National University of computer and emerging sciences

Sentimental Analysis

CS401

PROBLEM

Someone tweets: @VirginAmerica Hey first time flyer next week - excited! But I'm having a hard time getting my flights added to my Elevate account. Help?

- Is he/she happy or unhappy with the flight experience?
- If he/she is not happy with the, what went wrong?

My project let machine solve the task to identify sentiment of each tweet. Which means that i with this effort i will be able to specify weather the tweet done by the user was in the favour or against the comapany

Dataset

The dataset used is twitter airline data for Virgin American airline. . This dataset has almost 15000 rows along with time name sentiment value and remarks on each tweet. The data fraction is roughly 15% positive, 65% negative, and 20% neutral. I split the data into 67% as the training set and 33% for testing, and I used 5-fold cross-validation to choose hyper-parameters. I also assume that each tweet has only one label negative reason label.

MODELS

SVM

SVM is based on finding the maximum margin between different classes by determining w and b in w . x + b = 0 , Among many types i have used Linear Support Vector Machine

NAIVE BAYES

This method is based on probabilistic inference rules. Given x = (x1, x2 ,...xn) be some n features of some test sample. If the class is given then all features are independent of each other. For representation i used N-grams

VADER SENTIMENT LEXICON

The main assumption behind lexicon based methods is that the sentiment of a text is determined by any number of dominant words in that text. The dominance of a word in a text is calculated by applying a simple word counting technique. VADER is a lexicon with both polarity and intensity information attached to each entry. The basic structure is shown in the following table.The intensity of each word is calculated by averaging human evaluation vector gathered from ten experts' annotation. The lexicon only contains entries with a standard deviation less than 2.5.

RESULTS

NAIVE BAYES

Category	Precision	Recall	F1-measure
Positive	0.916	0.328	0.482
Negative	0.697	0.988	0.818
Neutral	0.718	0.246	0.367

LINEAR SVM

Recall	Precision	F1-measure
0.650	0.475	0.549
0.800	0.744	0.769
0.222	0.435	0.294

VADER LEXICON

Precision	Recall	F1-measure
0.812	0.678	0.739
0.831	0.918	0.873
0.667	0.559	0.608

CONCLUSION

Lexicon-based methods didn't perform well since it focuses on general cases and thus didn't bring any domain-specific knowledge. it only produced accuracy around 70%. On the other hand machine learning, algorithms performed well. Naive Bayes is faster then Linear support Vector Machine due to its simplicity. But Linear SVM yields better accuracy around 79%. whereas, Naive Bayes was only able to put on 76% accuracy on this dataset Linear SVM performs well since data is separated by support vectors word like good or bad

Muaz Maqbool 15-4053