SENTIMENT ANALYSIS

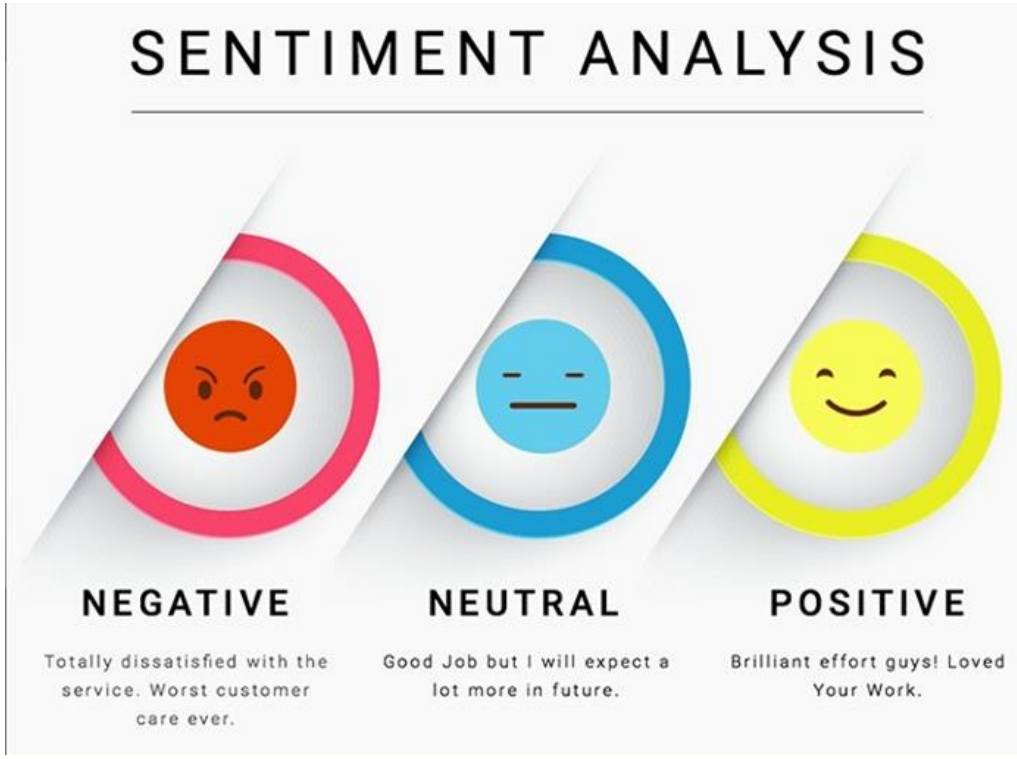
JUSTIN NAMBA, VIVEK GUPTA, ANMOL JAISING

MOTIVATION

- Sentimental analysis helps businesses understand what customers think of their products, services, buying experience.
- We test several algorithms to detect and evaluate customers' opinions.

SENTIMENT ANALYSIS

- Opinion mining
- Four types of sentiment analysis
  - Fined-grained
  - Emotion detection
  - Aspect-based
  - Intent analysis
- Two types of algorithms
  - Ruled-based
  - Automatic approach



DATASETS

- Hotel Dataset 1
  - 26 columns
  - 10,007 rows
- Hotel Dataset 2
  - 6 columns
  - 10,000 rows
- Amazon Dataset – Mobile phones
  - 6 columns
  - 400,000 rows

DATA PRE-PROCESSING

1. Select pertinent features
2. Remove rows with empty cells
3. Tokenization
4. Remove stop words
5. Stem words
6. Shuffle data
7. Get random sample

ALGORITHMS

- VADER
- Naive Bayes
- K-Nearest Neighbors

SOLUTION: VADER

- Valence Aware Dictionary and Sentiment Reasoner
  - Rule-based
  - Advantages
    - Polarity & intensity
    - Compound score

Sentiment Metric	Score
Positive	0.674
Neutral	0.326
Negative	0.0
Compound	0.735

Doc	Title Review	Negative	Neutral	Positive	Compound Score	Overall Polarity
1	measure.	0.0%	100.0%	0.0%	0.0%	Neutral
2	disappointed	100.0%	0.0%	0.0%	-47.7%	Negative
3	great hotel would stay	0.0%	42.3%	57.7%	62.5%	Positive
4	fantastic hotel ..... let staff	0.0%	52.6%	47.4%	55.7%	Positive
5	seattle excellence	0.0%	19.6%	80.4%	62.5%	Positive

Table 1: VADER: Title Reviews from Hotel 1

Doc	Text Review	Mode
1	first last stay property...	Negative
2	wife stayed hotel two nights, arriving tuesday...	Positive
3	family 5 travelling australia...	Neutral
4	visted san diego hotel solamar	
	first leg honeymoon...	Positive
5	hotel staff service great...	Positive

Table 2: VADER: Text Reviews from Hotel 1

RESULTS: VADER

Doc	Text Review	Mode
1	inn lake ok tourist area standards...	Positive
2	bad: came evening long business journey...	Positive
3	stayed hotel several times every time...	Negative
4	helpful staff check...	Positive
5	book prepaired!...	Neutral

