Correlating Sentiment in Tweets and News Headlines: Towards Retrospective Opinion Mining

Zubin John
Department of Computer Science
The Ohio State University
Columbus, OH 43210
johnz@cse.ohio-state.edu

***Abstract*—The growth of social media has led to its prominent use as a research tool. We use it in conjunction with building large datasets for experimentation and for generalizing trends and under-currents in public outlook. This citizen-authored form of journalism is quite intuitive in mirroring public opinion today but we have never put it to the test on past historic data. While our long term goal is to be able to pick out sentiment for historic events through co-relating them with modern events, this paper covers a preliminary study on feasibility of interpreting sentiment from tweets and news headlines.**

I. Task Definition

Twitter has often been used to link text sentiment to public opinion[1] and study does shown that it seems to be a satisfactory indicator of general public sentiment. The goal of the proposed system is to build a model trained along these lines using text sentiment from social media regarding modern-day events and then perform regression to gauge sentiment: outrage, affinity, apathy etc. related to major historic events. The purpose of this is to re-evaluate the efficacy of social media as a tool for public expression.

Further this incites study into why, if the model predicts reactions that do not conform with history, these irregularities do occur and is there a general trend to them. In an alternate time-line where social media is a perfect embodiment of the world's opinion space, was it societal development that generated this alternative outlook. If successfully constructed this model can engender many interesting questions and their answers. some of them are:

- If our model is historically accurate, can we divide and characterize the history of humanity using social media features?
- To what point in time can we project accurate predictions of public sentiment?
- Lastly, study the evolution of the freedom of speech and press from a digital outlook.

These however are the long-term goals for the project. As of now we will seek to perform the pre-liminary investigation into how an experimental system will perform on available data. Most of the data for this has been adapted from the 2007 SemEval 2013 Affective Task[3] . We have also used the research from the paper as a basis for our system. To this end we have achieved quite satisfactory results which outperform most of the systems cited in the paper, which came as a surprise. Since the task is quite abstract (sentiment from short text) the results are nowhere near acceptable state-of-the art standards. Our best systems correlate with the SemeEval results to give similar performance.

This brings us to establishing our hypothesis after experimentation; with sufficient evidence we can conclude that retrospective sentiment analysis is quite feasible with any of the two: tweets or news headlines by co-relation of the two past and present model data.

II. Related Work

Retrospective sentiment analysis by co-relation seems to be a task of interest to very few. There is currently no work in this area, surprisingly given its potential in Digital Humanities and as a powerful tool in Digital Journalism. We however have high hopes for this domain of Sentiment Analysis and would like to stress on it being as important as opinion mining for customer feedback, predictive recommendation systems and marketing-based applications that are being driven by sentiment analysis.

The SemEval 2013 Task 14 lays the foundation for most of our work. Mihalcea et al. have established two types of systems as a part of their work, the first targets valence score determination (polarity) of these news headlines while the second determines sentiments such as anger, disgust, fear, joy, sadness and surprise. The second approach implies a more lexicon-based approach and we are as yet unsure how this would tie in with our approach; as a result we have used polarity determination as the task for our system with minor adaptations on the data set. The similarity between their system and ours is that both of them rely on news from online media. The SemEval data-set is composed of news data scraped from news sites such as Google news, CNN and other web-sites. Most of the systems that have implemented valence scoring for the news headlines claim to have used unsupervised methods however a number of them use supervised methods (CLaC-NB and SWAT) however to poor effect. The best results have been exhibited by the unsupervised CLaC system which implements domain-independent headline valence detection and scoring. A short summary of results from their work has

	r	Acc.	Prec.	Rec.	F1
CLaC	**47.70**	**55.10**	**61.42**	9.20	16.00
UPAR7	36.96	55.00	57.54	8.78	15.24
SWAT	35.25	53.20	45.71	3.42	6.36
CLaC-NB	25.41	31.20	31.18	**66.38**	**42.43**
SICS	20.68	29.00	28.41	60.17	38.60

Table 1: System results for valence annotations

been presented in Table 1. For comprehensive descriptions of these systems please refer the literature.

Other work that use retrospective sentiment analysis in a manner; though in a marketing domain is Ding et al.[5] which performs stock market analysis on time series data. The work makes heavy use of the NLTK[4] and Naives-Bayes classifier for the task. The authors have made predictions on stock for data from January 2008 to April 2010 using SVM, linear regression and neural networks. The best results for these predictions were from the SVM classifier achieving an accuracy of 51% in a 5-day context window. They further boosted performance by including sentiment scores from a supervised Naive-Bayes system, They reported that this combined approach increased performance in 8 out of 12 cases and achieved an overall performance boost of 1.11% towards accuracy.

Previous research work such as Go et al.[6] also provides an effective solution to a possible problem for sentiment analysis on Tweets. Hand annotation is a difficult and expensive task, more so for tweets which are often abstract. Go et. al. suggests distance learning to extract sentiment from emoticons. They contrast SVM, Naive-Bayes and Maximum entropy learners to perform sentiment classification and achieve a performance of 82.2% on their data-set. This method performs quite well however would no be very effective in the absence of emoticons for training. Several other research groups have focused on using emoticons, and some of the best results have been achieved for this kind of training.

Perhaps the most well-known and cited work in this area is O'Connor et al.[1] which details the attempt to link text sentiment to public opinion time series. This preliminary work uses extremely simple text analysis to corelate polling data on consumer confidence and political opinion, subsequently predicting future movements in the polls.

III. Approach & Methodology

The system we are working with in this project is just preliminary work for a much larger project. The extended scope of our work will involve working with a broader data-set which clusters tweet-stories together based on events. We intend to build these event-models out of tweets under a hashtag or category. Once these tweets are clustered and an average score will be assigned for the polarity of each cluster. This will be determined by a suitable weighted valence function. We will then build similar events containing a cluster of headlines from historical news sources. A suitable regression algorithm will then be used to predict valence scores for these historical events.

After having painted a broader picture of our goals, in this mini project we intend to implement the sub-system for predicting valence scores/polarity from the trained corpus on our test data. We will be using different data-sets now and later. For the final data-sets we want to mine a large number of tweets under events defined by a hashtag, this will require use of the Twitter streaming API[13] to collect data over time. We intend to then translate our learners on this new data-set and observe results for those.

A. Data

For this take, we have adapted the SemEval 2007[3] data-set; however we have fit the data to our purpose. The original SemEval data consists of a set of 1000 head-lines with valence scores which could be used for training and a validation set of 250 news headlines with provided valence scores. Since we will be performing the task using a training set composed of news headlines and in tandem a training corpus of tweets we also used a data-set sourced by the CrowdFlower Open data Library[14]. The twitter corpus we make use of contains about 6090 tweets containing tweets around global warming. This data has been hand-annotated by a panel of human judges and contains excellent confidence scores for most of the data. However, since the classes are highly balanced due to the real-time nature of this data we have performed pre-processing on the data to make it better suitable for our purpose.

The pre-processing on the data includes only data with high confidence scores of greater than 60%. This does not solve the problem of balancing the classes however it does an excellent job of fortifying the strength of our training corpus. Since the training tweet set is highly selective and contains only a single event we will probably not achieve as good results as compared to having a multi-model for about 30 scenarios which would have been ideal. However, working with the data we have in present we have been able to achieve results equivalent to those achieved by the headlines based sentiment classifier. It does solve our need of determining whether the feature space and approach in mind is worth moving forward with or not.

B. Feature Space

Most of the systems in the SemEval task make use of features such as Unigrams, Bigrams and POS tags for training the classifier however for our purpose we will select a more informative feature space viz. the TF-IDF vector. The term-vector model has various advantages over the POS tags n-gram based feature spaces. Some of these are the ability to represent similarity amongst tweets and headlines with a stronger relation than n-grams or tags. Another is the relative

	Classes	SVR	Lin. Reg.	Ridge
News Headlines Based	Positive	**0.633855898024**	0.619694500057	0.632009289585
	Negative	0.496167626893	**0.561740051695**	0.561239129583
	Neutral	0.365367721177	**0.501974513489**	0.500974252503
	Average	0.498463749	0.561136355	**0.564740891**

