End-to-End Sentiment Analysis of Twitter Data

Apoorv Agarwal[1] *Jasneet Singh Sabharwal*[2]
(1) Columbia University, NY, U.S.A.
(2) Guru Gobind Singh Indraprastha University, New Delhi, India
apoorv@cs.columbia.edu, jasneet.sabharwal@gmail.com

Abstract
In this paper, we present an end-to-end pipeline for sentiment analysis of a popular micro-blogging website called Twitter. We acknowledge that much of current research adheres to parts of this pipeline. However, to the best of our knowledge, there is no work that explores the classifier design issues explored in this paper. We build a hierarchal cascaded pipeline of three models to label a tweet as one of Objective, Neutral, Positive, Negative class. We compare the performance of this hierarchal pipeline with that of a 4-way classification scheme. In addition, we explore the trade-off between making a prediction on lesser number of tweets versus F1-measure. Overall we show that a cascaded design is better than a 4-way classifier design.

Keywords: Sentiment analysis, Twitter, cascaded model design, classifier confidence.

1 Introduction

Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on microblogs.[1] One challenge is to build technology to detect and summarize an overall sentiment.

In this paper, we look at one such popular micro-blog called Twitter[2] and propose an end-to-end pipeline for classifying tweets into one of four categories: *Objective, Neutral, Positive, Negative*. Traditionally, *Objective* category is defined as text segments containing facts and devoid of opinion (Pang and Lee, 2004; Wilson et al., 2005). In the context of micro-blogs, we extend this definition to include intelligible text, like "SAPSPKSAPKOASKOP SECAFLOZ PSOKASPKOA". Note, since we are only concerned with sentiment analysis of English language micro-blogs, text in other languages will also fall under the intelligible category and thus under Objective text.

One option to classify tweets into one of the four aforementioned categories is to simply implement a 4-way classifier. Another option is to build a cascaded design, stacking 3 classifiers on top of each other: Objective versus Subjective, Polar versus Non-polar and Positive versus Negative. In this paper, we explore these possibilities of building a classifier. In addition, we study the trade-off between making predictions on lesser number of examples versus F1-measure. If the confidence of the classifier falls below a threshold, we reserve prediction on that example. In expectation, this will boost the F1-measure, because we are reserving prediction on *harder* examples. But a-priori the

[1] http://mashable.com/2010/04/19/sentiment-analysis/
[2] www.twitter.com

Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, pages 39–44,
COLING 2012, Mumbai, December 2012.

relation between the two (threshold and F1-measure) is unclear. Moreover, it is unclear in which of the aforementioned three designs, this trade-off is least. We present this relation graphically and show that one of the cascaded designs is significantly better than the other designs.

We use manually annotated Twitter data for our experiments. Part of the data, Positive, Negative and Neutral labeled tweets were introduced and made publicly available in (Agarwal et al., 2011). Annotations for tweets belonging to the *Objective* category are made publicly available through this work.[3]

In this paper, we do not introduce a new feature space or explore new machine learning paradigms for classification. For feature design and exploration of the best machine learning model, we use our previous work (Agarwal et al., 2011).

2 Literature Survey

In most of the traditional literature on sentiment analysis, researchers have addressed the binary task of separating text into Positive and Negative categories (Turney, 2002; Hu and Liu, 2004; Kim and Hovy, 2004). However, there is early work on building classifiers for first detecting if a text is Subjective or Objective followed by separating Subjective text into Positive and Negative classes (Pang and Lee, 2004). The definition of Subjective class for Pang and Lee (2004) contains only Positive and Negative classes, in contrast to more recent work of Wilson et al. (2005), who additionally consider Neutral class to be part of Subjective class. Yu and Hatzivassiloglou (2003) build classifiers for the binary task Subjective versus Objective and the ternary task Neutral, Positive and Negative. However, they do not explore the 4-way design or the cascaded design. One of the earliest work to explore these design issues is by Wilson et al. (2005). They compare a 3-way classifier that separates news snippets into one of three categories: Neutral, Positive and Negative, to a cascaded design of two classifiers: Polar versus Non-polar and Positive versus Negative. They defined Polar to contain both Positive and Negative class and Non-polar to contain only Neutral class. We extend on their work to compare a 4-way classifier to a cascaded design of three models: Objective versus Subjective, Polar versus Non-polar and Positive versus Negative. Note, this extension poses a question about training the Polar versus Non-polar model: should Non-polar category only contain Neutral examples or both Neutral and Objective. Of course, the 4-way classifier puts all three categories (Objective, Positive and Negative) together while training a model to detect Neutral. In this paper, we explore these designs.

In the context of micro-blogs such as Twitter, to the best of our knowledge, we know of no literature that explores this issue. Barbosa and Feng (2010) build two separate classifiers, one for Subjective versus Objective classes and one for Positive versus Negative classes. They present separate evaluation on both models but do not explore combining them or comparing it with a 3-way classification scheme. More recently, (Jiang et al., 2011) present results on building a 3-way classifier for Objective, Positive and Negative tweets. However, they do not explore the cascaded design and do not detect Neutral tweets. Moreover, to the best of our knowledge, there is no work in the literature that studies the trade-off between making less predictions and F1-measure. Like human annotations, predictions made by machines have confidence levels. In this paper, we compare the 3 classifier designs in terms of their ability to predict better given a chance to make predictions only on examples they are most confident on.

[3]Due to Twitter's recent policy, we might only be able to provide the tweet identifiers and their annotation publicly available: http://www.readwriteweb.com/archives/how_recent_changes_to_twitters_terms_of_service_mi.php

3 End-to-end pipeline with cascaded Models

The pipeline for end-to-end classification of tweets into one of four categories is simple: 1) crawl the tweets from the web, 2) pre-process and normalize the tweets, 3) extract features and finally 4) build classifiers that classify the tweets into one of four categories: Objective, Neutral, Positive, Negative.

We use our previous work for pre-processing, feature extraction and selection of suitable classifier (Agarwal et al., 2011). We found Support Vector Machines (SVMs) to perform the best and therefore all our models in this paper are supervised SVM models using *Senti-features* from our previous work.

The main contribution of this work is the exploration of classifier designs. Following is a list of possible classifier designs:

1. Build a **4-way** classifier. Note, in a 4-way classification scheme, a multi-class one-versus-all SVM builds 4 models, one for identifying each class. Each model is built by treating one class as positive and the remaining three classes as negative. Given an unseen example, the classifier passes this through the four models and predicts the class with highest confidence (as given by the four models).

2. Build a hierarchy of 3 cascaded models: Objective versus Subjective, Polar versus Non-Polar and Positive versus Negative. But there is one design decision to be taken here: while building the Polar versus Non-polar model, do we want to treat both Neutral and Objective examples as Non-polar or only Neutral examples as Non-polar? This decision affects the way we create the Polar versus Non-polar model. Note, a 4-way model, implicitly treats Neutral to be Non-polar and the remaining three classes to be Polar. This scenario is unsatisfying because a-priori there is no reason why Objective examples should be treated as Polar at the time of training. We explore both these options:

   (a) **PNP-neutral**: Polar versus Non-polar model, where only Neutral examples are treated as Non-polar whereas Positive and Negative examples combined are treated as Polar.

   (b) **PNP-objective-neutral**: Polar versus Non-polar model, where Neutral and Objective examples combined are treated as Non-polar whereas Positive and Negative examples combined are treated as Polar.

In this paper, we present results for each of the aforementioned design decisions in training models. Moreover, we explore the trade-off between predicting on fewer number of examples and its affect on the F1-measure. It is not hard to imagine, especially when the output of the classifier is presented to humans for judgement, that we might want to reserve predictions on examples where the classifier confidence is low. A recent example scenario is that of Watson (Ferrucci et al., 2010) playing the popular gameshow Jeopardy! Watson buzzed in to answer a question only if it was confident over a certain threshold. We perform classification with **filtering**, i.e. considering classifier confidence along with its prediction. If the classifier confidence is below a certain threshold, we do not make a prediction on such examples. If for some value of threshold, θ, we reserve predictions on say x test examples, and say the total number of test examples is N, then the **Rejection rate** is given by $\frac{x}{N} * 100\%$.

Class	# instances for training	# instances for testing
Objective	1859	629
Neutral	1029	344
Positive	1042	350
Negative	1020	327

Table 1: Number of instances of each class used for training and testing

4 Experiments and Results

In this section, we present experiments and results for each of the pipelines described in section 3: 4-way, PNP-only-neutral and PNP-objective-neutral.

Experimental Setup: For all our experiments, we use support vector machines with linear classifier to create the models. We perform cross-validation to choose the right C value that determines the cost of mis-classifying an example at the time of learning. We report results on an unseen test set whose distribution is given in Table 1.

4.1 Classifier design

As explained in section 3, it is not clear a-priori, which of the three design decisions (4-way, PNP-neutral, PNP-objective-neutral) is most appropriate for building and end-to-end pipeline for sentiment analysis of Twitter data. Our results show that that PNP-objective-neutral gives a statistically significantly higher F1-measure for Neutral category, while giving same ball-park F1-measure for other three categories as compared to the other two design options.

	4-way			PNP-neutral			PNP-objective-neutral		
Category	P	R	F1	P	R	F1	P	R	F1
Objective	0.70	0.87	**0.78**	0.77	0.76	0.76	0.78	0.76	0.77
Neutral	0.51	0.30	0.38	0.48	0.22	0.31	0.39	0.46	**0.42**
Negative	0.56	0.56	0.56	0.49	0.64	0.56	0.57	0.57	**0.57**
Positive	0.59	0.56	0.57	0.51	0.67	**0.58**	0.61	0.53	0.57
Average	0.59	0.57	0.57	0.56	0.57	0.55	0.59	0.58	**0.58**

Table 2: Results for different classifier designs as mentioned in section 3. Note all numbers are rounded off to 2 significant digits.

