Key work paper
“A Methodology to Quickly Perform Opinion Mining and Build Supervised Datasets Using Social Networks Mechanics”

Published in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, number 35 issue 9, pages 9797-9808, SEPT 1 2023

Manuel Francisco
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

Juan Luis Castro
Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain
castro@decsai.ugr.es

Abstract—**Social Networking Sites (SNS) offer a full set of possibilities to perform opinion studies such as polling or market analysis. Normally, artificial intelligence techniques are applied, and they often require supervised datasets. The process of building them is complex, time-consuming and expensive. In this paper, it is proposed to assist the labelling task by taking advantage of social network mechanics. In order to do that, it is introduced the co-retweet relation to build a graph that allows to propagate user labels to their similarity neighbourhood. Therefore, it is possible to build supervised datasets with significant less human effort and with higher accuracy than other weak-supervision techniques. The proposal was tested with 3 datasets labelled by an expert committee, and results showed that it outperforms other weak-supervision techniques. This methodology may be adapted to other social networks and topics, it is relevant for applications like informed decision-making (e.g. content moderation), specially when interpretability is required.**

Keywords— Terms—human-in-the-loop labelling, opinion mining, social network analysis, user profiling, supervised learning.

I. Introduction

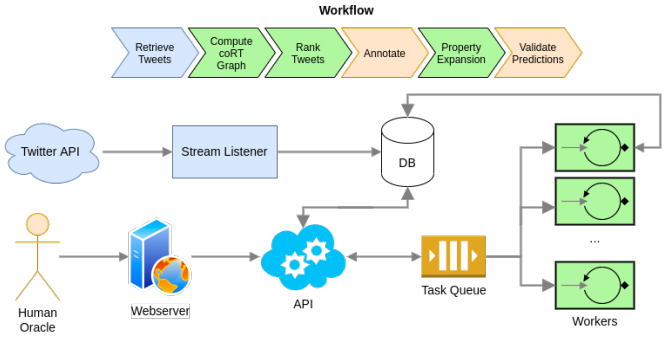
In this article it is proposed a methodology to reduce the effort required to produce a supervised dataset. It is made by using semantic network representations and label propagation mechanisms. This methodology can be divided in different steps that start with data retrieval. Meta-data regarding user interaction is used to compute a similarity graph, and most relevant documents are labelled using oracles. These labels are aggregated and propagated through the graph to obtain information regarding unknown users. Since not all labels are accurate, oracles are asked to validate new information and check for potential conflicts. This makes our proposal a human-in-the-loop approach to produce weak-supervised dataset with improved quality.

The main contributions of this article are (1) a methodology to reduce the effort when producing labelled dataset from SNS data, (2) a novel similarity-based network representation (the co-retweet graph), and (3) a label propagation mechanism that takes into account the relevance and coherence of the network. Both the network representation and the propagation mechanism are two possible implementations of our methodology, that we tested using (4) a proof-of-concept platform that follows aforementioned steps to build weak-supervised datasets from Twitter. Results show that this methodology is able to reduce labour costs by 75 percent and it outperforms other weak-supervision techniques.

II. Proposed Methodology

We propose the use of a system that would allow us to infer properties of unknown users from other previously annotated documents. The basic workflow would be:

1) Rank tweets by utility.
2) Ask an oracle to annotate the top n tweets.
3) Expand properties to other user profiles using a deep relation.
4) Rank automatically-generated user annotations by utility.
5) Ask an oracle to validate the top m autoannotations.
6) Repeat from step 2 until necessary.



Grant PID2021-125788OB funded by MICCIU/AEI/10.1309/501100011033 and by “ERDF A way of making Europe”.

Fig. 1. Workflow of the system

And the keys of the system are:

A. Similarity Graph

It calculates the retweets that any two users have in common. As we stated above, the retweet mechanic implies that the user is interested in the topic and also that they agree with the opinion of the original author. Therefore, if users A and B have retweeted something from C, the retweet graph would have three directed links, from A to C and from B to C; with our proposal, both links would be undirected and a new edge would arise between A and B. Note that original tweets are considered copies (retweets) of themselves. Consequently, each node in the co-retweet graph would stand for a Twitter user and edges connecting

B. Property expansion

Given a similarity graph (in our case, co-retweet graph), it is possible to infer property values for new users in the neighbourhood of known users through a process of weighted

III. Development and Experiments

In order to check the performance of our proposal against current weak-supervision techniques. We applied several classic machine learning algorithms (Support Vector Machine (SVM), Random Forest (RF), AdaBoost (AB) and Multinomial Naive Bayes (MNB). We applied our methodology to datasets of tweets regarding Spanish National Elections (spanish), Madrid Regional Elections (madrid) and USA Presidential Elections (usa).

IV. Results and Discussion

Results shows that the number of automatic annotations grows quickly with the first annotations due to the ranking strategy, and it stabilises after the annotation process reach non-influential users. For example, for the Spanish National Elections dataset, the top 25% has more than 191 connected users, hence the grow. The ratio between number of accepted annotations versus total annotations has a mean value of 0.89, with an almost negligible standard deviation of 0.0045. It keeps steady regardless of the number of manual annotations. This points out that the chosen user representation (co-retweet graph) behaves coherently. Our proposal significantly improves the results of other weak-supervision techniques when the number of annotations is low. When annotating tweets, it is common to assume that the cost of evaluating a tweet is uniform [42], since documents have 240 characters as much. The number of manual annotations is called effort. Table 6 shows the effort required to reach at least .75 in f1-score. Note that we stopped several experiments before they reached the threshold. Our proposal is the method that requires less human annotations therefore it is ideal to reduce labelling costs.

V. Conclusion

SNS are used frequently to study the opinion of customers, citizens, voters and almost any other role a human can assume while using Social Media. However, these analysis present limitations that should be considered, such as users age range and digital literacy. Normally, artificial intelligence techniques can be used to perform these analyses, which often require supervised datasets. Labelling datasets is an expensive task that needs a lot of resources, regardless that it is conducted in-house or outsourced to companies or freelancers.

Most SNS have similar mechanics, such as liking and befriending, that may offer more information than the contentitself. We introduced the co-retweet, that is built upon user interactions and represents the network as a similarity graph. Therefore, it constitutes a way to infer knowledge of unknown users from others in their neighbourhood. In order to do this, we presented a methodology that iteratively propagates labels through a similarity graph, generating predictions that can be reviewed automatically, in most cases, or manually. Our experiments show that our proposal outperforms other weak-supervision techniques when the effort is low, and it behaves at least as well as the number of manual annotations grows.

These results are relevant in the field of opinion mining for applications like market analysis, recommendation system and trend predictions, and it is particularly useful since it significantly improves the required time to build a supervised dataset without sacrificing interpretability or quality.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

Reference

[1] M. Francisco, and J..L. Castro, “A Methodology to Quickly Perform Opinion Mining and Build Supervised Datasets Using Social Networks Mechanics”, IEEE Trans. on Knowledge and Data Engineering 35 9, 9797-9808, 2023.

Grant PID2021-125788OB funded by MICCIU/AEI/10.1309/501100011033 and by “ERDF A way of making Europe”.