REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
06-09-2017	Final Report	1-Oct-2013 - 31-May-2017

4. TITLE AND SUBTITLE
Final Report: Content-based Covert Group Detection in Social Networks

5a. CONTRACT NUMBER
W911NF-13-1-0416

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Carnegie Mellon University
5000 Forbes Avenue

Pittsburgh, PA 15213 -3589

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
64114-NS.7

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU		Katia Sycara
					19b. TELEPHONE NUMBER
					412-266-8882

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

RPPR Final Report
as of 03-Oct-2017

Agency Code:

Proposal Number: 64114NS

Agreement Number: W911NF-13-1-0416

INVESTIGATOR(S):

Name: Katia Sycara
Email: katia@cs.cmu.edu
Phone Number: 41226688825
Principal: Y

Organization: **Carnegie Mellon University**
Address: 5000 Forbes Avenue, Pittsburgh, PA 152133589
Country: USA
DUNS Number: 052184116
EIN: 250969449

Report Date: 31-Aug-2017
Date Received: 06-Sep-2017
Final Report for Period Beginning 01-Oct-2013 and Ending 31-May-2017

Title: Content-based Covert Group Detection in Social Networks

Begin Performance Period: 01-Oct-2013
End Performance Period: 31-May-2017
Report Term: 0-Other
Submitted By: Katia Sycara
Email: katia@cs.cmu.edu
Phone: (412) 266-88825
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 4 **STEM Participants:** 4

Major Goals: Along with the benefits of providing ways for people to connect and socialize on-line, social media provides opportunities for various groups to spread messages to try to spread influence and shape opinions according to interests of the group. Community detection in social media is a challenging problem. A community can be defined as a group of users that (1) interact with each other more frequently than with those outside the group and (2) are more similar to each other that to those outside the group. Research on detection of groups that disseminate content to push their own agenda faces many challenges, due among many others, to the difficulties of modeling and detecting cues for strategic messages of groups. To help close this research gap, the overall goal of this project is to develop techniques and algorithms for detection of goal-driven covert groups that spread strategic information.

Accomplishments: General conclusions from the results: (a) Multi-label classification method RAkEL, no matter with or without postprocessing, can perform better than single-label classifiers, i.e., KNN and SVM in this study on both data sets. However, due to the different characteristics of different data sets, the scales for the Hamming loss are different. Among all the methods, RAkEL with weighted summation strategy performs best. For example, RAkEL+wsum is improved 29.65% compared with KNN on Obama care data set. (b) RAkEL with postprocessing using either summation strategy or weighted summation strategy can generate better results than the original RAkEL method which validates the effectiveness of our proposed postprocessing strategies. And weighted summation strategy is more effective on both data sets. For example, RAkEL+wsum and RAkEL+sum can get 7.6% and 5.21% improvement on Obama care data set, respectively © In all the results, we can observe that using the combination of (1+2)-gram, POS and STAT features performs best. From 1-gram features to (1+2)-gram, POS and STAT features, the more features are introduced, the better performance can be achieved.

Training Opportunities: Four graduate students were supported by the project. These were MS students in the Language Technologies department of Carnegie Mellon University. The students took courses in natural language processing, data mining in various multi-media data sets, text retrieval, text summarization and other relevant courses. They also extracted large number of data from social media such as Twitter, Fac3ebook blogs etc. they have developed and implemented novel algorithms.

Results Dissemination: The results were disseminated via presentations at conferences, and scientific journals.

Honors and Awards: Nothing to Report

RPPR Final Report
as of 03-Oct-2017

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Anika Gupta
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yulong Pei
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Rahul Iyer
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yuezhang Li
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

ARTICLES:

