





# Sentiment analysis for amazon review

# What is Sentiment analysis

Sentiment analysis interprets and classifies positive, neutral, and negative attitudes within text data. The applications of sentiment analysis include identifying customer satisfaction levels, understanding users' opinions or attitudes toward a topic, etc.

# What is Sentiment analysis

## Movie review:

-  • unbelievably disappointing
-  • Full of zany characters and richly applied satire, and some great plot twists
-  • this is the greatest screwball comedy ever filmed
-  • It was pathetic. The worst part about it was the boxing scenes.

## Product review:



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner  
\$89 online, \$100 nearby ★★★★★ 377 reviews  
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

### Reviews

Summary - Based on 377 reviews



### What people are saying

ease of use	<div><div></div><div></div><div></div><div></div><div></div></div>	"This was very easy to setup to four computers."
value	<div><div></div><div></div><div></div><div></div><div></div></div>	"Appreciate good quality at a fair price."
setup	<div><div></div><div></div><div></div><div></div><div></div></div>	"Overall pretty easy setup."
customer service	<div><div></div><div></div><div></div><div></div><div></div></div>	"I DO like honest tech support people."
size	<div><div></div><div></div><div></div><div></div><div></div></div>	"Pretty Paper weight."
mode	<div><div></div><div></div><div></div><div></div><div></div></div>	"Photos were fair on the high quality mode."
colors	<div><div></div><div></div><div></div><div></div><div></div></div>	"Full color prints came out with great quality."

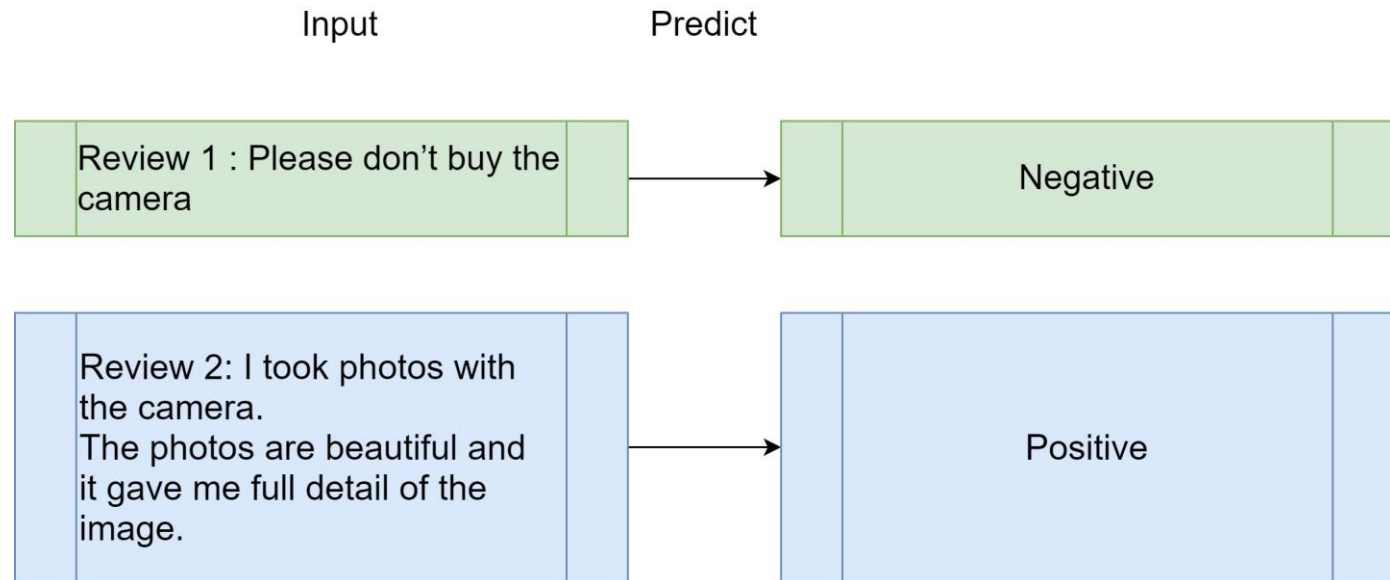
# What is Sentiment analysis

- **Movie:** is this review positive or negative?
- **Products:** what do people think about the new iPhone?
- **Public sentiment:** how is consumer confidence? Is despair increasing?
- **Politics:** what do people think about this candidate or issue?
- **Prediction:** predict election outcomes or market trends from sentiment

