Final Report Proposal

Name: Guangyu Lin & Huihuang Zheng & Ge Gao
Proposed Topic: Sentiment Analysis in Twitter: Tweet classification according to a two-point scale with Named Entity Recognizer
Instructor: Ray Mooney
Course Name: CS 388 Natural Language Processing

Introduction: Tweets and texts are often used to share opinions and sentiments that people have about what is going on in the world around them. We believe that a freely available, annotated corpus that can be used as a common testbed is needed in order to promote research that will lead to a better understanding of how sentiment is conveyed in tweets and texts. Our primary goal in this task is to create such a resource: a corpus of tweets marked with their message-level polarity, in general and towards a specific topic, with adding Named Entity Recogniser as pre-process.

Problem Description: Given a tweet known to be about a given topic, classify whether the tweet conveys a positive or a negative sentiment towards the topic, which also required to filter out tweets that were not about the topic. We train and test with or without adding Named Entity Recogniser as pre-process.

Evaluation:As such, it is thus a binary classification task, in which each tweet must be classified as belonging to exactly one of the two classes C={Positive, Negative}, which also required to filter out tweets that were not about the topic. As an evaluation measure, for this task we will adopt macroaveraged recall, i.e.,

$$\rho^{PN} = \frac{\rho^{Pos} + \rho^{Neg}}{2} \tag{1}$$

ρ^{Pos} defines as the fraction of Positive tweets that are predicted to be such; in terms of the confusion matrix of Table, this means that $\rho^{Pos} = \frac{PP}{PP+NP}$. ρ^{PN} ranges in [0,1], where 1 is achieved only by the perfect classifier (the classifier that correctly classifies all items), 0 is achieved only by the perverse classifier (the classifier that misclassifies all items), while 0.5 is

- the value obtained by a trivial classifier (i.e., the classifier that assigns all tweets to the same class - be it Positive or Negative), and

- the expected value of a random classifier.

The advantage of ρ^{PN} over 'standard' accuracy is that it is more robust to class imbalance, since for standard accuracy the score of the majority-class classifier is the relative frequency (aka 'prevalence') of the majority class, that may be much higher than 0.5 if the test set is imbalanced. The advantage of ρ^{PN} over F1 is that it is more robust to class imbalance, since for F1 the score of the trivial acceptor may be much higher than 0.5 if the test set is

imbalanced and the Positive class is the majority class. Another advantage of ρ^{PN} over F1 is that ρ^{PN} is invariant with respect to switching Positive with Negative, while F1 is not.

		Actual	
		Pos	Neg
Predicted	Pos	PP	PN
	Neg	NP	NN

References

[1] Bella, A., Ferri, C., Hernandez-Orallo, J., & Ramirez-Quintana, M. J. (2010). Quantification via probability estimators. In Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010), pp. 737-742, Sydney, AU.

[2] Esuli, A., & Sebastiani, F. (2010b). Sentiment quantification. IEEE Intelligent Systems, 25(4), 72-75.

[3] Wei Gao and Fabrizio Sebastiani. Tweet Sentiment: From Classification to Quantification. Proceedings of the 6th ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015), Paris, FR, 2015.