Table 4: VADER: Text Reviews from Hotel 2

Doc	Title Review	Negative	Neutral	Positive	Compound Score	Overall Polarity
1	ok place short visit	0.0%	57.7%	42.3%	29.6%	Positive
2	staff accommodating problems	47.4%	52.6%	0.0%	-40.2%	Negative
3	terrible customer service	60.8%	39.2%	0.0%	-47.7%	Negative
4	lake delton resort	0.0%	100.0%	0.0%	0.0%	Neutral
5	150 day ridiculous was	55.6%	44.4%	0.0%	-36.1%	Negative

Table 3: VADER: Title Reviews from Hotel 2

RESULTS: VADER – HOTEL 2 DATASET

Doc	Text Review	Mode
1	first nokia phone owned love...	Neutral
2	exactly expecting!...	Neutral
3	phone 3 weeks mostly love aside annoying little things...	Positive
4	've bought factory unlocked device, received locked version...	Negative
5	liked phone...	Neutral

Table 5: VADER: Text Reviews from Amazon

RESULTS: VADER – AMAZON DATASET

SOLUTION: NAÏVE BAYES

- Supervised machine learning classification algorithm
- Uses bayes theorem which assumes features are independent
- Data to classifier is fed in form of reviews with assigned labels (positive, negative and neutral) based on ratings
- Ratings > 3.8 are labeled positive, ratings between 3.0 and 3.8 are labeled neutral and those with < 3.0 are labeled negative.
- Requires fewer data to train and identify features of model

Doc	Text Review	Mode
1	The food was delicious...	Positive
2	The rooms were not clean...	Negative
3	The hotel was expensive but...	Positive
4	The quantity of the food was not enough	Positive

Table 9: Naive Bayes on Hotel1 dataset

Doc	Text Review	Mode
1	The food was delicious...	Positive
2	The rooms were not clean...	Positive
3	The hotel was expensive but...	Positive
4	The quantity of the food was not enough	Neutral

Table 10: Naive Bayes on Hotel2 dataset

RESULTS: NAÏVE BAYES – HOTEL 1 & 2 DATASETS

Doc	Text Review	Mode
1	Very clean set up and easy set up.	Positive
2	The camera quality was not that great...	Positive
3	There was only one little blemish on the side, but who cares	Negative
4	I'm really disappointed about my phone and service	Negative

Table 11: Naive Bayes on Amazon dataset

RESULTS: NAÏVE BAYES – AMAZON DATASET

SOLUTION: K-NEAREST NEIGHBORS

- Supervised classification algorithm.
- Finds the k-closest training example in the datasets.
- Easy to implement. Uses hamming distance to compute distance
- Advantages:
  - No training period
  - New data can be added anytime
- Disadvantages:
  - Does not work well with large datasets
  - Requires feature scaling.

Doc	Title Review	Average Rating
1	It is also good to here	Very Positive
2	We hope to see you back..	Neutral
3	great hotel would stay	Very Positive
4	fantastic hotel .....let staff	Very Positive
5	Settle excellence	Very Positive
6	We would hesitate to ..	Very Negative

Table 7: K-Nearest Neighbours on Hotel 2

Doc	Title Review	Average Rating
1	All great around	Very Positive
2	We Miss You Already	Neutral
3	Wonderful visit	Very Positive
4	The Most Disgusting Motel!!!!!	Very Negative
5	This hotel was perfect for us	Very Positive

Table 6: K-Nearest Neighbours on Hotel 1

RESULTS: K-NEAREST NEIGHBORS – HOTEL 1 & 2 DATASETS

Doc	Text Review	Mode
1	Very clean set up and easy set up.	Positive
2	The camera quality was not that great...	Positive
3	There was only one little blemish on the side, but who cares	Negative
4	I'm really disappointed about my phone and service	Negative

Table 11: Naive Bayes on Amazon dataset

RESULTS: K-NEAREST NEIGHBORS - AMAZON DATASET

Lessons Learned

VADER
- Advantages
  - Gives more information (polarity, intensity, & compound score)
- Disadvantages
  - Only works on a single sentence at a time.

Naive Bayes
- Advantages
  - Requires fewer data to train and identify features of model
  - Gets trained faster compared to other models
- Disadvantages
  - Zero frequency: A label-feature combination not present in sample data, thereby giving posterior probability as 0.

K-Nearest-Neighbors
- Advantages:
  - No training period
  - New data can be added anytime
- Disadvantages:
  - Does not work well with large datasets
  - Requires feature scaling

CONCLUSION

- Performed sentiment analysis on hotel and amazon datasets
- Reviewed three algorithms: VADER(rule-based), Naive Bayes(supervised ML algorithm) and K-Nearest Neighbors (unsupervised ML algorithm)
- VADER determines polarity as well as intensity of a sentiment
- Naive Bayes is a probabilistic classifier with independence assumption between features. It is faster in training the models.
- K-NN is an effective learning algorithm with ease of implementation on dataset of small size. With easy classification, K-NN is selected to be the widely use classifier

FUTURE WORK

- Making use of rule-based approaches
  - Regular Languages
  - Regular Expressions
- Experiment with other machine learning algorithms
  - Random Forest
  - Decision Tree