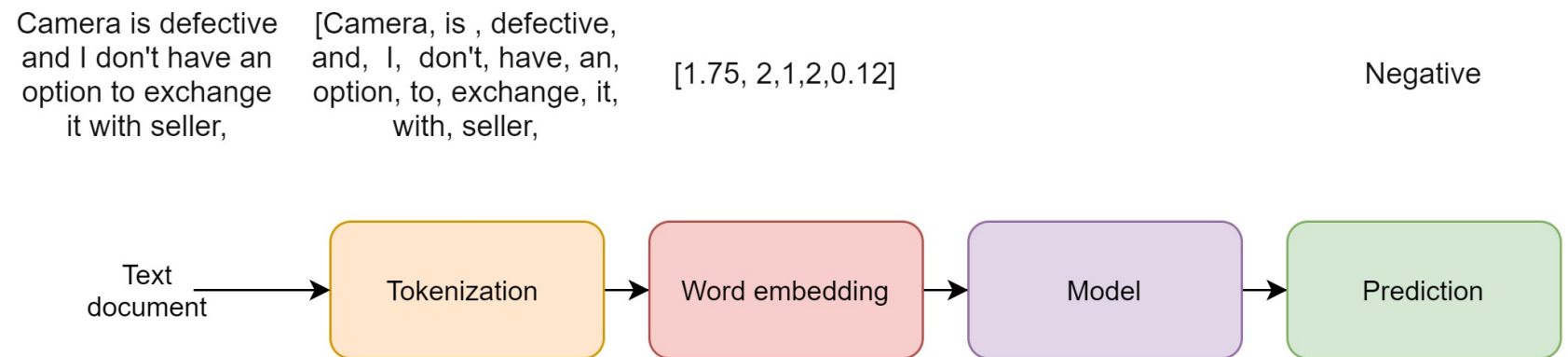
Our study:

Amazon  
review:

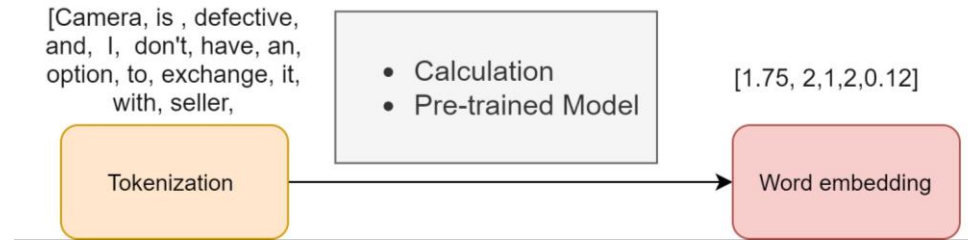
Is the review of  
this text positive  
or negative?



# Overview

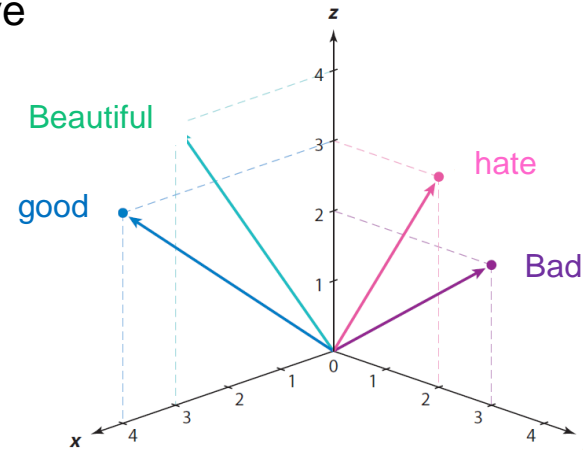


# How to get different word embedding??



Word Embeddings are the texts converted into numbers and there may be different numerical representations of the same text.