Table 2: Comparing classifiers for Headlines based training

	Classes	SVR	Lin. Reg.	Ridge
Tweet Based	Positive	0.628532155052	0.606755790003	**0.632930885534**
	Negative	0.271250212613	0.372479172343	**0.454538069988**
	Neutral	0.353523015954	0.384377561117	**0.429272976864**
	Average	0.417768461	0.454537508	**0.505580644**

Table 3: Comparing classifiers for Tweet based training

ease with which dimensionality reduction is possible on a term vector. POS tags may also inherently be a good feature space for the task as they are able to deliver information regarding the tense and action in the short phrases we encounter. Verbs are supposedly great indicators of sentiment features and in most lexicon based approaches, it is the verbs that are the most contributing feature to the overall polarity of the sentence. Finally, we believe that LIWC[15] would also have made a great feature addition to our project. However, with the unavailability of the library at hand we were unable to integrate this as well, into our system.

C. Classifiers

The classifiers for the system are as important a component as the feature space in itself. We pored over our decision of which classifiers to use for our implementation; from the survey of the implemented systems in Mihacea et al.[3] it was clear that most Bayesian learners did not perform well on the task irrespective of the features used and hence we choose to look at alternative systems for our approach. We wanted to exploit the linear relationship of our data and short-listed common regression methods to implement.

- Linear SVM Classification
- Linear Regression
- Ridge Regression

Intuitively, these linear classifiers should complement our feature space of their term frequency vector. The measure of performance we will be using to evaluate the different systems is accuracy of prediction. The loss function is the root-mean square error: which is computed as one or zero over each class. With the accuracy scores for each class in the classifier we can then sum over the classes to give a measure of average performance for the classifier. Most classifiers tend to perform the best with the positive sentiment classes as oddly enough the general sentiment over most of our training data is positive; which makes sense for tweets (as a general trend people tend to tweet positive events in their life; also most of the topical tweets in twitter buzz are about positive events), however we would expect news headlines to be largely neutral. Neutral and Negative sentiments are unbalanced in both out training data-sets and hence the poor results can be partially attributed to the data.

IV. Results and Discussion

Table 2 and Table 3 show the results for the two different training sets, our headline based corpus results are recorded in Table 2 and the tweet-based-corpus trained classifiers have been recorded in Table 3. We can clearly see that our headlines trained classifiers perform comparatively to the systems cited in the SemEval paper. The best accuracy reported by our system is with the Ridge Regression based learner. The average accuracy over all the classes comes out to be **56.5%** for the Ridge Regressor based model which out performs all the systems from the Mihalcea paper. Our take to this is that the strength of our feature-space and the suitability of Ridge Regression for this task is what gives superior results.

Now to the more interesting part, new-headlines based training and test analysis for such systems has met mixed reactions from the research community. So much so that, perhaps the unconvincing nature of the results have discouraged further exploration in the field. However, our results indicate that their is possibility of achieving better results with the task given tweaking in the learning space and methods, as well as better tuned data-sets. If a preliminary probe can out-perform existing systems; it only goes to show that there is scope for better results.

The novel idea here is the training of the system based on a tweet corpus. Untried as it was, our results go on to show that if some day news-headline based systems achieve human-level accuracy at the task; tweet-based sentiment prediction will not be far behind. A testament to this theory, our results from experimentation - our ridge regression based tweet-sentiment classifier achieved better scores than the CLaC-NB and SICS systems from Mihalcea et al. Neither of the work is exactly state-of-the art. While our system does achieve an average accuracy of **50.6%**, existing systems in research have already achieved results of **60.8**% using a Senti-word model (see Agarwal et al.[16]). All in all, twitter-based sentiment classification on short text has a strong influence that is bound to grow in the research sphere.

