TEXT DOCUMENTS AND POTENTIAL INTENT

A variety of factors affect an individual’s expression of intentionality [5], [6], [7]. Multiple potential intentions complicate natural language interpretation in short-text documents. Therefore, to make the intent mining problem computationally tractable, we exploit a classification form of this problem for mining specific intent classes, and define a multiclass intent classification problem.

Intent classification (focused on future action) is a form of text classification. However it is different from the well-studied problems of topic classification [8] (focused on matter) as well as subjective text classification such as sentiment or emotion classification (focused on the current state of affairs) [9]. For instance, in a message “I wanna watch awesome Fast & Furious 7. Yh, Vin Disel is COOLest!!!”, topic classification focuses on the noun, the movie ‘Fast & Furious 7’; sentiment and emotion classification is focused on the positive feeling of the author’s message expressed with the adjective ‘awesome’. In contrast, intent classification concerns the author’s intended future action, i.e. going to watch the movie.

We observe two key challenges with intent classification on social media. First, informal language use causes *ambiguity* in interpreting user expressions in short-text messages, weakening predictor-class relationships (e.g., 'wanna help' appears as a strong intent signal but exists in messages of two complementary intent classes, 'seeking' and 'offering'). Second, *sparsity* of instances of specific intent classes in the corpus creates data imbalance (e.g., our prior study on binary intent classification [4] observed that expressions of 'offering' intent were only a fraction of those with 'requesting' intent (1:7 ratio) during Hurricane Sandy event in 2012). Furthermore, both intent classes of 'seeking' and 'offering' may co-occur within a single message.

Our novel method for intent classification exploits a rich feature representation for learning, created by integrating top-down processing using knowledge-guided patterns with bottom-up processing using a bag-of-tokens model. We experiment with pattern sets from a variety of knowledge sources including psycholinguistics, social behavior in conversations, and contrast mining-guided patterns (see Section IV). Our experimental datasets focus on the context of cooperation during crisis. Therefore, the relevant intent classes for organizational tasks of resource prioritization are $\{$'*seeking*','*offering*'$\}$, an abstraction known to help avoid a second crisis of resource management (as reported by NPR after the Hurricane Sandy[1]). Our specific contributions are the following:

- To our knowledge, this is the first study of intent classification on social media for crisis event datasets in both the popular frameworks of multiclass classification, One-vs-One (OVO) and One-vs-All (OVA).
- We demonstrate the power of integrating different knowledge-guided patterns into the traditional text mining approach of bag-of-tokens model for improving intent classification performance on unconstrained, short natural language text (absolute gain in F1 score up to 7%).
- We show the need for integrating social behavioral knowledge into the analysis of intent in a social context, such as dialogue management indicators (e.g., tx, anyway) that are often removed as *stopwords* in traditional text mining tasks.

Below we discuss related work in Section II, a problem statement in Section III, and a description of our approach in Section IV. Section V describes the experimental setting, and Section VI discusses results, limitations and future work.

II. Related Work

Work related to intent classification amalgamates multiple issues involving data, domain and problem variants as follows:

Search log data. Researchers have designed approaches to mine intent in queries using data from user search logs, including clicks, click sequence graphs and query terms, with broadly identified content categories such as navigational, informational and transactional [1], [2]. A major limitation of these approaches for our problem context is the dependence on (an unavailable) large data set of user behavior.

Well-formed text data. Prior work spans varying problem areas including analysis of presidential speeches [9], and product reviews [10], [11]. In contrast with the short-text documents of platforms like Twitter, reviews and large text documents provide more explicit information about the applicable context of intent, and typically comply with syntactic structure that enables the success of established methods of Natural Language Processing.

Short-text social data. Earlier research focused on mining transactional intent consistent with commercial interests [12], [13], [14]. The limited motives pertain to the transactional intent of buying and selling, which are different from the critical actions involved in our problem context of cooperation. Therefore, the nature of other kinds of complex intent (such as the broad intent class of helping) requires more thorough investigation. The closest works on crisis data analytics on Twitter has dealt with the identification of problems-aid report [3], and request-offer messages [4], [15] using only binary classifiers.