Table 2 presents the result for the three design choices. For predicting the Objective class, all three designs perform in the same ball-park. For predicting the Neutral class, PNP-objective-neutral is significantly better than 4-way and PNP-neutral, achieving an F1-measure of 0.42 as compared to 0.38 and 0.31 respectively. For predicting the remaining two classes, Positive and Negative, the performance of the three designs is in the same ball-park.

4.2 Trade-off between Rejection rate versus F1-measure

Figure 1 presents a plot of rejection rate (on x-axis) versus mean F1-measure (on y-axis) for 4-way design (dotted green curve) and for PNP-objective-neutral (solid blue curve). The plot for the third design (PNP-neutral) is in the middle of these two curves and is omitted for clarity.

First thing to note is that the rejection rate always increases faster than in F1-measure.

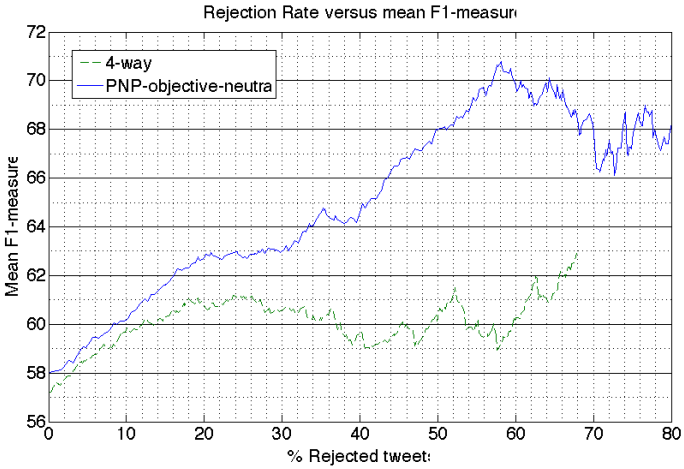

Figure 1: Rejection rate (on x-axis) versus mean F1-measure (on y-axis) for 4-way design (dotted green curve) and for PNP-objective-neutral (solid blue curve).

$$P = \frac{tp}{tp + fp}; R = \frac{tp}{tp + fn}; F1 = \frac{2PR}{P + R} = \frac{tp}{2tp + fp + fn}$$

where, P is precision, R is recall, $F1$ is F1-measure, tp is number of true positive, fn is number of false negative, and fp is number of false positive.

In the best case scenario, reserving predictions will lead to decrease in number of false positives and false negatives, without affecting true positives. So as x (number of test examples on which we reserve predictions) increases, rejection rate increases, and $fp + fn$ decreases (all linearly). Therefore, $F1 \propto \frac{1}{2+\frac{1}{x}} = \frac{x}{2x+1}$. It is easy to check that x grows faster than $F1$.

Second, the increase in the mean F-measure for PNP-objective-neutral grows at a higher rate as compared to the 4-way classifier. What this translates to is that the 4-way classifier is classifying true positives with lower confidence as compared to the cascaded model design, PNP-objective-neutral. Differently put, PNP-objective-neutral is eliminating more false positive and false negatives, which it is not confident about, as compared to 4-way. Comparing the maximum mean F-measure achieved by both designs, we see that 4-way achieves the mean maximum F-measure of 0.63 at a rejection rate of 67.81% as compare to PNP-objective-neutral, which achieves a higher maximum mean F-measure of 0.71 at a lower rejection rate of 58.12%.

Conclusion and Future Work

We conclude that overall PNP-objective-neutral is a better design. In the future we would like to study the nature of examples that the classifier deems *hard* and its correlation with what humans think is *hard*. For this we will need human confidence on their annotations.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon. Association for Computational Linguistics.

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaefer, N., and Welty, C. A. (2010). Building watson: An overview of the deepqa project. *AI Magazine*, 31:59–79.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *KDD*.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. *49th Annual Meeting of Association of Computational Linguistics*, pages 151–160.

Kim, S. M. and Hovy, E. (2004). Determining the sentiment of opinions. *Coling*.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. *ACL*.

Turney, P. (2002). Thumbs up or thumbs down? seman- tic orientation applied to unsupervised classification of reviews. *ACL*.

Whissel, C. M. (1989). *The dictionary of Affect in Language*. Emotion: theory research and experience, Acad press London.

Wilson, T., Wiebe, J., and Hoffman, P. (2005). Recognizing contextual polarity in phrase level sentiment analysis. *ACL*.

Wu, Y., Zhang, Q., Huang, X., and Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541, Singapore. Association for Computational Linguistics.

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Conference on Empirical methods in natural language processing*, 10:129–136.