Finally, let us observe the class-wise performance of all our systems. The figure visualizes the performance of each individual over the classes. It is easily visible from *Figure 1* that the best average performance over all classifiers is for

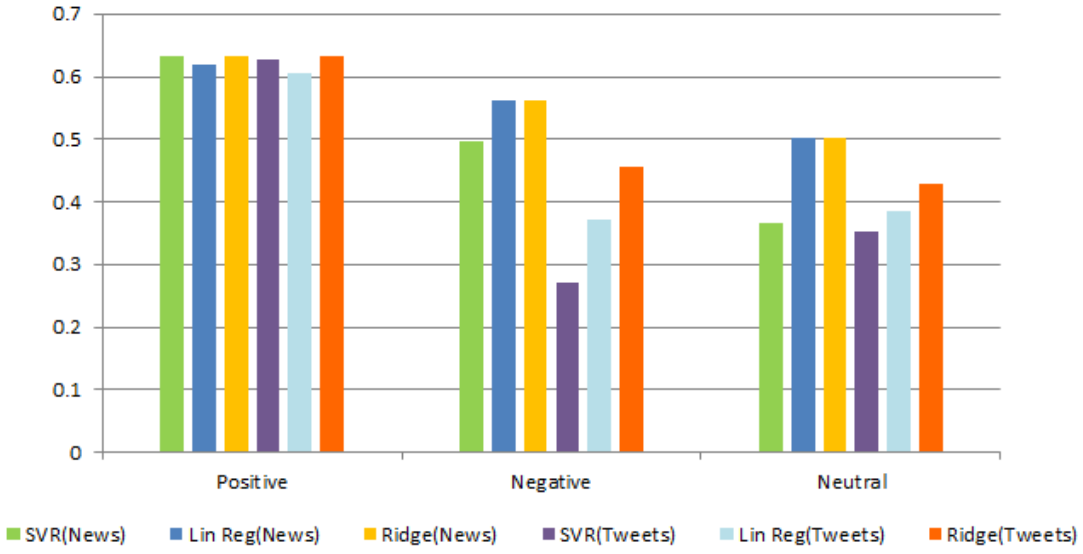

Fig 1: Classifier and class performance distribution.

the the `positive` sentiment class. Whether this is because of a better fit of the data towards the said class or because of over-fitting of data remains to be explored. However, the second does not seem to be the case as our term vector seems to display linguistic diversity without statistical dominance of a particular feature and hence a class bias. This matter does however require thorough examination before stating pre-mature conclusions.

Our concluding visualization again records the average classification accuracy for the classifiers from Table 2 & 3 however in a more visually appealing sense. It also enforces our observation that majority of the classifiers are well matched to the task; hence if any further improvement is to be made it must be done on the data or the feature space and not on the underlying classification model.

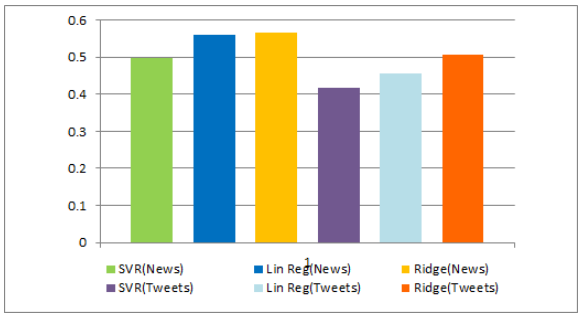

Fig 2: Average classifier accuracy distribution

V. Extensions & Future Work

The redundant point here being that, the work and experimentation conducted here is for a project of a larger scope. The final expectation of the system is to predict and hence engender retrospective analysis on historical data. However, the first road-block here being the ability to classify sentiment from tweets. With the current level of sentiment analysis systems it is difficult to envision any useful findings on historic data through textual analysis. If we were to achieve systems that could state-of-the-art sentiment extraction from tweets we might move ahead with retrospective prediction.

If we were to achieve satisfactory tweet classification the way to go ahead would be to build event models from tweets representing current high-profile events and train classifiers on these models to match them up with events retrospectively. Each event model would have a set of features and a sentiment valence attached to it. The learner would then use the features of a model to identify events of a similar nature in our historic data and go on to associate sentiment or public reaction on a larger scale to these historic events. It would have an application in tasks such as the GDELT corpus[2] which attempts to record all of humanity's history in one place denoting time, locality and tone information. Such aa system would contribute to a great degree due to its ability to classify events together irrespective of spacial or temporal setting and perhaps offer further insight to the degrees of public reaction associated with event.