Publication Type: Journal Article Peer Reviewed: N **Publication Status:** 1-Published
Journal: PLoS ONE
Publication Identifier Type: DOI Publication Identifier: 10.1371/journal.pone.0102199
Volume: 9 Issue: 7 First Page #: 1
Date Submitted: 8/30/17 12:00AM Date Published:
Publication Location:
Article Title: Polarity Related Influence Maximization in Signed Social Networks
Authors: Dong Li, Zhi-Ming Xu, Nilanjan Chakraborty, Anika Gupta, Katia Sycara, Sheng Li
Keywords: information diffusion, signed networks, influence maximization
Abstract: Influence maximization in social networks has been widely studied motivated by applications like spread of ideas or innovations in a network and viral marketing of products. Current studies focus almost exclusively on unsigned social networks containing only positive relationships (e.g. friend or trust) between users. Influence maximization in signed social networks containing both positive relationships and negative relationships (e.g. foe or distrust) between users is still a challenging problem that has not been studied. Thus, in this paper, we propose the polarity-related influence maximization (PRIM) problem which aims to find the seed node set with maximum positive influence or maximum negative influence in signed social networks. To address the PRIM problem, we first extend the standard Independent Cascade (IC) model to the signed social networks and propose a Polarity-related Independent Cascade (named IC-P) diffusion model. We prove that the influence function of the PRIM prob
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ACM Conference on Information and Knowledge Management
Date Received: 30-Aug-2017 Conference Date: 27-Oct-2013 Date Published:
Conference Location: San Francisco, CA
Paper Title: Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction
Authors: Dong Li, Zhiming Xu, Yishu Luo, Sheng Li, Anika Gupta, Katia Sycara, Shengmei Luo, Lei Hu, Hong Ch
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 2-Awaiting Publicat
Conference Name: ACM Conference
Date Received: 30-Aug-2017 Conference Date: 23-Jul-2017 Date Published: 26-Jul-2017
Conference Location: Washington DC
Paper Title: Detecting type of Persuasion : Is there structure in persuasion?
Authors: Rahul R Iyer, Katia Sycara, Yuezhang Li
Acknowledged Federal Support: **Y**

Final Report (2017)

Award Number: W911NF-13-1-0416
Award Title: Content-based Covert Group Detection in Social Networks
Period of Performance: August 1 2016-July 31 2017
Performing Organization: Carnegie Mellon University
PI: Katia P. Sycara

Background and Research Goals

Along with the benefits of providing ways for people to connect and socialize on-line, social media provides opportunities for various groups to spread messages to try to shape opinions according to interests of the group. Community detection in social media is a challenging problem. A community can be defined as a group of users that (1) interact with each other more frequently than with those outside the group and (2) are more similar to each other that to those outside the group. Research on detection of covert groups that disseminate content to push their own agenda faces many challenges, due among many others, to the difficulties of modeling and detecting cues for strategic messages of groups that may not be connected. To help close this research gap, the **overall goal** of this project is to (a) develop techniques and algorithms for detection of goal-driven covert groups that spread strategic information and may not be explicitly connected.

Research Accomplishments

During the research program we performed work on various topics related to data mining in social media including: we performed work, on (a) diffusion in social networks, (b) influence maximization in signed social networks, (c) community detection in social networks, (d) detecting and classifying persuasive arguments in social media, (e) hierarchical category embedding, and (f) simultaneous identification of tweet purpose and position.

A. Diffusion in Social Networks

Modelling the temporal process of diffusion in social networks is a critical research task that has not been adequately addressed in prior work. To model and predict the temporal dynamics of the diffusion process, we developed we a novel information diffusion model (GT model), which considers the users in network as intelligent agents. Each agent jointly considers all its interacting neighbors and calculates the payoffs for his different choices to make strategic decision. We introduced the time factor into the user payoff, enabling the GT model to not only predict the behavior of a user/agent but also to predict when the agent will perform the behavior. Both the global influence and social influence are explored in the time-dependent payoff calculation, where a new social influence representation method is designed to fully capture the temporal dynamic properties of social influence between users. Experimental results on Sina Weibo (36 million microblogs published by 251,639 users) and Flickr (43 million favorite markings over 11 million photos published by 2.5 million users) validate the effectiveness of our methods and show that our model outperforms other state of the art algorithms in precision and recall.

B. Influence Maximization in Signed Social Networks

Online social networks such as Twitter, Facebook and Google+ have developed rapidly in recent years. They support social interaction and information diffusion among users all over the world. These online sites present great opportunities for large-scale viral marketing. Viral marketing is a cost effective marketing strategy that promotes products by giving free or discounted items to a selected group with high influence, in the hope that through the word-of-mouth effects, a large number of users will adopt the product. Motivated by viral marketing, influence maximization has emerged as a fundamental problem concerning the diffusion of products, opinions, and innovations through social networks. Diffusion models are used to explain and simulate the spread of information in social networks. The influence maximization problem is to determine a small subset of seed nodes in a network graph that maximizes the influence, under a given diffusion model. To date, however, all works consider influence maximization in unsigned social networks which only have positive relationships between users (e.g. friend or trust). However, the polarity of relationships in social networks is not always positive. There are also signed social networks containing both positive and negative relationships (e.g., foe or distrust) simultaneously. Influence maximization in signed social networks is a key problem that has not been studied and it is the focus of this paper. We developed a model of influence diffusion in signed networks and prove that the influence function is monotonic and submodular. Therefore a greedy algorithm can be used and be shown to achieve an approximation ratio of 1-1/e. Experimental results on two signed social network datasets, Epinions (131,828 users and 841,372 relationships) and Slashdot (77,350 users and 516,575 relationships) show the effectiveness and superior performance of our approach.

C. Community Detection in Social Networks

Many existing works on community detection focus only on social relations or content. However, neither social relations nor content alone can indicate the community membership accurately. On one hand, in the real-world social media such as Twitter, compared with the large amount of users, the social relations for each user are extremely sparse and two users may belong to the same community even if there are no relations between them. On the other hand, the content on social media is diverse and noisy which will influence the content analysis and may lead to failure in detecting communities. Therefore, combining the relations and content may be a better strategy for community detection. In summary, there are three challenges in community detection task: There are three general challenges in community detection: (1) Combination Strategy. It is insufficient to determine the community membership using only social relations or only content. Thus, the model should combine social relations and content in detecting communities. (2) Model Flexibility. This challenge requires that the model can capture different social information without changing the model form so the model can be generalized to modeling different social networks. (3) User Similarity. Community detection essentially is a user clustering problem in which the user similarity plays an important role. Thus, the model should consider the user similarity explicitly.

Additionally, the user similarity calculation should use both user relations and content. In this paper, we organize users and messages in a userword-message tripartite graph. We cluster the users into different communities using not only the user-word-message relations and the user pairwise relation but also user similarity, message similarity and user interaction. In order to cluster the users, we employ a constrained nonnegative matrix tri-factorization (NMTF) framework to cluster users and messages simultaneously by combining the relations and content, and propose three types of graph regularization to model user similarity, message similarity and user interaction explicitly.

This proposed method can deal with the three challenges introduced above. First, we utilize two NMTF components to model the use-rword relation and the message-word relation respectively and one NMTF component to model the user pairwise relation. The NMTF method performs well in co-clustering tasks with multiple relations so the combination of these NMTF components can fuse social relations and content. Second, by integrating graph regularization, the NMTF framework is flexible in incorporating rich social information such as retweet and citation in social networks. In particular, we introduce three types of graph regularization based on user similarity, message similarity and interaction respectively. Other social information can also be integrated using similar graph regularization without changing the form of this framework. Third, we model the user similarity and message similarity explicitly in the graph regularization. To exploit the similarities, we construct a two-layer graph based on user pairwise relations, message pairwise relations and user-message relations, and then propose a random walk method on this graph to calculate the user similarity and message similarity. This method employs both user relations and content to calculate the user similarity and message similarity in the networks. Furthermore, in order to validate the effectiveness of the proposed method, experiments were conducted on three real-world data sets. The data sets were the following. The first data set was Politics-UK: This data set consists of 419 Members of Parliament from the United Kingdom who belong to five different political groups. There were 72,693 messages. The second data set is Politics-IE: This data set has 348 Irish politicians and political organizations. These are assigned to seven disjoint groups according to their affiliation. This data set contained 49,546 messages. The third data set is DBLP: This data set contains 6604 authors 8293 papers from 16 top conferences which cover 4 research fields including Machine Learning, Information Retrieval, Data Mining and Database. There are 8,293 messages. Our method outperformed state of the art methods based on content only or combination of content and relation.

D. Detecting and Classifying Persuasive Arguments

The amount of information shared on online social media has been growing at unprecedented rates during recent years. The uncontrolled nature of social media makes them vulnerable to exploitation for spreading spam, rumors, slander, and other types of misinformation. An increasing number of people rely, at least in part, on information shared on social media (SM) to form opinions and make choices on issues related to lifestyle, politics, health, and product purchases, and this reliance provides motivation for a variety of parties to promote information. There has been some work on detecting such promotional campaigns on social media such as Twitter. We want to perform a similar task by detecting persuasion in the texts. In order to successfully promote a campaign, the information being promoted, pertaining to products, political views etc., has to be persuasive (in order to persuade the readers) and also the promoter should have a wide influence. The work we performed was focused on to persuasion detection in text. Once a good model is built for this, we can incorporate the model into a social network graph and integrate techniques for detecting influential members of the social network.

