Hugh Cunningham
Andrew Lang Adams

CS224U Milestone Report

Project Goals

We will investigate whether combining language modeling with social contextual information can improve sentiment classification of user reviews. In particular, we will use the Yelp Academic Dataset as our source of experimental data and use the "friend" edges in this dataset's social graph to establish social relationships. The intuition behind this approach is that users who share some social relationship are likely to share similar opinions of the businesses that they patronize: friends who visit a restaurant together might share the same collective experience in their reviews, or the review of one user may color subsequent reviews from his or her friends. We will build a classifier that combines social graph features with language features from the review text and compare its performance to baseline classifiers that use only language features.

Previous Approaches

Previous work in sentiment analysis by Thomas et. al (2006) and Tan et. al (2011) inspires our direction in this project. Thomas et. al successfully use relationships between pieces of text to supplement language features in the sentiment classification of segments of Congressional speeches. While the relationships between pieces of text (as used in Thomas et. al) clearly differ from the relationships between authors of that text (which Thomas et. al do not examine), the methodology of the former type of relationships could be useful in determining the effects of the latter type. Thomas et. al train an SVM on language features of speech segments for sentiment classification, employ an additional SVM to identify agreement edges between speech segments, and add weights established from agreement edges to the sentiment classification of their initial SVM. It is not clear yet whether the existing social graph edges in our data set will be sufficient on their own to model social context or whether, like Thomas et. al, we will require more complex methods to appropriately establish the relationships we are looking for. However, the classification framework in Thomas et. al provides a useful blueprint for effectively combining language features with relationships across text segments in a simple model.

The goals pursued by Tan et. al are more similar to our own; using social network information from to improve upon the sentiment classification of classifiers that use only textual information. Tan et. al model two different types of social connection from Twitter in what they refer to as the "@-network" and the "follower/followee network" and additionally suggest different social phenomena driving each. For the "@-network", in which users mutually tweet at one another, Tan et. al posit that *homophily* (similarity and connection tend to co-occur) will produce shared sentiment among connected users. In cases where the social connection between users is follower/followee, Tan et. al suggest that the follower's approval or desire to pay attention to the followee may result in the follower adopting the sentiment of the followee. At this point we plan on modelling only the former type of relationship and social phenomenon. Yelp does appear to possess data of the follower/followee type in the case where one user is a "fan" of or "compliments" another user, but these connections are not included in our data set except in terms of raw counts. While the goals and results that Tan et. al achieve serve as inspiration for our own efforts, the complexity of their approach will prevent us from proceeding directly from their work; Tan et. al use a fairly complicated graph approach that exceeds our own command of graph algorithms. However, where Thomas et. al provide a blueprint for building the classifier, Tan et. al provide guidance in modelling social relationships.

Current Approach

As indicated previously, we plan to use the social graph of the Yelp Academic Dataset to model social relationships and to incorporate these into a sentiment classifier using only textual features of the reviews in this dataset. The dataset is freely available at http://www.yelp.com/dataset_challenge/ and contains >300k reviews and >150k edges in its social graph. The data implicitly models social relationships as binary relationships between pairs of users: two users are either friends or they have no identifiable social connection. These binary relationships may prove to be sufficient given our modest goals, but we would like to explore deeper models of social relationships between users. For example, social graph edges might be given higher weight based on the number of friends that a pair of users has in common or based on the number of businesses that the pair of users have both reviewed. Intuitively, a pair of users who share most of their friends may be thought of as more similar in that they belong to a particular group of users and a pair of users that has reviewed a large number of the same businesses may be thought to have similar taste.

In addition to the including information on the overlap between sets of friends and sets of businesses reviewed, we are interested in including temporal information in our model. In particular, we may say that if a user u_1 reviewed a business after a friend of his or hers u_2 reviewed the same business then u_2's review might have influenced the sentiment of u_1's, but not the reverse. Similarly, we may be able to identify if a pair of friends visited a business together, if, for example, their reviews of the same business are created on the same day. Two people who visit a business together seem very likely to share a similar opinion of the experience. However, as we will note below, further work is needed before we will fully realize how to extract or incorporate such features.

Progress and Further Work

Our progress to date resides primarily in the realm of decision-making: we have decided on a problem and a goal (defined above), we have decided on a dataset (Yelp Academic Dataset), and we have decided on a language to utilize in constructing our models (Python). We have acquired our dataset, begun to explore it, and identified potential features as discussed above. However, we still have work to do in identifying which examples should be used for training and test data. For instance, we may use only reviews from users whose set of friends is non-empty since our model would incorporate no additional information beyond the text of the review. Similarly, we must decide whether we will construct a multiclass classifier based on the star rating of each review (1-5 stars) or limit ourselves to a binary classifier (e.g., negative =1-2 stars, positive = 4-5 stars, and 3-star reviews omitted).

Based on the results of Tan et. al, we know that a social graph, even a sparse one, can yield significant gains in sentiment classification provided there is a strong correlation between social connectedness and shared sentiment. To that end, exploration of the data to establish the extent of this correlation, similar to that performed by Tan et. al, will be very important to establish expectations for and evaluation of our final results. Further exploration is also likely to suggest steps we need to take with regard to cleaning the data for our purposes.

Our choice of Python for programming is logical as it is the only language in which we both have recent, non-trivial experience, and we are aware of popular packages for machine learning (e.g., scikit-learn). However, while we are both aware of such packages, neither of us has any direct experience in using them. We are both currently working on familiarizing ourselves with the landscape of Python libraries related to our task and with scikit-learn in particular.

It also remains to decide what our baseline classifier will be and to implement it. This decision will be influenced by what we are able to glean from the programming libraries that we leverage, but previous approaches by Thomas et. al and Tan et. al suggest that an SVM using only textual features will provide a sufficient and meaningful baseline. In the same vein, we have yet to choose metrics for evaluation, but Thomas et. al and Tan et. al represent useful suggestions for this decision (Thomas et. al use precision and accuracy, Tan et. al F1 and accuracy).

Obviously, the bulk of the work that lies ahead is in the actual implementation of both our model and baseline(s) since we have yet to implement anything. We are confident that we will be able to implement our baseline classifier and a binary model of social relationships without much difficulty. We anticipate that the main source of obstacles in our work will be in extending this model of social relationships in any or all of the ways discussed in the previous section. The project is admittedly still in its nascency and we will be glad of any deficiencies or oversights that you might point out.