Twitter Sentiment Analysis

http://www.cs.columbia.edu/~julia/papers/Agarwaleta l11.pdf

Problem Definition

Microblogging websites such as Twitter have evolved to become a source of varied kind of information

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

Our challenge is to build technology to detect and summarize an overall sentiment

Previous Work

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity.

Starting from being a document level classification task (Turney, 2002; Pang and Lee, 2004), it has been handled at the sentence level (Hu and Liu, 2004; Kim and Hovy, 2004) and more recently at the phrase level (Wilson et al., 2005; Agarwal et al., 2009)

Another significant effort for sentiment classification on Twitter data is by Barbosa and Feng (2010). They use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for tuning and another 1000 manually labeled tweets for testing.

Our System

The traditional model for sentiment Analysis is unigram or Bag of Words model

In addition to the unigram modelo mining sentiments we introduce certain more features depending upon the polarity of the words. These features can be divided into 11 categories

f1,f2,f3,f4,f5,f6,f7,f8,f9,f10,f11

The features

$\mathbb{N}$	Polar	POS	# of (+/-) POS (JJ, RB, VB, NN)	f_1
		Other	# of negation words, positive words, negative words	f_2
			# of extremely-pos., extremely-neg., positive, negative emoticons	f_3
			# of (+/-) hashtags, capitalized words, exclamation words	f_4
	Non-Polar	POS	# of JJ, RB, VB, NN	f_5
		Other	# of slangs, latin alphabets, dictionary words, words	f_6
			# of hashtags, URLs, targets, newlines	f_7
$\mathbb{R}$	Polar	POS	For POS JJ, RB, VB, NN, $\sum$ prior pol. scores of words of that POS	f_8
		Other	$\sum$ prior polarity scores of all words	f_9
	Non-Polar	Other	percentage of capitalized text	f_{10}
$\mathbb{B}$	Non-Polar	Other	exclamation, capitalized text	f_{11}

Data

I use data from one kaggle Problem(Partly Sunny with plenty of hastags)

https://www.kaggle.com/c/crowdflower-weather-twitter/data

Training Data contains 77946 lines of tweet with the associated results

s2,s3,s4

Result s2 : Negative s3: Neutral, s4: Positive

Data Preprocessing And Representation

we designed two new resources for pre-processing twitter data:

1) an emoticon dictionary and

2) an acronym dictionary

Emoticons Dictionary

An emoticon s a pictorial representation of a facial expression using punctuation marks, numbers and letters, usually written to express a person's feelings or mood.

Emoticons can generally be divided into three groups: Western or horizontal (mainly from America and Europe), Eastern or vertical (mainly from east Asia), and 2channel style (originally used on 2channel and other Japanese message boards)

Emoticons Dictionary

We mined Wikipedia and AFINN for building the emoticons dictionary

THough AFINN provides a list of emoticons it was too limited for the given application

The Augmentation from Wikipedia was done manually as we had to figure out the happiness score from the emojis

:-I :I					😐😑	Straight face[5] no expression, indecision[8]
:$					😳😞😖	Embarrassed,[6] blushing[7]
:-X :X	:-# :#	:-& :&			🤐😶	Sealed lips or wearing braces,[4] tongue-tied[8]
O:-) O:)	0:-3 0:3	0:-) 0:)	0;^)		😇🙏	Angel,[4][5][9] saint,[8] innocent
>:-) >:)	}:-) }:)	3:-) 3:)	>;)		😈	Evil,[5] devilish[8]
l;-)	l-O				😎🥱	Cool,[8] bored/yawning[9]
:-J					😏😒	Tongue-in-cheek[11]
#-)					—	Partied all night[8]
%-) %)					😵😕🤢	Drunk,[8] confused
:-###.. :###..					🤒🤮◻	Being sick[8]
<:-I					—	Dumb, dunce-like[9]

Abbreviations and slang Dictionary

We crawled noslang.com to build a dictionary of abbrevaitions and slangs like lol,rofl

The dictionary has over 5000 words and serves well for the purpose

Tree Representation for Tree Kernel

This was motivated by http://disi.unitn.it/moschitti/Tree-Kernel.htm and the tree representation in the tree_rep.txt mimics the Penn TreeBank Format

Basic Rules of Tree generation:

a) if the token is a target, emoticon, exclamation mark, other punctuation mark, or a negation word, add a leaf node to the “ROOT” with the corresponding tag

b) if the token is a stop word, we simply add the subtree “ (STOP (‘stop-word’))” to “ROOT”.
c) if the token is an English language word, we map the word to its part-of-speech tag, calculate the prior polarity of the word and add the subtree (EW (‘POS’ ‘word’ ‘prior polarity’)) to the “ROOT”.

d) For any other token we add subtree “(NE ())” to the “ROOT”. “NE” refers to non-English

e) For any token which has repeated characters like cool we add a tag EMP (emphasis) to the root

Polarity of Words

A number of our features are based on prior polarity of words.

For this we use SentiWordnet and extend it using WordNet.

SentiWordNet is a lexical resource for opinion mining.

It assigns to each synset three sentiment scores: positivity, negativity, objectivity.

Word Polarity

If a word is not directly found in the sentiwordnet we retrieve all synonyms from Wordnet.

We then look for each of the synonyms in sentiwordnet.

If any synonym is found in sentiwordnet, we assign the original word the same pleasantness score as its synonym.

If none of the synonyms is present in sentiwordnet, the word is not associated with any prior polarity.

Results

Evaluation data, metric and experimental Results

Model	Accuracy
Unigram	54.34
Unigram +features described(f1-f11)	59.23
Unigram+features described(f1-f11) Tree Kernel	60.43

Conclusion

We investigated two kinds of models: tree kernel and feature based models and demonstrate that both these models outperform the unigram baseline.

For our feature-based approach, we do feature analysis which reveals that the most important features are those that combine the prior polarity of words and their parts-of-speech tags.

We tentatively conclude that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres