**ABSTRACT**

The aim of this study is to explore authorship attribution methods in Greek tweets. We have developed the first Modern Greek Twitter corpus (GTC) consisted of 12,973 tweets crawled from 10 Greek popular users. We used this corpus in order to study the effectiveness of a specific document representation called Author’s Multilevel N-gram Profile (AMNP) and the impact of different methods on training data construction for the task of authorship attribution. In order to address the above research questions we used GTC to create 4 different datasets which contained merged tweets in texts of different sizes (100, 75, 50 and 25 words). Results were evaluated using authorship attribution accuracy both in 10-fold cross-validation and in an external test set compiled from actual tweets. AMNP representation achieved significant better accuracies than single feature groups across all text sizes.

**Aims of the research**

- To perform authorship attribution experiments in tweets written in Modern Greek.
- Create the first Modern Greek Tweets Corpus (GTC) in order to use it as a reference corpus for studying social media language including authorship attribution, sentiment analysis and linguistic variation.
- Explore the effectiveness of a specific document representation called Author’s Multilevel N-gram Profile (AMNP), which comprises of a combined vector of increasing size and different level n-grams.
- Investigate alternative ways to construct training sets for authorship attribution in Twitter data.

**The Greek Twitter Corpus -GTC**

We compiled a new corpus of tweets written in Modern Greek (see Table 1). In order to extract tweets from the specific users we used the twitterR package. During the corpus preprocessing we removed all @replies, RTs and manual retweets (RTs').

**Experimental Methodology**

**Results**

Authorship attribution in Greek tweets can be performed with remarkable accuracy when we use a training set in which the basic text units contain merged tweets. Best results were obtained using 100-word and 75-words text chunks (0.952 and 0.918 respectively) and using a 10-fold cross-validation procedure. However, when we used the external tweets as test set accuracy rates moved to the opposite direction with smaller text-size chunks producing better attribution rates than bigger ones (see figure 2).

**CONCLUSIONS**

- Authorship attribution in tweets of Modern Greek is a feasible task. Our top performance (0.951 accuracy in 10-fold cv using 100-word text chunks) is a good indication that the tweet’s linguistic structure is a significant carrier of authorship information.
- AMNP representation proved highly efficient compared to single n-gram feature groups in all text sizes.
- Optimal results are achieved when both training and testing sets of authorship attribution contained merged tweets.
- The text-size threshold for using AMNP seems to be the 50 words per text chunk.