The goal here is to detect if a given piece of text is persuasive. If it is, then we can look into the type of persuasion strategy being used, such as threat/promise, outcome, reciprocity etc. Here, in this work, we look at 11 different persuasion strategies. These will be listed later. From the social network graph (graph of users), one can determine if the message being spread by a user is intended to be spam and if he is successful in his/her promotion, by looking at how persuasive the message is, and also by looking at how influential the user is (this can be determined by that user's pagerank score from the graph).

Blog Authorship Corpora: This is a subset of the blog authorship corpus. Each directory corresponds to a blog. Each blog has sub-directories corresponding to a day. Inside the day sub-directory, there may be multiple posts; posts are identified with an underscore followed by a number. Out of around 25048 posts, only around 457 were annotated with persuasive acts. Each blog post has been broken down into different "text tiles", which are a few sentences long, and each of these tiles are annotated with a persuasive tactic (if present). A total of 11 persuasive tactics, such as social pressure, reasoning, empathy, favors/debts etc. were used to annotate categories.
We tested various approaches for classification of persuasive arguments. We used both supervised and unsupervised methods. In the supervised cases, we have split the dataset into 100 posts for training and the rest for testing. The methods used and the results are described below.

1. Simple Supervised: This is the simplest approach taken. In this case, the 100 training posts were taken and simple textual features were extracted: unigrams, bigrams, without punctuation. An SVM was then run on these features to learn a model that could be applied to the holdout set. This model was then used to test on the remaining posts. Accuracy: 51.4 %. This accuracy very good considering this is not a binary but rather an 11-wise classification.

2. Supervised Document Vectors: In this case, first the different posts (all posts were considered), separated into different categories based on the type of persuasion (11 categories in total). Then, the Doc2Vec model (Distributed Representations of Sentences and Documents, Quoc Le and Tomas Mikolov, 2013) was applied to convert all the posts to vectors and the prototype vector for each category was chosen as the mean of all the vectors in that category. Now, there are 11 vectors which represent the different categories. To classify future blog posts, one would compute the vector of that post and then compute the similarity (cosine) to the prototype vectors. The category which has the highest similarity is the one that is chosen. Accuracy: 56.2 %. This accuracy is indeed higher than a simple SVM on simple textual features. We tested this model because of how successful the vector embedding models are.

3. Document Vector: This is similar to the above method, except that all the posts need not be annotated. So, this is kind of like an unsupervised approach. The Doc2Vec for all the posts is computed. Then, a random vector from each category is chosen as the representative of that category. Since, we just need to label 11 posts in this case, one can consider this to be almost unsupervised as these 11 posts can be considered as some sort of a seed. Similar to the above method, to classify a new post, one would compute its vector's similarity with the vectors for each of the categories. The category with the highest similarity is chosen. Accuracy: 53.7 % . This compares favorably with the previous method, where all the vectors were used. This is probably because each persuasion tactic has a unique sentence structure and the mean of all vectors of a category (since the embeddings take into account the context) is close to each vector of that category.

4. Domain-Independent: In this case, we attempted a domain-independent classification. The idea here is that persuasive documents may have a certain characteristic sentence structure and we want to exploit that. This can also be considered almost unsupervised. We pick 11 posts in total, one for each category. We perform phrase-structure parsing to obtain the parse-tree for each of these posts. We then convert these parse trees into a string, and remove the leaf nodes (the terminal symbols: words, so that we get just the structure of the sentence). So now, we obtain the representative strings for each of the category. To classify a new post, we first compute its parse-tree-string. Then, we compute the normalized edit-distance for this string to the strings of each of the category. The category with the least edit-distance is chosen. Accuracy: 54.3 %. This approach is comparable to the vector-embedding approaches indicating that there is indeed a structure to the tactics which can be exploited.

E. Hierarchical Category Embedding

Hierarchies provide a natural way to categorize and locate knowledge in large-scale knowledge bases (KBs). For example, WordNet, Freebase and Wikipedia use hierarchical taxonomy to organize entities into category hierarchies. These hierarchical categories could benefit applications including concept categorization, object categorization, document classification, and link prediction in knowledge graphs. In all of these applications, it is essential to have a good representation of categories and entities as well as a good semantic relatedness measure.