Classification problem variants. Prior research on short-text social media has mainly focused on binary intent classification in both commercial and crisis domains due to the complexity of intent prediction from noisy text. Multiclass classification remains an open and challenging problem. Optimizing multiclass classification depends on the data and problem domain. Therefore, researchers have studied mainly two different learning schemes, a.) A standalone multiclass learner, and b.) Binarization to enable the use of multiple binary (base) learners, followed by their combination [16], [17]. In the standalone multiclass learner, the complexity of simultaneously learning decision boundaries for a large number of classes is a major challenge, while the binarization method simplifies learning due to only the two-class nature for the base learners. Furthermore, binarization can be parallelized for addressing scalability concerns. The most popular schemes for the binarization framework are decomposition based—OVO and OVA. OVO creates a base learner for each class pair (${}^{K}C_2$ learners for K classes). On the other hand, OVA creates a base learner for each class (K learners for K classes), by considering the target class instances in a positive training set, and instances of remaining classes in the negative set. Although these approaches have been investigated on the UCI gold standard datasets [16], [18], within the context of intent classification on social media, these schemes are yet to be examined.

III. Problem Statement and Hypotheses

We define a Citizen Community CC as a group of users on social media who participate in discussions of a real world event by posting relevant short-text messages (documents). A relevant short-text document m_i is defined as a social media message containing any event-related keyword. Event-related keywords are manually provided while crawling the dataset from social media.

[1]http://www.npr.org/2013/01/09/168946170/thanks-but-no-thanks-when-post-disaster-donations-overwhelm

Problem Statement:
Given a corpus A of n short-text documents m_i generated in a citizen community CC, $A = \{m_i \mid 1 \leq i \leq n\}$, and a set of K intent classes, $C = \{C_j \mid 1 \leq j \leq K\}$; predict an intent class in C for each $m_i \in A$.

Our experiments on the crisis event datasets study a class set of {*seeking*, *offering*, *none (neither seeking nor offering)*}. We propose to evaluate the following two hypotheses:

H1. Psycholinguistic knowledge can guide the design of semantic-syntactic patterns to enrich the expressivity of informational context for intent classification.

H2. Intent classification can be improved by fusing top-down knowledge-guided patterns with bottom-up frequency-based representation in noisy social data.

IV. Approach

We first discuss the data collection method, followed by rich feature design, class labeling process, and learning method.

A. Data Collection

Using a keyword-based crawling method, we collected a set of short-text documents as *tweets* from the Twitter Streaming API like prior studies [4]. For a real-world event, we defined a set of relevant keywords. For each keyword *k*, the Twitter API provided tweets containing any of the form: #*k*, #*K*, *k*, and *K*.

B. Rich Feature Design

We investigate three approaches for feature representation. Approach v1 is to represent the feature space using only bottom-up processing, which exploits the implicit semantics of the local content of a document via an unordered bag-of-tokens representation. On the other hand, approach v2 is to represent the feature space generated by top-down processing, which exploits the semantics acquired outside the context of the local content via a set-of-patterns derived using different knowledge sources. Lastly, approach v3 combines the power of the two processing paradigms to create a hybrid.

1) v1: Bottom-Up Processing: The prior literature for both binary and multiclass classification has employed a basic approach to text classification problems by exploiting *local* content to extract n-gram features [8], [12]. Therefore, we use them for a baseline.

- (*T*) Textual Features:

We use the bag-of-tokens model, a well-known content exploitation approach in text mining. Each short-text document m_i is represented as,

$m_i = \{ (w_i, f(w_i)) \mid w_i \in W, f(w_i) \in [0,1]\}$, where w_i is an n-gram token, and $f(w_i)$ is a function for choice of the feature.

We create features using a dictionary W of n-gram tokens w_i that is acquired by tokenizing the documents of corpus A. We employ term frequency function as $f(w_i)$ for each n-gram token feature.

2) v2: Top-Down Processing: Approach v1 limits a learning algorithm in identifying and deriving relationships between features to the provided training data *locally*. Also, this process can be complex and time consuming, given that textual data can generate a large dictionary of n-gram features. Taking a top-down approach, rule-base and pattern-aided classification have been studied for text mining problems [10], [19], [15]. We use patterns (similar to rule antecedents) as binary features to classify intent. Therefore, we present three diverse ways to derive patterns *a priori* for informing the learning task of intent classification: *Declarative Knowledge*, *Social Behavior Knowledge*, and *Contrast Mining-guided Patterns*.

