**1. ABSTRACT**

On Twitter, hashtags are used by users to tag the content of their 140 character long tweets which aids in categorization and indexing. Unfortunately due to the vague way hashtags are chosen and defined, many tweets are tagged incorrectly or are not tagged at all resulting in a loss of potentially useful information. We propose ???, a supervised learning hashtag recommendation system that relies on a dual ensemble approach where one implementation generates appropriate hashtags based on the tags of existing training data and the other implementation extracts hashtags based on part-of-speech tags of word phrases in the tweet. We tested our Twitter data set consisting of 60,000 tweets and demonstrated that ??? is successful in recommending relevant hashtags by achieving a ??% precision when asked to find the top 2 most likely hashtags for a tweet.

**Keywords**

twitter, nlp, hashtag recommendation

**2. INTRODUCTION**

Twitter’s ever increasing popularity and establishment as the most popular micro-blogging platform has led it to become a very popular source of current information. Unlike traditional blogging platforms, Twitter limits the size of their user’s publications (called Tweets) to 140 characters forcing the content of the tweets to be precise and most importantly short and timely. This combined with their user base of over 236 million active users per month allows it to be an extremely useful potential source in tracking real-time trends. [Give examples and real life application of trend tracking] To aid in the tracking of tending topics, Twitter utilizes hashtags which are inserted within the tweet itself to help identify the topic of the tweet. Researchers can track rising trends by simply looking at which hashtags are suddenly being used frequently. A hashtag can be identified if it is prefixed by a # symbol followed directly without white space by a string and can be placed anywhere within the body of the tweet [citation needed]. For example the annual bicycle race that takes place in France could result in the hashtag #tourdefrance or #tdf or #france, all of which would be appropriate hashtags for the event. Our example also demonstrates the difficulties for a user in choosing the most correct hashtag as there are an arbitrary number of possible relevant hashtags which may result in two tweets about the same topic being tagged completely different. In many cases, the user may forgo using hashtags all together which would make it impossible to track trends.

As the use of hashtags becomes more popular in blogging sites, the need for a hashtag recommendation system becomes more prevalent to keep the information organized. Tumblr, another blogging platform, has a simple recommendation system based on the content of the user’s blog [CITATION]. Various other researchers have proposed similar recommendation systems for Twitter to solve the tagging problem. These algorithms can be generally categorized into two groups. The first being algorithms that recommend pre-existing hashtags either from a set of predetermined topics [CITATION] or based from the existing hashtags of training tweets. While this system is effective for text categorization problems, using predefined hashtags ignores rising trend tweets as overall more popular hashtags may be favored over the more correct and timely hashtag. In addition, due to the arbitrary nature of hashtags, the correct hashtags may often not exist within the corpus of known hashtags and an incorrect hashtag may be assigned instead. The second group attempts to extract hashtags directly from the words and phrases of the tweet itself. This solves the aforementioned rising trend problem as the hashtag extracted will not be from an older and possibly irrelevant tweet. In section 3 of the paper, we will discuss past implementations that we have based our approach from.

Our recommendation system, fully described in section 4, is a combination of an implementation that generates hashtags from learned data as well as an implementation that extracts hashtags from tweet content based on part-of-speech tags. The rest of the paper will be formatted as such: in section 5 we will discuss our evaluation metrics and results, and we will conclude in section 6.

**3. RELATED WORKS**

The most naïve approach to hashtag recommendation is to retrieve the hashtag(s) of the training tweet(s) that best matches the test tweet. Various studies have applied the cosine similarity and TF-IDF metrics popular in information retrieval to the problem with various degrees of success [CITE THE TRINITY + TFIDIF SOURCE]. The problem with using traditional information retrieval is that these approaches ignore the fact that documents/tweets to hashtags is a 1 to many relationship which greatly decreases the number of hashtags that can be “retrieved”. Otsuka, Wallace and Chiu proposed a novel approach that combats this problem while still loosely sharing a similar structure to the TF-IDF approach called Hashtag Frequency - Inverse Hashtag Ubiquity (HF-IHU). However, unlike TF-IDF which looks at the association between words and documents, HF-IHU looks at the association between the words of a tweet and the hashtags associated with a tweet.

Blah blah blah

**4. ???**

Our proposed hashtag recommendation system consists of an ensemble of two independent approaches. The first approach is a modified version of HF-IHU system proposed by Otsuka [CITATION]. ??? relies on the creation of three data structures: Term-to-Hashtag-Frequency, Hashtag-Word-Popularity and Inverse-Tweet-Frequency. The first data structure called Term-to-Hashtag-Frequency (THF) maps all of the terms of the tweet corpus to another mapping from hashtags to the frequency that the primary key term appears with the hashtag. Hashtag-Word-Popularity (HWP) is a mapping from hashtags to a set of all the unique terms that appear with the hashtag. Finally, Inverse-Tweet-Frequency (ITF) maps unique terms in the corpus to a set of all the tweets that contain the word.

**5. EXPERIMENTAL EVALUATION**