We developed two models to jointly learn entity and category representations from large scale knowledge bases (KBs). They are Category Embedding model and Hierarchical Category Embedding model. The Category Embedding model (CE model) extends the entity embedding method of (Hu et al., 2015) to jointly learn entity and category embeddings and Hierarchical Category Embedding model (HCE model) extends CE model with category hierarchies. The final learned entity and category vectors can capture meaningful semantic relatedness between entities and categories. We evaluate our models from two applications: concept1 categorization (Baroni and Lenci, 2010) and dataless hierarchical classification (Song and Roth, 2014).

The main contributions of our work are summarized as follows. First, we achieved jointly learning entity and category embeddings with CE model. Second, we extended CE model to HCE model and achieves an improvement in embedding quality. Third, we developed a concept categorization method based on nearest neighbor classification, avoiding issues arising from disparity in the granularity of categories that plague traditional clustering methods. Fourth, we constructed a new concept categorization dataset from Wikipedia. Fifth, we showed the potential of utilizing entity embeddings on dataless classification. Overall, our model produced state of the art performance on both concept categorization and dataless hierarchical classification.

Experiments

In the experiments, we use the dataset collected from Wikipedia on Dec. 1, 20153 as the training data. We preprocess the category hierarchy by pruning administrative categories and deleting bottom-up edges to construct a DAG. The final version of data contains 5,373,165 entities and 793,856 categories organized as a DAG with a maximum depth of 18. The root category is “main topic classifications”. We train category and entity vectors in dimensions of 50, 100, 200, 250, 300, 400, 500, with batch size B = 500 and negative sample size k = 10. With the training dataset defined above, we conduct experiments on two applications: concept categorization and dataless hierarchical classification.

1. Concept Categorization

1.1 Datasets

There are two datasets used in this experiment. The first one is the Battig test set introduced by (Baroni and Lenci, 2010), which includes 83 concepts from 10 categories. The Battig test set only contains single-word concepts without any multiple-word concepts (e.g., “table tennis”). Hence, using this dataset restricts the power of concept categorization to single-word level. We use this dataset because it has been used as a benchmark for most previous approaches for concept categorization.
3https://dumps.wikimedia.org/wikidatawiki/20151201/

Due to the limitations of the Battig test set, we construct a new entity categorization dataset DOTA (Dataset Of enTity cAtegorization) with 450 entities categorized into 15 categories. All the categories and entities are extracted from Wikipedia, so the resulting dataset does not necessarily contains only single-word entities. Thus, the dataset can be split into two parts, DOTA-single that contains 300 single word entities categorized into 15 categories and DOTA-mult that contains 150 multiple-word entities categorized into the same 15 categories.

Results

In the experiments, we used scikit-learn (Pedregosa et al., 2011) to perform clustering. We tested k-means and hierarchical clustering with different distance metrics (Euclidean, cosine) and linkage criterion (ward, complete, average). All these choices along with the vector dimensionality are treated as our models' hyper-parameters. For selecting hyper-parameters, we randomly split the Battig and Dota datasets to 50% of validation data and 50% of test data, evenly across all categories. We trained all the embeddings on the same Wikipedia dump and tuned hyper-parameters on the validation set. For experiments on Dota dataset, since the ground truth is contained in our Wikipedia corpus, we deleted all category-entity links contained in Dota dataset from our category hierarchy to train HEE, TransE and HCE embeddings to make a fair comparison.

Our hierarchical category embedding (HCE) model outperforms other methods in all datasets. Our model achieves a purity of 89% on Battig and 90% on DOTA-all.

2 Dataless Hierarchical Classification

20Newsgroups Data(20NG):The 20 newsgroups dataset (Lang, 1995) contains about 20,000 newsgroups documents evenly categorized to 20 newsgroups, and further categorized to six super-classes. We use the same label description provided by (Song and Roth, 2014).

RCV1 Dataset: The RCV1 dataset (Lewis et al., 2004) contains 804,414 manually labeled newswire documents, and categorized with respect to three controlled vocabularies: industries, topics and regions. We use topics as our hierarchical classification problem. There are 103 categories including all nodes except for the root in the hierarchy, and the maximum depth is 4. To ease the computational cost of comparison, we follow the chronological split proposed in (Lewis et al., 2004) to use the first 23,149 documents marked as training samples in the dataset. The dataset also provides the name and the description of each category label.