- (*DK*) Declarative Knowledge-guided Patterns:

Declarative knowledge includes facts, and in this context, knowledge about the expression of different intent classes. Using domain expertise, one can provide patterns for specific lexical rules to express intents. Another way is to rely on studies of human expression from linguistics and psychology, which can inform the design of domain independent rules.

Within the crisis domain, our prior study of binary intent classification [4] created request-offer behavior patterns from search logs of domain experts at American Red Cross. In another study [15], we used characteristics of communication from psycholinguistics to define a pattern set of intent expressions. We leverage these pattern sets from the prior studies in the multiclass classification context for creating binary features.

The pattern design relies on semantic-syntactic knowledge of intent expression. For example, a subject with the main verb "have" and any noun suggests an *offering* intent. However, the same text preceded by the auxiliary verb "do" and the pronoun "you" suggests a *seeking* intent. Similarly, word order such as verb-subject position also plays a crucial role in intent expression, and provides stark contrast to the unordered bag-of-tokens model for feature representation. Such patterns for expressing intent can help address the *ambiguity* challenge by endorsing the likelihood of an intent association for a short-text document.

The pattern design leverages a lexicon of verbs, given that verbs imply a plan for action. Using Schank's P-Trans primitive [20], which reflects the transfer of property, we acquire *seeking-offering* intent related verb classes. Specifically, our verb lexicon includes the Levin verb [21] categories of {give, future having, send, slide, carry, sending/carrying, put, removing, exerting force, change of possession, hold/keep, contact, combining/attaching, creation/transformation, perception, communication}. Our patterns also include classes of auxiliary verbs (e.g., 'be', 'do', 'have'), the modals (e.g., 'can', 'could', 'may', 'might', 'would'), question words ('wh'-words and 'how'), and the conditional ('if').

We extend the seed patterns by an exhaustive representation of synonymous verbs preserving the tense, using the WordNet knowledge base [22] (we will provide a list of 29 seed patterns with datasets upon request). We create a binary feature for each pattern using pattern matching on m_i.

Determiners (the)
Determiners (a, an)
Subject pronouns (she, he, we, they)
Mixed subjects/object pronouns but centered on individual (my, I, me)
Relative pronouns (that, this, these, those)
Possessive (mine, yours, his, hers, ours, theirs)
Relative pronouns (who, what, which, whom, whose)
Intensive/reflexive pronouns (myself, yourself, himself, herself, itself, ourselves, themselves, yourselves)
Dialogue management indicators (thanks, yes, ok, sorry, hi, hello, bye, anyway, how about, so, what do you mean, please, {could, would, should, can, will} followed by pronoun)
Hedge words (kinda, sorta)
Ambiguous pronoun (you)
Ambiguous pronoun (it)
Object pronouns (us, them, him, her)

TABLE II
CONVERSATION INDICATORS FOR SOCIAL BEHAVIOR KNOWLEDGE

An example of a seed pattern is: $\backslash b(I|we|they|he|she)\backslash b.*$ $\backslash b(like|want|likes|wants)\backslash b. \quad * \quad \backslash b(to)\backslash b. \quad *$ $\backslash b(\{$LEVIN-VERB-LEXICON-FOR-*give*-CATEGORY$\})\backslash b$

- (*SK*) Social Knowledge-guided Patterns:

Conversations are the foundation of social context. In online socio-technical systems, citizens generate intentional content in the expectation of a cooperative listening audience. This differs from user actions that may or may not have a motive for social interaction (e.g., search intent). Exploiting such a social aspect of conversational behavior as a knowledge source can improve the informational context of intent classification. For instance, examination of determiners follows [23], which asserted that "the" assumes a previously established topic. Often discarded as stopwords, we rather consider such conversational indicator categories to be social behavior patterns (see Table II), as studied for conversation classification in [24]. To factor in the degree of conversationality, instead of creating a binary feature for the pattern, we create term frequency based numerical features for each of the conversational indicator categories. Additionally, we also created a feature of word counts on m_i.

- (*CTK*, *CPK*) Contrast Mining-guided Patterns:

In declarative knowledge and social behavior knowledge, it is possible to miss predictor relationships due to the challenge of creating an exhaustive pattern set. Therefore, our goal is to incorporate the power of data mining to discover contrasting patterns for each of the intent classes as *a priori* knowledge in the learning process. Contrast patterns are those patterns that occur significantly more often in a class of interest than in other classes. Such patterns can boost data representation for learning predictor-class relationships [25]. The patterns should be sequential due to the importance of token (word) order in intent expressions. Prior literature has observed the importance of sequential pattern-aided text classification [19], [26].

We first mine sequential patterns within a labeled dataset of an intent class, followed by contrasting such class-wise pattern sets against each other to derive interesting, and novel emerging patterns [25], [27]. Incorporating this technique can help address the *sparsity* challenge of efficiently capturing context for imbalanced intent classes.

Formally, we define a *Sparse-Contrast-Strength(P,C_j)* measure for selecting useful contrasting patterns P (of given intent class C_j) as features; this will involve several definitions.

a. An item is a token; a token is a word or an n-gram (depending on experiment setting). A pattern P is a finite set of items. A document d matches a pattern P if every item of P occurs in d. The support of a pattern P in dataset A, denoted by $support(P, A)$, is

$$support(P, A) = |T_P|/|A|,$$
$$\text{where } T_P = \{d \in A \mid d \text{ matches } P\}.$$

Here $|A|$ denotes the cardinality of A.

b. Assume A and B are two datasets and P is a pattern. Then the support ratio of P from B to A, also called *growth rate*, is

$$gr(P, A, B) = support(P, A)/support(P, B).$$

c. Given a minimum support threshold $minSup$ and a minimum support ratio threshold $minGr$, a pattern P is a contrast pattern (a.k.a. emerging pattern) of C_i if $support(P, C_i) \geq minSup$ and $gr(P, C_i, \bar{C}_i) \geq minGr$. Here, $\bar{C}_i$ denotes the complement of C_i, i.e. $\bar{C}_i$ is the set of all documents of the application not in C_i. A Jumping Emerging Pattern (JEP) is a contrast pattern that has an infinite *gr*. That is, a JEP is present in C_i but not in its complement.

d. The contrast measure value of a pattern P for class C_j is

$$\begin{aligned} &\textit{Sparse-Contrast-Strength}(P, C_j)\\ &\quad = \textit{support}(P, C_j) * \textit{Contrast-Growth}(P, C_j)\end{aligned}$$

where,

$$\begin{aligned} &\textit{Contrast-Growth}(P, C_j)\\ &\quad = 1/(|C_j| - 1) \textstyle\sum_{C_k, k \neq j} gr(P, C_j, C_k)/\\ &\qquad\qquad\qquad\qquad (1 + gr(P, C_j, C_k))\end{aligned}$$

and *Contrast-Growth*$(P, C_j, C_k) = 1$ if $gr(P, C_j, C_k)$ is infinite (a case of jumping emerging pattern).

We will use a ranking method to select patterns – we select the top-$X\%$ patterns (X is a parameter) ranked by the *Sparse-Contrast-Strength*(P, C_j) measure. Regarding computation, instead of computing all frequent patterns which can be costly, we compute per class frequent patterns with support threshold ST_j for an intent class C_j. After computing frequent patterns, we prune for minimal patterns in each C_j, and then find the contrast measure of the remaining patterns. (A pattern P is a minimal contrast pattern of C_j if it does not contain other contrast patterns of C_j.)

We create a binary feature per selected pattern using pattern matching on m_i. We denote the feature set as *CTK* when items are the text tokens for the patterns, and as *CPK* for the case when items are part of speech (POS) tokens.

3) v3: Hybrid Processing: Our hybrid approach combines bottom-up processing (v1), and top-down processing (v2) representations to obtain features for improving the expressivity of informational context. The feature set contains both bag-of-tokens as well as set-of-patterns. Therefore, this approach exploits *a priori* knowledge as patterns external to the bag-of-tokens representation for the learning process. It allows efficient learning of expressive and diverse patterns.

C. Class Labeling Process

Three human judges annotated each tweet using six labels for ground truth: {*Request to get*, *Offer to give*, *Both request and offer*, *Report of past donations*, *None of the above*, *Cannot judge*}. We merged the labels to design the intent classes {*seeking*, *offering*, *none* (i.e. neither seeking nor offering)}. Hence, we excluded 'Both request and offer', and 'Cannot judge' labeled tweets,such that 'Request to get' represents the *seeking* intent class, 'Offer to give' represents the *offering* intent class, and 'Report of past donations' and 'None of the above' represents the *none* class.

D. Learning Method for Classification

We note two ways to address *ambiguity*, and *sparsity* in the learning process—first, by improving the *expressivity* for representing content (e.g., feature design), and second, via *algorithmic* choice for learning (e.g., boosting). We focus on improving the *expressivity* by a rich feature space obtained using our hybrid approach (v3) to reduce *ambiguity* in content interpretation. For the *algorithmic* choice, we use an ensemble learning approach for base learners of the multiclass binarization frameworks OVO and OVA. The ensemble approach helps in learning with imbalanced data to address the *sparsity* issue of the multiclass problem.

V. EXPERIMENTS

Data. We collected two diverse crisis event datasets:

1) Dataset-1: 4.9 million tweets for Hurricane Sandy in the US in 2012 (October 27 to November 7), and
2) Dataset-2: 1.9 million tweets for Typhoon Yolanda in the Philippines in 2013 (November 7 to 17).

For *Dataset-1*, we borrowed the raw labeled data with similar labeling design from our prior study focused on binary intent classification [4], which resulted in a total of 3,135 unique labeled tweets (Levenshtein similarity $<$ 0.75). It consisted of 52% *seeking*, 6% *offering*, and the remaining 42% *none* class labels. For *Dataset-2*, given *sparsity* in the data for intent classes observed in the Dataset-1, we created a biased sample for labeling to get more human judged labels. We selected 2,000 unique tweets with four diverse random samples of 500 tweets each: first from all the tweets in Dataset-2, and others from *donation* classified, *seeking* classified, and *offering* classified tweets of Dataset-2 using binary classifiers of the prior study [4]. Given the biased sampling, we used the strict criterion of 'all agree' on a label decision for the three human judges. Table III shows similar class label distribution (*seeking* more prevalent than *offering*) across the datasets.

CLASS	Dataset-1	Dataset-2
seeking	1,626 (52%)	197 (26%)
offering	183 (6%)	91 (12%)
none	1,326 (42%)	475 (62%)

TABLE III
LABELS FOR TWO REAL-WORLD CRISIS DATASETS

Features. For feature set *T*, we used bi- and tri-grams for a bag-of-tokens model after text preprocessing on the corpus A as performed in related binary classification work [12], [4]. Text preprocessing steps include platform characteristics generalization (explained next), followed by removal of non-ASCII characters, stopword removal, stemming and string to word vectorization with default parameters using the WEKA API [28]. We also generalized platform specific characteristics by replacing retweets ("RT @user_name" by _RT_), user mentions ("@user_name" by _MENTION_) as well as numbers (by _NUM_) and hyperlinks (by _URL_). We employed a normalized term frequency function for bag-of-tokens features of *T*. For feature set *CTK*, we used the preprocessed text corpus A and employed the SPADE [29] algorithm for mining sequential frequent patterns per class. Minimum support (ST_j) per class was chosen equally to be 10% after examining various settings. The extreme setting of 2% leads to spurious patterns and the opposite extreme of 50% (a conventional choice) yields very few patterns due to noise. Similarly, the minimum *gr* was chosen as 1.5. In case of the feature set *CPK*, we first extracted POS tags of the corpus A using the ARK-NLP tool [30]. Minimum support 50% works for POS tags given that they represent abstract syntactic classes (e.g., Adjective) that are frequent across multiple itemsets. Similarly, *gr* was chosen as 2.0. We computed contrast strength for the patterns per class and observed a majority of emerging patterns, and therefore, chose X=100 for the ranked patterns. We created binary features for the extracted patterns of *CTK* and *CPK* by employing pattern matching (1 for the match, 0 otherwise). For the feature set *DK* also, we employed pattern matching for the exhaustive pattern set to create binary features. For the feature set *SK*, we used a term frequency function for the linguistic indicator categories listed in Table II (e.g., dialogue management) for each $m_i \epsilon A$.

The resultant hybrid feature space represented a rich representation of the relevant informational context. For example, in a message 'My home state was hit hard by hurrican Sandy. To help with the hurricane victims in Brooklyn, Staten Island, NY.. http://t.co/VbnfoMbl', a diverse contrast pattern for *seeking* class was found, 'help .* victim .* _URL_ .*'. The pattern captured a long distant relationship between items, which could take complex processing time and effort to get identified with the n-gram tokens representation alone. Similarly, for the second example in Table I, despite having a partial question/seeking form, 'I wanna give' pattern associated with the *offering* help intent class from the declarative pattern set aids in resolving *ambiguity*, by strongly endorsing the likelihood of *offering* intent.

Learning. We experimented with the combination of all the feature sets (*T*, *DK*, *SK*, *CTK*, *CPK*) to design OVO and OVA based multiclass classifiers. We used the ensemble learning algorithm of Random Forest [28] (with 10 trees, 100 features and depth level 5 nodes) for base learners. Using the Scikit-Learn python library[2] for experiments, we performed 10-fold cross-validation (CV) to evaluate performance measures of accuracy and F1 score, following the evaluation in prior works on multiclass classification. Accuracy and F1 score reflect performance improvement across the classes including minority classes. To investigate our hypothesis of exploring the influence of knowledge-guided pattern features, we created bottom-up processing (approach v1) based feature set *T* as the baseline.

VI. Results and Discussion

10-fold cross-validation results in Figure 1 show a performance gain in both accuracy and F1 score attributable to addition of top-down processing feature set (*DK, SK, CTK, CPK*) to the baseline feature set (*T*). Therefore, we performed statistical significance test using two-tailed *t*-test for comparing results of the bottom-up processing feature representation (*T*), and the proposed approach of hybrid representation (*T*, *DK*, *SK*, *CTK*, *CPK*). We noted statistically significant ($p <$ 0.05) performance differences in F1 score and accuracy for both datasets and learning frameworks. The absolute gain in F1 score and accuracy for the hybrid approach are up to 7% and 6% respectively.

Significance of Prior Knowledge. We ranked features from the hybrid processing in both the datasets using a Chi-squared test with default parameters in the WEKA tool. More than half of the most discriminating, best ranked 1% features belong to the top-down processing representation. It informs about the value of knowledge guidance in learning the feature space for identifying newer and better statistical predictor-class relationships. Accuracy and F1 score improvements support our hypothesis H1. Dialog management and subject pronoun category features present social behavior knowledge that exist among the top features. They support our argument that knowledge of social behavior in conversations can improve context for identifying intent in the social setting.

Performance in the Multiclass Classification Frameworks. We observed significant improvement in F1 and accuracy scores in both frameworks, OVA and OVO. Interestingly the gain observed in both the cases is significant. OVA suffers from imbalance created by the framework design itself, and OVO suffers from the class dependence issue owing to the design of pairwise classification. Despite these challenges, classification performance improves in both frameworks using the hybrid approach, supporting our hypothesis H2.

Limitations. We have studied two diverse events related to the crisis domain. However, other domains can be explored using our multiclass framework in future that may be differently influenced by knowledge-guidance in intent classification.

[2]Open source Machine Learning library: http://scikit-learn.org/stable/

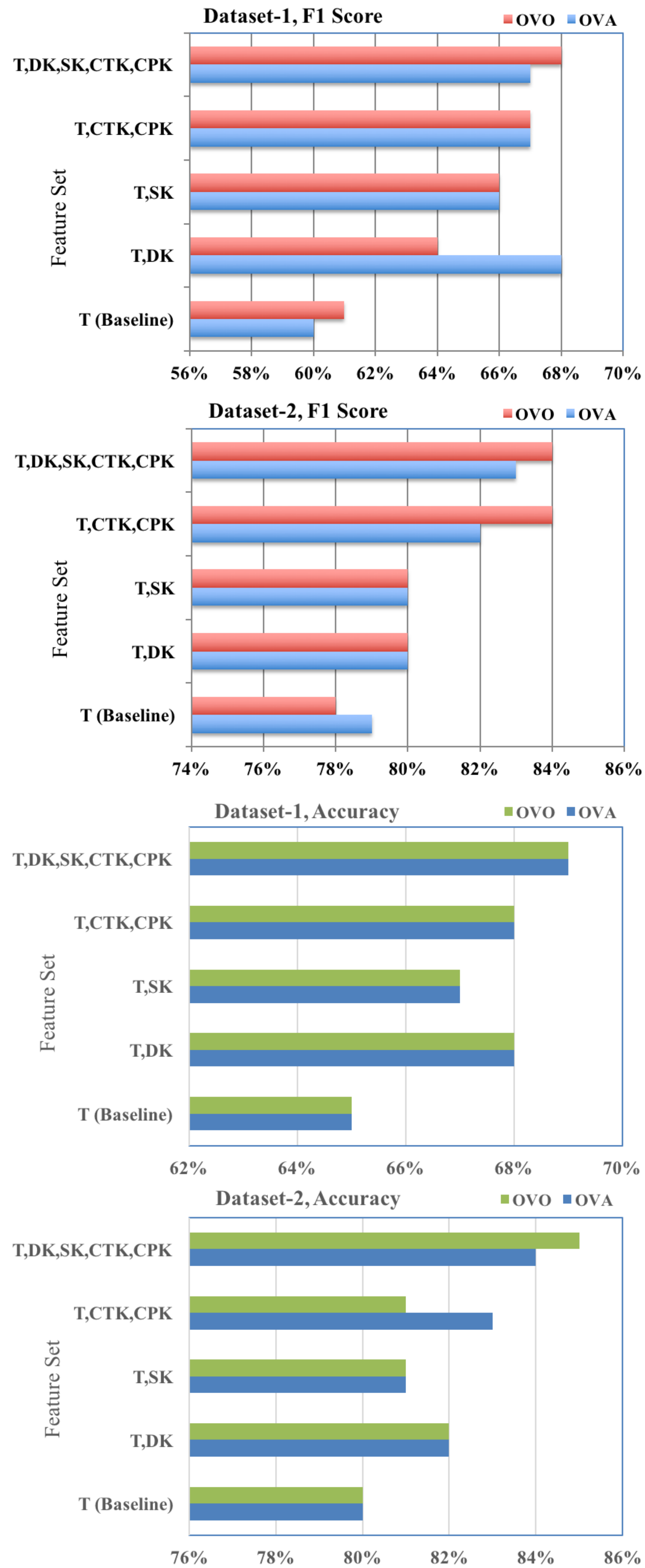

Fig. 1. 10-fold cross validation results show statistically significant gain for the proposed hybrid approach v3 (feature set: $\{$*T,DK,SK,CTK,CPK*$\}$) compared to the baseline approach v1 for both datasets within OVA and OVO learning frameworks.

We also note the varying performance of knowledge-guided feature sets between the Dataset-1 (Hurricane Sandy, US based) and Dataset-2 (Typhoon Yolanda, Philippines based), which may be due to socio-cultural differences of expressing intent in the diverse events. It requires further investigation on assessing the need for building socio-culture-region specific models of intent mining. We have shown the importance of contrast patterns using equal support thresholds for all classes, however, we shall explore the effects of varying class-wise thresholds. We did not show results for feature selection-based and modified data sampling-based (e.g., under/oversampling) learning schemes, given that it can be subjective. We prefer to first answer questions of any improvements of feature representation in a general setting for a learning task. We shall explore various algorithmic choices, including cost-sensitive learning combined with the ensemble framework in the future. Also, there can be multiple valid intent expressions within a message—an instance of multi(-intent) label classification problem, e.g., *acknowledging* appearing with *seeking* intent. We shall extend the multiclass framework for the multilabel setting. We also limited intent classes crucial to the crisis response coordination but future studies could incorporate more classes without scalability concerns, given the use of binarization frameworks.

VII. Conclusion

We presented a novel approach to classify intent of short-text on social media using a hybrid approach that combines knowledge-guided patterns with syntactic features based on bag of n-gram tokens. Specifically, we explored a variety of knowledge sources (declarative, social behavior about conversations and contrast patterns) to create pattern sets for examining improvement in the multiclass intent classification. Our experiments on the two crisis event datasets demonstrated a statistically significant gain in the F1 score and accuracy in both the popular multiclass frameworks of One-vs-One and One-vs-All. Application of the presented multiclass framework for intent classification on social media can help organizations efficiently filter content to meet their information needs as well as engage targeted citizens. For example, resource scarcity (*seeking* intent) and availability (*offering* intent) data collection can help decision making for resources during a crisis response.

Acknowledgment

We thank US National Science Foundation for the SoCS grant IIS-1111182 'Social Media Enhanced Organizational Sensemaking in Emergency Response' to support this research. We also acknowledge our colleagues at Kno.e.sis Center, and collaborators at OSU and QCRI for the invaluable discussion and feedback to improve our results.

References

[1] A. Broder, "A taxonomy of web search," in *ACM Sigir forum*, vol. 36. ACM, 2002, pp. 3–10.

[2] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo, "Classifying and characterizing query intent," in *Advances in Information Retrieval*. Springer, 2009, pp. 578–586.

[3] I. Varga, M. Sano, K. Torisawa, C. Hashimoto, K. Ohtake, T. Kawai, J.-H. Oh, and S. De Saeger, "Aid is out there: Looking for help from tweets during a large scale disaster." in *ACL (1)*, 2013, pp. 1619–1629.

[4] H. Purohit, C. Castillo, F. Diaz, A. Sheth, and P. Meier, "Emergency-relief coordination on social media: Automatically matching resource requests and offers," *First Monday*, vol. 19, no. 1, 2013.

[5] I. Ajzen, "The theory of planned behavior," *Organizational behavior and human decision processes*, vol. 50, no. 2, pp. 179–211, 1991.

[6] B. F. Malle and J. Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology*, vol. 33, no. 2, pp. 101–121, 1997.

[7] S. A. Sloman, P. M. Fernbach, and S. Ewing, "A causal model of intentionality judgment," *Mind & Language*, vol. 27, no. 2, pp. 154–180, 2012.

[8] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *ACL*, 2012, pp. 90–94.

[9] M. Kröll and M. Strohmaier, "Analyzing human intentions in natural language text," in *Proceedings of the fifth international conference on Knowledge capture*. ACM, 2009, pp. 197–198.

[10] J. Ramanand, K. Bhavsar, and N. Pedanekar, "Wishful thinking: finding suggestions and 'buy' wishes from product reviews," in *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. ACL, 2010, pp. 54–61.

[11] Z. Chen, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Identifying intention posts in discussion forums," in *NAACL*, 2013, pp. 1041–1050.

[12] B. Hollerit, M. Kröll, and M. Strohmaier, "Towards linking buyers and sellers: detecting commercial intent on twitter," in *WWW companion*. ACM, 2013, pp. 629–632.

[13] C. S. Carlos and M. Yalamanchi, "Intention analysis for sales, marketing and customer service." in *COLING (Demos)*, 2012, pp. 33–40.

[14] J. Wang, G. Cong, X. W. Zhao, and X. Li, "Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets," in *AAAI*, 2015.

[15] H. Purohit, A. Hampton, S. Bhatt, V. L. Shalin, A. P. Sheth, and J. M. Flach, "Identifying seekers and suppliers in social media communities to support crisis coordination," *JCSCW*, vol. 23, pp. 513–545, 2014.

[16] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognition*, vol. 44, no. 8, pp. 1761–1776, 2011.

[17] A. Sen, M. Islam, K. Murase, and X. Yao, "Binarization with boosting and oversampling for multiclass classification," *IEEE Transactions on Cybernetics*, 2015.

[18] N. García-Pedrajas and D. Ortiz-Boyer, "An empirical study of binary classifier fusion methods for multiclass classification," *Information Fusion*, vol. 12, no. 2, pp. 111–130, 2011.

[19] S. Jaillet, A. Laurent, and M. Teisseire, "Sequential patterns for text categorization," *Intelligent Data Analysis*, vol. 10, no. 3, p. 16, 2006.

[20] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive psychology*, vol. 3, no. 4, pp. 552–631, 1972.

[21] B. Levin, *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[22] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[23] W. Chafe, "Cognitive constraints on information flow," *Coherence and grounding in discourse*, vol. 11, pp. 21–51, 1987.

[24] H. Purohit, A. Hampton, V. L. Shalin, A. P. Sheth, J. Flach, and S. Bhatt, "What kind of #conversation is twitter? mining #psycholinguistic cues for emergency coordination," *Computers in Human Behavior*, vol. 29, no. 6, pp. 2438–2447, 2013.

[25] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in *SIGKDD*. ACM, 1999, pp. 43–52.

[26] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *SIGIR*. ACM, 2006, pp. 244–251.

[27] E. Loekito and J. Bailey, "Using highly expressive contrast patterns for classification-is it worthwhile?" in *Advances in Knowledge Discovery and Data Mining*. Springer, 2009, pp. 483–490.

[28] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[29] M. J. Zaki, "Spade: An efficient algorithm for mining frequent sequences," *Machine learning*, vol. 42, no. 1-2, pp. 31–60, 2001.

[30] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *ACL*, 2011, pp. 42–47.