Sentiment Analysis of Microblog Data Streams

...

Steven Kester Yuwono (A0080415N)
Choo Jia Le (A0116673A)

List of features

No.	Features	Descriptions
1	**Text**	**The content of the tweet (with/without POS tag)**
2	Contains Topic	1 if the tweet contains the topic, 0 otherwise
3	Positive Lexicon count	The number of positive lexicon present in the text of the tweet
4	Negative Lexicon count	The number of negative lexicon present in the text of the tweet
5	Retweets	The number of retweet for this particular tweet
6	Timezone	The timezone of the user's location when this tweet is posted
7	Friends Count	The number of users this account is following (AKA their "followings")
8	Followers Count	The number of followers this account currently has.
9	Favourites Count	The number of tweets this user has favorited in the account's lifetime.
10	User ID	ID of the person posting this tweet

Methodology (1/2)

- Preprocessing of tweets
  a. Replace all newline and tab characters with space.
  b. Use regular expression matching to generalize all URL mentioned in the tweet. (In other words, replaces http://t.co/someID urls with a standardize “[url]” string.)
  c. Use Stanford Tokenizer to tokenize the text.
  d. Convert all letters to lowercase.

Methodology (2/2)

- Explored 3 machine-learning algorithms
  - Naive Bayes (Stanford CoreNLP)
  - Maximum Entropy (Stanford CoreNLP)
  - SVM (SVM-Light, multiclass)
- Limitation for SVM: only accept real numbers features
- Resolved by using TF/IDF of each word as a feature
- **Maxent** is the best classifier out of the three.

Advantage of Maxent

- Obtain weights of all features (including each word in the text)
  - e.g. [“good”,positive] = 0.3 ; [“slow”, negative] = 0.2)
- Get top 1000 features and print a list of possible sentiment words (positive or negative)
- Used to improve the system’s performance further

Negative	Positive
angst	#awesome
#palmface	#prime
#ceo	power
#fail	king
malfunctioningagain	biggest
wait	innovations
#notcool	lmfao
#neednewipadguide	singing
#ripstevejobs	telling
#thenonsensepersists	pulling
#fatfuckingchance	details
n't	simple
wonky	upgrade
#fuckingpissed	replaced
fucks	#genius
wont	#honest
whhyy	gratis
eclipsed	
still	
humpt	
pissing	
#crankywithnophone	

Testing

- To develop and tune our system, we used the development set given, with 10-fold cross validation.
- For testing, we used both training and test set combined together with 10-fold cross validation
- Using F1-Score to assess the performance of the system

Results

Features	Naive Bayes	Maxent	SVM
Text only	**0.4013**	**0.7892**	**0.2605**
Text + POS tag	**0.3804**	**0.7758**	**-**

The performance (F1 score of all the classes and 10-fold cross validation) of the 3 classifiers

Results

Features	Maxent Unigram	Maxent Bigram
Text only (1)	0.7892	0.7964
Text + Topic (1+2)	0.7894	0.7964
Text + Topic + PosLex + NegLex (1+2+3+4)	0.7951	0.8013
Text + RetweetCount + Timezone (1+5+6)	0.7892	0.7986
Text + UserID (1+10)	0.7941	0.7961
Text + Topic + PosLex + NegLex + UserID (1+2+3+4+10)	**0.7964**	**0.8033**
Text + FriendCount (1+7)	0.4932	0.5801
Text + FollowersCount (1+8)	0.4621	0.4636
Text + FavouritesCount (1+9)	0.4793	0.4857

Results

Maxent N-Gram (N)	**Best features** Text + Topic + PosLex + NegLex + UserID (1+2+3+4+10)
1-gram (unigram)	0.7964
2-gram (bigram)	**0.8033**
3-gram (trigram)	0.8026
4-gram	0.8003

Maxent classifier performance with best features and N-gram

Results

- Manually looked through all the possible additional sentiment terms for both positive and negative lexicon, learnt from the development set

Maxent N-Gram (N)	**Best features + learnt lexicon** Text + Topic + PosLex + NegLex + UserID (1+2+3+4+10)
1-gram (unigram)	0.7991
2-gram (bigram)	**0.8061**
3-gram (trigram)	0.8041
4-gram	0.8028

Conclusion

- Only topic, positive lexicon count, negative lexicon count, and user ID are useful
- User ID is useful because some users tend to post either negative tweets or positive tweets quite consistently.
- Friends' count, followers' count, and favourites' count is the least useful features, as shown by the significant drop in the classifier's performance when these features are used.
- POS tag is found to be not very useful
- The best classifier is the bigram classifier, showing the best result as compared to unigram, trigram and even 4-gram