Results

We perform dataless hierarchical classification using the same ESA settings, combined with different embedding methods for similarity calculation. Experiments were performed with different embedding methods in dimensionality of {50,100,200,250,300,400,500}. The performance increase from our method, compared to other state of the art methods ranges from 2% improvement to 10% improvement in prediction accuracy.

F. Simultaneous Identification of Tweet Purpose and Position

Tweet classification has attracted considerable attention recently. Most of the existing work on tweet classification focuses on topic classification, which classifies tweets into several predefined categories, and sentiment classification, which classifies tweets into positive, negative and neutral. Since tweets are different from conventional text in that they generally are of limited length and contain informal, irregular or new words, so it is difficult to determine user intention to publish a tweet and user attitude towards certain topic. In this paper, we aim to simultaneously classify tweet purpose, i.e., the intention for user to publish a tweet, and position, i.e., supporting, opposing or being neutral to a given topic. Tweets have been classified into different classes of purpose, e.g., social interaction with people, promotion or marketing, information sharing, etc. For position classification tweets are classified into positive, negative and neutral. We transform this problem to identify tweet purpose and position simultaneously into a multi-label classification problem. Our method is advantageous in two aspects: (1) It is more efficient to use multi-label classification methods to simultaneously identify tweet purpose and position since only one unified classifier needs to be trained. (2) The correlation between tweet purpose and position can be captured by multi-label classification methods to improve the accuracy for classification. Besides, aiming to tackle the issue that some tweets in the data are predicted to contain no labels or multiple labels using multi-label classification method, two different postprocessing strategies have been proposed. In order to validate the effectiveness of this problem transformation and postprocessing strategies, we build two data sets collected from Twitter and experiments are conducted on the data sets.

The work makes the following contributions: (a) We define the task to identify tweet purpose and position simultaneously and transform this problem to a multi-label classification problem. (b) We propose two postprocessing strategies i.e., summation and weighted summation, for the classification task and by incorporating the strategies into the multilabel classification method, the classification performance can be improved. (c) We test our approach on two real-world data sets to validate the classification method with postprocessing and the results demonstrate the effectiveness of the problem transformation and postprocessing strategies.

Data Sets: In order to build the Twitter data set, we collected the tweets in two topics, i.e., Obama care and death penalty. We used Twitter Search API1 with the queries *Obama care* and *death penalty*. Then we preprocessed these tweets by removing (1) non-English tweets, (2) tweets less than 5 words, and (3) duplicated tweets. After removing the irrelevant tweets to these two topics, we labeled 1000 tweets for each of the two topics.

Methods: Label powerset (LP) method (Boutell et al, 2004) is a simple but effective multi-label learning method which considers each unique set of labels that exists in the training set as one of the classes of a new single-label classification task and then the multi-label classification problem can be transformed into several singlelabel classification problems. RAkEL (Random k-Labelsets) multi-label classification method (Tsoumakas et al, 2007) is based on LP. RAkEL solves the problems in label powerset (LP) method that the large number of labelsets when the number of labels is large and the inability to predict labelsets not observed in the training set while keeping the advantage of capturing label correlations. The RAkEL method breaks a large set of labels into a number of small-sized labelsets randomly, and for each of the labelsets, a multi-label classifier will be trained using LP method. For an unlabeled instance, the final decision is based on the combination of the results generated by all LP classifiers using the majority voting rule.

We extend the power of RAkEL proposing post-processing strategies for RAkEL. For the tweets assigned no label or multiple labels for purpose and/or position, we find *K* tweets from the training set which are most similar to the original tweet and use the labels from these *K* tweets to make new prediction. Two strategies are used to make new prediction: (1) summation strategy; and (2) weighted summation strategy.

In order to classify tweets, each tweet in the data set is represented as a vector of features. We use some commonly used text classification features, including n-grams (1-gram and 2-gram), punctuation (number of occurrence of exclamation marks, question marks and colons), part-of-speech (POS) and Twitter specific features (e.g. hashtags, the @ symbol, the RT symbol for retweet, URLs and email addresses). We also combine punctuation features and Twitter-specific features as the statistical features and use *STAT* to denote this combination. *POS* is used to denote all the POS features.

Experiments: In the experiments, we randomly choose 600 tweets as the training set and the rest 400 tweets as the test set for each data set. And we compare 5 different methods in the experimental study including: (a) KNN: Since our proposed post-processing strategies are based on KNN model, we use KNN as one of the baselines in the comparison. (b) SVM: the RAkEL method is based on LP method and LP method will use single- label classifers to make predictions. In the experiments, SVM is applied as the single-label classifier, so SVM is used as another baseline. (c) RAkEL, (d) RAkEL+sum: RAkEL with summation strategy, (e) RAkEL+wsum: RAkEL with weighted summation strategy. Since multi-label classification method is employed in the experiments, the Hamming loss is applied as the evaluation measure. However, the single-label classifiers used in the experiments like KNN and SVM cannot be evaluated directly using Hamming loss. Therefore, the purpose labels and position labels generated by two individual classifiers are combined and the combined labels are in the same form of results generated by multi-label classifiers

Results: Here we report some general conclusions from the results: (a) Multi-label classification method RAkEL, no matter with or without postprocessing, can perform better than single-label classifiers, i.e., KNN and SVM in this study on both data sets. However, due to the different characteristics of different data sets, the scales for the Hamming loss are different. Among all the methods, RAkEL with weighted summation strategy performs best. For example, RAkEL+wsum is improved 29.65% compared with KNN on *Obama care* data set. (b) RAkEL with postprocessing using either summation strategy or weighted summation strategy can generate better results than the original RAkEL method which validates the effectiveness of our proposed postprocessing strategies. And weighted summation strategy is more effective on both data sets. For example, RAkEL+wsum and RAkEL+sum can get 7.6% and 5.21% improvement on *Obama care* data set, respectively © In all the results, we can observe that using the combination of *(1+2)-gram, POS and STAT* features performs best. From *1-gram* features to *(1+2)-gram, POS and STAT* features, the more features are introduced, the better performance can be achieved.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. Computational Linguistics, 36(4):673–721.

Zhiting Hu, Poyao Huang, Yuntian Deng, Yingkai Gao, and Eric P Xing. 2015. Entity hierarchy embedding. InProceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), volume 1, pages 1292–1300.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. Journal of machine learning research, 5(Apr):361–397.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

Fabian Pedregosa, Ga¨el Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. The Journal of Machine Learning Research, 12:2825–2830.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In AAAI, pages 1579–1585.

G. Tsoumakas and I. Katakis. Multi-label classi_cation: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3(3):1{13, 2007.

M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. Pattern recognition, 37(9):1757{1771, 2004.

Publications

Li D, Xu Z-M, Chakraborty N, Gupta A, Sycara K, et al. "Polarity Related Influence Maximization in Signed Social Networks". PLoS ONE 9(7): e102199. doi:10.1371/journal.pone.0102199, 2014.

Li, D., Xu, Z., Luo, Y., Li, S., Gupta A., Sycara, K., Luo S., Hu L., Chen, H. "Modeling Information Diffusion over Social Networks for Temporal Dynamic Prediction", ACM Conference on Information and Knowledge Management, San Francisco, CA., October 27-November 1, 2013.

Pei, Y., Chakraborty, N., Sycara, K., Nonnegative Matrix Tri-Factorization with Graph Regularization for Community Detection in Social Networks, International Joint Conference on AI (IJCAI), Buenos Aires, Argentina, July 25-31, 2015.

Li, Yuezhang, Iyer, Rahul, Sycara, Katia "Joint Embedding of Hierarchical Categories and Entities for Concept Categorization and Dataless Classification", submitted to the 26th International Conference on Computational Linguistics (COLING 2016).

Iyer, Rahul, Li, Yuezhang, Sycara, Katia "Detection of Persuasive Arguments in Social Media", the 17th Workshop on Computational Models of Natural Argument, London, UK, June 16-17, 2017.

Iyer, Rahul, Sycara, Katia, Li, Yuezhang "Automated Detection of Persuasion Tactics", invited submission for the Journal on Argument and Computation.

Iyer, Rahul., Pei, Yulong, Li, Yuezhang, Sycara, Katia Simultaneous Identification of Tweet Purpose and Position, to be submitted to SIAM International Conference on Data Mining (SDM18), San Diego CA., May 3-5, 2018.

Personnel Supported (partially)

Yulong Pei (student)
Rahul Iyer (student)
Yuezhang Li (student)
Katia Sycara (faculty)

Transitions

N/A