While quite interesting as well as beneficial to digital humanities, the task as of now has certain restrictive barriers. Firstly, as we have already stated is the poor performance of short-text sentiment analysis on the Twitter data task. The second is a well organized and annotated data-set for training. As we have observed, most of the available training data is unfit or rather unbalanced in terms of classes. Further, as of now we have only considered training on a single event i.e. sentiment scores associated with tweets on Global Warming (highly negative issue however, we have very few quality negative tweets, our tweet set seems to favor positive sentiment). A good way to go ahead would be to dedicate

more time and effort to collection of data. The idea is to mine tweets related to major events in modern context and build an event (set of statistical features) to sentiment mapping. This will be our training set, for a supervised learning task. For testing the task we would build a set of about 100 events which have well-documented public reactions in the GDELT data-set. After using our training set to measure the predicted public sentiment we would validate this against the historically recorded tone from the data-set. The limitation of the data-set is that it only contains global event data stretching back to 1979. However, it makes for a good initial setup to be able to predict sentiment using tweets for events belonging to the pre-Social Media era.

Despite the difficulties listed, the scenario is not quite that dark. A significant amount of development now surrounds this domain of study. Sentiment analysis from Twitter as a journalistic tool has received a spotlight recently in technological growth. This could be partly due to big News corporations getting enlightened to the productivity of this tool in Digital practice: media giants Bloomberg and the New York Times have already started their own wings of research in Big Data and its impact on Journalism. Perhaps more apt to our scenario is Thomson-Reuters move to tapping into Twitter for Big Data Sentiment Analysis[17]. This totally offers creditability that retrospective analysis of sentiment has both a useful application as well as feasibility.

Finally, the Sem-Eval 2013 task 2[18] by Nakov, Kozareva and Ritter offers a great opportunity at another attempt at our task. The Sem-Eval 2013 data-set collects data from Twitter and assigns sentiment tags of `positive, negative or neutral` at phrase and message level for a variety of different entities, products and events. The data-set recently released in agreement with Twitter Data Policies consists of about 12K -20K tweets. Our immediate goal is to integrate this data into our very next iteration of the proposed system.

VI. Conclusion

The take-away from this project is firstly that Tweet-based classifiers have as much aptitude to the sentiment analysis task as do news-headlines based classifiers and the former though quite frequently implemented now in modern research has a lot to offer in the sphere of retrospective sentiment analysis; and is definitely a direction worth considering. Secondly, short text-sentiment analysis is difficult and there is no doubt that systems require a lot more tuning before they perform as well as their larger context counter-parts such as review sentiment analysis , customer feedback analysis or their likes. However, any advancement in these would be a boost to our system; likewise any head-way in retrospective sentiment analysis using tweets entails parallel advancement in short-text sentiment analysis. Our system does perform well on the data-sets for which scores have been marginal: indicating that better feature-spaces need to be explored and perhaps that is one avenue towards achieving growth in the short-text analysis task.

References

[1] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11,* 122-129.

[2] Leetaru, K., & Schrodt, P. A. (2013, April). GDELT: Global data on events, location, and tone, 19792012. In *Paper presented at the ISA Annual Convention* (Vol. 2, p. 4).

[3] Strapparava, C., & Mihalcea, R. (2007, June). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 70-74). Association for Computational Linguistics.

[4] S. Bird, E Loper, and E Klein. *Natural Language Processing with Python*. OReilly Media Inc., 2009. http://ntlk.org/

[5] Ding, T., Fang, V., & Zuo, D. Stock Market Prediction based on Time Series Data and Market Sentiment.

[6] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford,* 1-12.

[7] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79-86, 2002.

[8] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65). ACM.

[9] J. Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics.*

[10] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1-135, 2008.

[11] Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics, 35(3)*, 399-433.

[12] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis.

[13] https://dev.twitter.com/docs/api/streaming

[14] http://www.crowdflower.com/open-data-library

[15] http://liwc.net/index.php

[16] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30-38). Association for Computational Linguistics.

[17] http://techcrunch.com/2014/02/03/twitter-raises-its-enterprise-cred-with-thomson-reuters-sentiment-analysis-deal/

[18] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter.