Hugh Cunningham
Andrew Lang Adams

CS224U Literature Review

We review seven papers solving problems in sentiment analysis, broadly defined as the analysis of an author's attitude toward a topic or towards his or her audience. Beginning with a review of state of the art techniques from Socher et. al for sentiment analysis in the sense of opinion classification, we then examine techniques developed by Tan et. al and Thomas et. al that leverage links between conversational agents to aid in this sentiment analysis task. The use of social relationships to augment text-level sentiment analysis leads us to investigate the use of natural language processing to classify socials relationships in work by Bramsen et. al and Wang et. al, and finally work by Ranganath et. al and Jurafsky et. al on detecting flirtation and extracting social meaning. This collection defines a unifying theme of examining problems in the interaction between sentiment classification and social relationship classification.

Socher et. al 2013 —Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Socher et. al's 2013 paper introduces two tools for sentiment analysis - the Stanford Sentiment Treebank and the Recursive Neural Tensor Network (RNTN). Each of these alone makes a significant contribution to the advancement of sentiment analysis, and their combination in this paper defines the state of the art in positive/negative sentence classification.

The fully labeled parse trees of the Stanford Sentiment Treebank corpus facilitate the analysis of compositional effects of sentiment in language. Use of the sentiment treebank shows performance exceeding 80% for single sentence positive/negative classification for even the baseline techniques (Naive Bayes, SVM, etc.) considered in this paper. The construction of the sentiment treebank also makes way for more fine-grained sentiment classification, as parse labeling used a continuous scale from 'very negative' to 'very positive' sentiment in the extremes.

The RNTN represents a further iteration on previous work by Socher et. al which used both standard recursive neural network (RNN) and Matrix-Vector RNN (MV-RNN) models for the sentiment analysis task. The RNTN overcomes problems identified with these previous approaches (a lack of interaction between input vectors in standard RNN and a large number of parameters in the MV-RNN), and exceeds the performance of these models for both binary, positive/negative sentence classification as well as the more fine-grained sentiment task undertaken in the paper. The RNTN also shows a great deal of promise in capturing negation by far exceeding the performance of both the standard RNN and MV-RNN in negation detection. The performance of recursive neural models in classifying fine-grained sentiment in short phrases is encouraging for future work in this area, particularly given the proliferation of short natural language communications on Twitter or via SMS.

Tan et. al 2011 — User-Level Sentiment Analysis Incorporating Social Networks

Tan et. al approach the problem of user-level sentiment analysis motivated by the idea that connections between users in social networks may improve classification of those users' utterances. Intuitively, users that are connected to one another through some personal relationship may share similar sentiments with regard to many topics, and one user's following another on Twitter suggest the follower's approval of the sentiments of the followee. To this end Tan et. al use connections in user graphs as features for classifying each user's sentiment on specific topics ("Obama", "Sarah Palin", "Glenn Beck", "Lakers", and "Fox News") as either positive or negative.

The paper experiments with four different representations of user connections - @-network and *t*-follow, each in both directed and undirected graphs - and compares two algorithmic approaches to sentiment classification using an SVM. Though performance improvements over the SVM are modest for both @ graphs, Tan et. al demonstrate that these social network connections do provide useful features for sentiment classification. In particular, they find that significant performance gains are achievable where the underlying graph shows strong correlation between user connectedness and shared sentiment even in the face of sparse graph connections. This suggests a potentially fruitful synergy in combining natural language processing with social network analysis for the task of sentiment classification.

Thomas et. al 2006 — Get out the vote: Determining support or opposition from Congressional floor-debate transcripts

Thomas et. al leverage the relationships between discourse segments in classifying the sentiment of individual segments. For example, if a given speech indicates agreement with the sentiment of another speech, then, intuitively, the former should receive the same label as the latter. The paper aims to analyze sentiment in multi-speaker discussions of Congressional bills, and defines "positive" sentiment as support for the bill (voting records provide a convenient source of labeling for most speeches).

The paper points out an important ambiguity in the type of language used in Congressional floor-debates: persuasive language is often characterized by the presentation of evidence in support of an attitude rather than the explicit statement of that attitude. This highlights the problem of some text (e.g., discussions) being more difficult to classify with regard to sentiment than other text (e.g., reviews). The authors observe that the former such text should be easily classifiable if it provides evidence of agreement with a text that is more clearly identifiable as either positive or negative.

Employing agreement links as features in classification yields significantly improved performance over the baseline SVM classification, and the authors note that constraining all speech segments uttered by the same speaker to receive the same label improved accuracy even further. This latter observation suggests that agreement links are particularly useful for classifying sentiment at the speaker-level.

Bramsen et. al 2011— Extracting Social Power Relationships from Natural Language

The Bramsen et. al paper focuses on classifying the relationships between people in email threads, and social power relationships in particular, from linguistic features. This task departs from the conventional idea of sentiment analysis as classifying the opinion of a writer toward the classification the attitude of the writer's communication. The authors do not classify their own work as sentiment analysis as such, and prefer the designation 'Social Power Modeling' (SPM), but they acknowledge a certain synergy in using sentiment analysis and SPM to determine the opinions of respected members of a community.

Even if not specifically designated as sentiment analysis, Bramsen et. al's work utilizes a familiar approach: they train a model to classify natural language communication in a binary fashion as either 'UpSpeak' (from subordinate to superior) or 'DownSpeak' (from superior to subordinate) based on n-gram modeling and POS tagging. Broadly construed, such a model attempts a classification of the attitude of a communication. Bramsen et. al use n-grams binned according to their relative frequency in each class as their primary features and achieve performance comparable to the baseline bag-of-words models even while limiting their classifier to only eight binned n-grams for features. Performance improved with the inclusion of polite imperatives in the feature set. The paper demonstrates that natural language processing provides a promising avenue for the classification of social relationships between people and for the classification of the attitude of the writer toward the reader in written communication.

Jurafsky et. al 2009— Extracting Social Meaning: Identifying Interactional Style in Spoken Conversation

Using a professionally transcribed spoken corpora of 991 4-minute speed dates and incorporating lexical, dialogical, and prosodic features, Jurfafsky et. al sought to classify conversational style and social intention—namely, whether an individual in speed-dating is judged by an interlocutor of the opposite sex to be *friendly, awkward,* or *flirtatious.* For these binary classifications, Jurafsky's model made effective predictions with up to 75% accuracy (with 50% as the baseline for a binary classification). As the only published paper in the ACL library on romantic flirtation, the paper has implications for social science, but it also has strong implications for extraction of meaning more generally speaking, because of the group demonstrated the ability to extract dialogic features with quite simple techniques.

Jurafsky's group used (fairly) standard lexical and prosodic features: their prosodic features were extracted from RMS amplitude and F0 features, and their lexical features mainly drew upon the Pennebaker's widely influential LIWC lexicons (The two additional lexical features beyond the LIWC features were "past tense auxillary" to heuristically detect storytelling and 'metadate', which reflected discussion about the speed-date itself).

As for dialog and disfluency features, Jurafsky et. al utilized the following features: **backchannels** (the number of utterances like "uh-huh", "oh, okay"), **appreciations** (utterances like "wow", and "that's true"), **repair questions** (utterances like "wait, excuse me") and **collaborative completion,** which is when the speaker completes an utterance begun by the alter.

Admittedly, Jurafsky et. al submit that their techniques for extracting these dialog and disfluency features were "shallow". To extract the presence of repair questions, for instance, Jurafsky et. just regexed for the presence of "wait" or "excuse me". To extract for collaborative completion, they used a "simple heuristic [that] was errorful, but did tend to find completions". Specifically, they used an interpolated trigram model and marked any turn with a p >.01 (an arbitrarily chosen threshold) as a collaborative completion. Despite the rudimentary means of extraction, collaborative completions and other dialogic and disfluency features were among the most predictive features (suggesting an avenue for future refinement).

As for the social science takeaways, Jurafsky et. al's model demonstrated many insights, such as "both genders convey friendliness by laughing more, and using collaborative completions" and "men ask more questions when (labeled as) flirting, women ask fewer.

Ranganath et. al 2009 —It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates

Using the same set of features and data set, this paper by Ranganath, Jurafsky, and McFarland extends the previous research done by Jurafsky et. al on detecting *perceived* flirtation. In this paper, Ranganath et. al build a model to detect whether speakers themselves *intend* to flirt, and they also explore the differences in ability and inability of the sexes to *correctly* perceive intended flirtation cues. In terms of performance, their automatic model detects flirtation better than do male (71.5%

model accuracy versus 56.2% male human accuracy) and female (69% model accuracy versus 62.2% female human accuracy) interlocutors. Technically speaking, this research advanced a neural net technique for capturing domain specific lexical cues that the LIWC features failed to effectively capture.

While this research paper postulates many interesting social science takeaways akin to those found in the group's previous paper, the most relevant contribution to sentiment analysis in conversational dyads, is the use of a lexical dimension reduction. In Jurafsky et. al, the group found that the LIWC lexical features were much less useful than prosodic and dialogic features in detecting perceived flirtation. Ranganath et. al hypothesized that this was because flirtation lies in a different class of words than previously investigated. Due to issues of sparsity, Ranganath et. could not generalize lexical features by using a high-dimensional vector of word. Instead, they reduced the 2000 most commons words into a high-level low dimensional space of 30 features, utilizing an autoencoder (neural net). The success of the autoencoder suggests that it "shows potential for being a promising feature extraction method for social tasks where cues are domain specific.", and it would be interesting to see future work utilize this technique.

Wang et. al 2012— "Love ya, jerkface" using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens

In this paper, Wang et. al investigate how dialogue between conversational agents changes over time—particularly, in their use of positive and impolite language. This paper suggests a result believed to be true by academic psychologists (such as Tickle-Degnen and Rosenthal) but contrary to "the vast majority of computational approaches to rapport-building in dialogue". Namely, this paper suggests that "positivity decreases over the course of a relationship."

As a data set, Wang et. al hand annotated a history of chat logs between 130 high school peer tutors and their tutees. 54 of these dyads signed up as **friends** whereas only 6 signed up as **strangers.** For all of the logs, the two raters encoded language behavior labels indicative of impoliteness or positivity. For instance, labels included *Insults* ("you are so weird"), *Condescensions* ("Tutee: nothing you have done has affected me what so ever"), and *Pet name* ("whats up homie?").

Having encoded their data, Wang et. al built a model to predict positivity or impoliteness in the next turn. In addition to these aforementioned language behavior labels, Wang also integrated POS and lexical features. To overcome sparsity in their model, Wang et. al utilized a sparse log linear model with a Lasso penalty. Overall their results were good: the p values for the full models (behavior labels, POS, lexical features) predicting impoliteness in the friend dyads and for predicting positivity in the friend and stranger dyads were respectively statistically significant (.003, .001, .019), although the positivity model did not predict impoliteness among stranger dyads with statistical significance. Interestingly, the model predicting impoliteness utilizing just behavior labels (excluding POS and lexical features) predicted impoliteness among friend dyads with statistical significance (.017)—the only prediction based solely on language behavior labels that was statistically significant. Wang et. al conclude that "these results suggest that positivity is a predictable behavior among strangers, who may all express uniform positivity across all dyads, while it is the impoliteness that is predictable among friends". Wang et. al find other interesting divergences between the stranger and friend dyads—for instance, tutee bragging predicts with high likelihood (.7) tutor positivity, whereas tutor bragging predicts with an exceedingly high likelihood (.96) the tutee responding with impoliteness (the rationale being that a tutee is perhaps threatened by a tutor's bravado, whereas the tutor is encouraged by his/her tutee's confidence).

Comparison

Each of the papers reviewed demonstrates an interest in moving beyond simple bag-of-words models for natural language tasks towards using feature sets that leverage a fuller understanding of the contextual information of an utterance. For Socher et. al this understanding is developed from the use of the sentiment treebank and the RNTN to model compositionality of sentiment. Ranganath et. al also turn to deep learning techniques in using autoencoders for detecting social intentions. In leveraging connections between Twitter users for sentiment classification, Tan et. al recognize the similarity between their approach and that taken by Thomas et. al in using agreement links between dialogue segments. Furthermore, both papers find that the inclusion of links between agents yields significant improvement over baseline models.

While Socher et. al make significant advancements in the sentiment classification of individual utterances (both binary and fine-grain), the work of Tan et. al and Thomas et. al may be more informative in understanding sentiment at the level of the individual agent. User-level sentiment forms the core of Tan et. al, but for Thomas et. al agent-level sentiment results from the restriction of segment labeling to one label per speaker. This constraint effectively gives a sentiment label to the speaker, which matches the assignment of labels in Tan et. al. For some applications, this level of sentiment analysis may be the more pertinent one (e.g., a single utterance may not be a good proxy for determining whether a voter supports a particular politician or legislation).

Bramsen et. al's topic differs from that of Socher et. al, Tan et. al, and Thomas et. al in that the authors do not explicitly address the problem of classifying opinions from utterances. The link between Bramsen et. al and the other papers mentioned above lies in the attempt of Bramsen et. al to use statistical language modeling to extract the types of relationships that could be the bases of features in explicit opinion classification. Wang et. al shows that the task of sentiment analysis may be heavily influenced by the social roles of interlocutors, such as the power roles that form the basis of Bramsen et. al. For example, bragging by tutee predicts with high likelihood tutor positivity, whereas tutor bragging predicts the opposite sentiment (impoliteness) with an exceedingly high likelihood. In Jurafsky et. al, likewise, sentiment could mean one thing or it could mean the opposite, depending on social role (i.e. Friendly women are very disfluent, whereas it is *un*friendly men who are are disfluent).

Future Work

The problems that the above papers address suggest a number of avenues for further exploration. Using approaches like those in the work of Bramsen et. al or Ranganath et. al might reveal social relationships that could be leveraged in sentiment analysis (i.e., as features in a classifier like those designed by Tan et. al or Thomas et. al). Furthermore, none of the papers by itself addresses how power relationships or roles might determine sentiment (Wang et. al, which deals with *prediction* of next turn sentiment being the paper that comes closest). There may be a great deal of interaction between how one agent's power influences the sentiment of another agent. For instance, along the lines of the work of Thomas et. al, it may be of interest to see how the opinions of more senior congressmen and congresswomen influence those of their juniors. Several other problem areas are currently attempting to leverage the techniques introduced by Socher et. al (e.g., entailment), and attempts to extend the sentiment treebank or RNTN to a different task in sentiment analysis could be fruitful.

As an alternative tack, it could be very interesting to investigate how metaconversation referencing the social relationship itself affects sentiment. In Jurafsky et. al, for instance metadate information was an important negatively weighted feature in predicting flirtation perception. As well, in Ranganath et. al the top negative weighted words for predicting flirtation were “long”, “school”, “phd”, “years”, “stanford, “research”, “education”; while Ragnanath attributed these negative weightings to “conversation[s] focused on the mundane details of grad student life”, it is possible that this type of metaconversation wherein agents explicitly reference their roles could be generalized as indicating lack of positive mutual sentiment (i.e. not just lack of flirtation).

Finally, there exists the possibility of replicating some existing studies while making iterative improvements. For instance, some of the approaches in Jurafsky et. al and Wang et. al while fruitful were not particularly sustainable. Namely, Jurafsky's approach for extracting collaborative completion and other dialogue features (trigram, regex) were errorful and shallow, and Wang et. al's approach of hand labeling language behavior could be similarly improved upon in a more generally applicable and sustainable manner.

Implementing the algorithms from Socher et. al also presents a challenging but intriguing possibility.

Bramsen, Philip, et al. "Extracting social power relationships from natural language." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

Jurafsky, Dan, Rajesh Ranganath, and Dan McFarland. "Extracting social meaning: Identifying interactional style in spoken conversation." *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009.

Ranganath, Rajesh, Dan Jurafsky, and Dan McFarland. "It's not you, it's me: detecting flirting and its misperception in speed-dates." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.

Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2013.

Tan, Chenhao, et al. "User-level sentiment analysis incorporating social networks." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.

Thomas, Matt, Bo Pang, and Lillian Lee. "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts." Proceedings of the 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006.

Wang, William Yang, et al. "Love ya, jerkface: using sparse log-linear models to build positive (and impolite) relationships with teens." Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, 2012.