ABSTRACT
Twitter messages are a potentially rich source of continuously and instantly updated information. Shortness and informality of such messages are challenges for Natural Language Processing tasks. In this paper we present a hybrid approach for Named Entity Extraction (NEE) and Classification (NEC) for tweets. The system uses the power of the Conditional Random Fields (CRF) and the Support Vector Machines (SVM) in a hybrid way to achieve better results. For named entity type classification we used AIDA [8] disambiguation system to disambiguate the extracted named entities and hence find their type.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Linguistic processing; I.7 [Document and Text Processing]: Miscellaneous

General Terms
Algorithms

Keywords
Named Entity Extraction, Named Entity Classification, Social Media Analysis, Twitter Messages.

1. INTRODUCTION
Twitter is an important source for continuously and instantly updated information. The huge number of tweets contains a large amount of unstructured information about users, locations, events, etc. Information Extraction (IE) is the research field which enables the use of such a vast amount of unstructured distributed information in a structured way. Named Entity Recognition (NER) is a sub-task of IE that seeks to locate and classify atomic elements (mentions) in text belonging to predefined categories such as the names of persons, locations, etc. In this paper we split the NER task into two separate tasks: Named Entity Extraction (NEE) which aims only to detect entity mention boundaries in text; and Named Entity Classification (NEC) which assigns the extracted mention to its correct entity type. For NEE, we used a hybrid approach of CRF and SVM to achieve better results. For NEC, we first apply AIDA disambiguation system [8] to disambiguate the extracted named entities, then we use the Wikipedia categories of the disambiguated entities to find the type of the extracted mention.

2. OUR APPROACH
2.1 Named Entity Extraction
For this task, we made use of two famous state of the art approaches for NER: CRF and SVM. We trained each of them in a different way as described below. The purpose of training is only for entity extraction rather recognition (extraction and classification). Results obtained from both are unionized to give the final extraction results.

2.1.1 Conditional Random Fields
CRF is a probabilistic model that is widely used for NER [5]. Despite the successes of CRF, the standard training of CRF can be very expensive [6] due to the global normalization. In this task, we used an alternative method called empirical training [9] to train a CRF model. The maximum likelihood estimation (MLE) of the empirical training has a closed form solution, and it does not need iterative optimization and global normalization. So empirical training can be radically faster than the standard training. Furthermore, the MLE of the empirical training is also a MLE of the standard training. Hence it can obtain competitive precision to the standard training. Tweet text is tokenized using special tweets tokenizer [1]. For each token, the following features are extracted and used to train the CRF: (a) The Part of Speech (POS) tag of the word provided by a special POS tagger designed for tweets [1]. (b) If the word initial character is capitalized or not. (c) If the word characters are all capitalized or not.

2.1.2 Support Vector Machines
SVM is a machine learning approach used for classification and regression problems. For our task, we used SVM to classify if a tweet segment is a named entity or not. The training process takes the following steps:

1. Tweet text is segmented using the segmentation approach as described in [4]. Each segment is considered a candidate for a named entity. A set of features is extracted for each segment and the SVM is trained to distinguish true positive entities from false positive ones. We enriched the segments by looking up a Knowledge-Base (KB) (here we use YAGO [3]) for entity mentions as described in [2]. The purpose of this step is to achieve high recall. To improve the precision a bit, we applied some filtering hypothesis (such as removing
We used Precision, Recall and F1 measures as evaluation criteria through a 4-fold cross validation approach for training and testing. All our experiments are done out of vocabulary (does not appear in training set), we apply AIDA disambiguation system on the extracted mentions. AIDA provides NEC we used AIDA disambiguation system to disambiguate the extracted named entities and hence find their type.

### 5. REFERENCES