Aim to create a representation for words that capture their meanings, semantic relationships and the different types of contexts they are used in → Make the computer know whether this review context is positive or negative



How to get different word embedding??

Calculation from text in the review directly

### Simple bag of words

- Review 1 : Please don't buy the camera
- Review 2: I took photos with the camera.  
The photos are beautiful and it gave me full detail of the image.
- → build a vocabulary from all the unique words in the above reviews. 'Please', 'don't', 'buy', 'the', 'camera', 'I', 'took', 'photos', 'with', 'are', 'beautiful', 'it', 'gave', 'me', 'full', 'detail', 'of', 'image'.

	Please	don't	buy	the	camer a	I	took	photos	with	are	beauti ful	and	it	gave	me	full	detail	of	image
Review w1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Review w2	0	0	0	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1

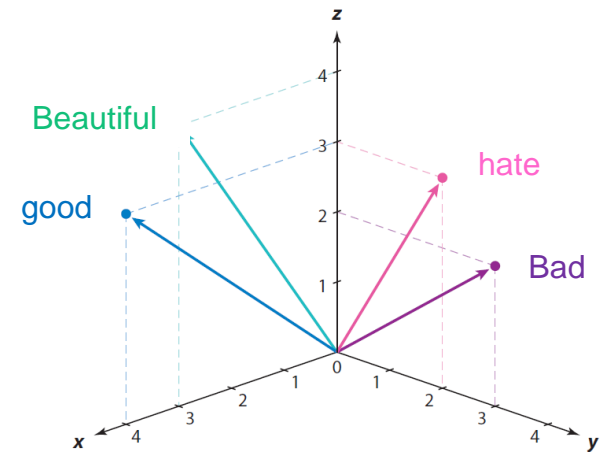
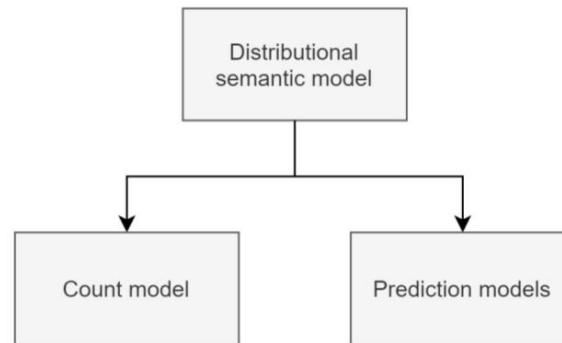
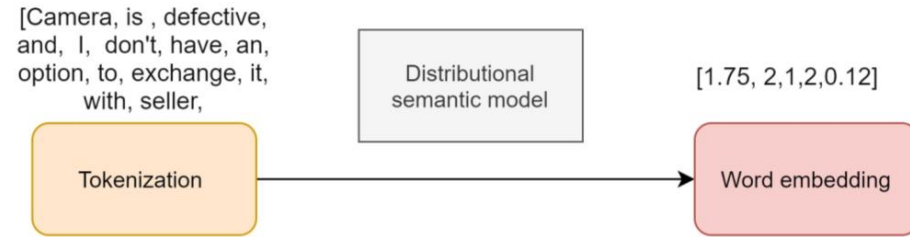
Each word is linked to a vector index and marked as 0 or 1 or more depending on how many times it occurs in document.

→ No semantic relationship and does not consider word sequence



How to get different word embedding??

Use co-occurent matrix from corpus (count model)



Beautiful is closer to good

Hate is closer to bad

→ Positive phrases co-occur more with 'good'  
Negative phrases co-occur more with 'bad'

How to get different word embedding??

Use co-occurent matrix from corpus (count model)

1. Get word to word co-occurrence matrix from the corpus (a corpus is a language resource consisting of a large and structured set of texts.)

	<i>bite</i>	<i>buy</i>	<i>drive</i>	<i>eat</i>	<i>get</i>	<i>live</i>	<i>park</i>	<i>ride</i>	<i>tell</i>
<i>bike</i>	0	9	0	0	12	0	8	6	0
<i>car</i>	0	13	8	0	15	0	5	0	0
<i>dog</i>	0	0	0	9	10	7	0	0	1
<i>lion</i>	6	0	0	1	8	3	0	0	0

2. Transformed the raw frequencies into significance weights to reflect the importance of the contexts (raw numbers are highly skewed) → Positive pointwise mutual information (PMI)

- Measures how much the probability of a target–context pair estimated in the training corpus is higher than the probability we should expect if the target and the context occurred independently of one another.

$$\text{PPMI}(t, c) = \max \left( 0, \log_2 \frac{p(t, c)}{p(t)p(c)} \right).$$

	<i>bite</i>	<i>buy</i>	<i>drive</i>	<i>eat</i>	<i>get</i>	<i>live</i>	<i>park</i>	<i>ride</i>	<i>tell</i>
<i>bike</i>	0	0.50	0	0	0	0	1.09	1.79	0
<i>car</i>	0	0.80	1.56	0	0	0	0.18	0	0
<i>dog</i>	0	0	0	2.01	0	1.65	0	0	2.16
<i>lion</i>	2.75	0	0	0	0.26	1.01	0	0	0

How to get different word embedding?? -

Use co-occurent matrix from corpus (count model)

With weighted least-square regression

GloVe learning

1. Get the word embedding from the matrix model
2. Training is performed on aggregated global word-word co-occurrence statistics from a corpus (a corpus is a language resource consisting of a large and structured set of texts.)

The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

Weighting term  
The, it, of : give less weights  
Rare words : higher weights

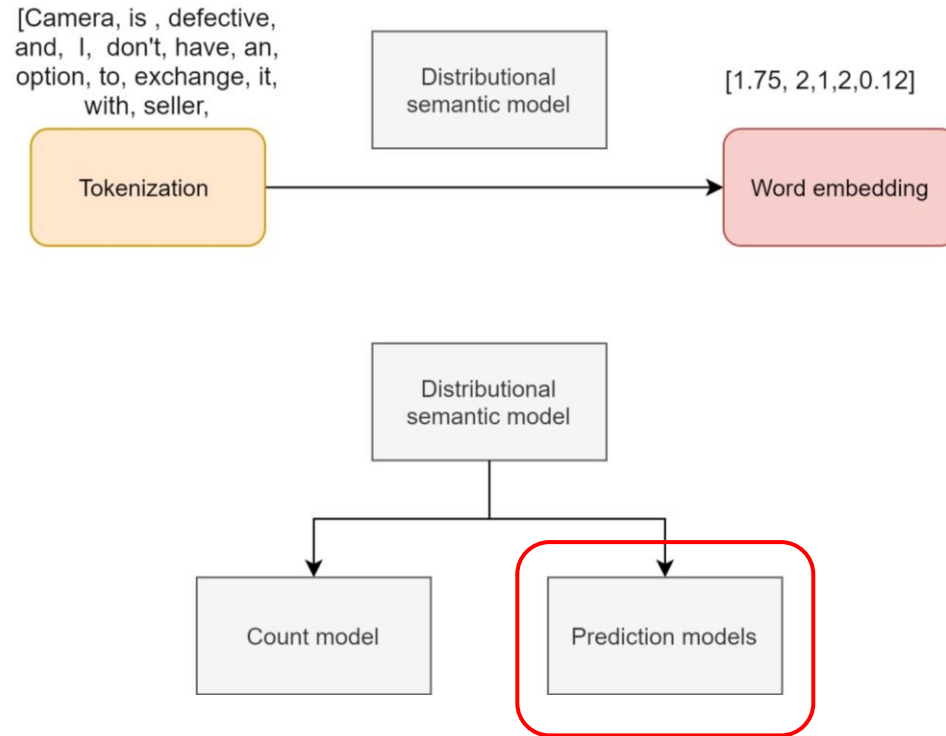
where  $w_i$  and  $b_i$  are the word vector and bias respectively of word  $i$ ,  $\tilde{w}_j$  and  $\tilde{b}_j$  are the context word vector and bias respectively of word  $j$ ,  $X_{ij}$  is the number of times word  $i$  occurs in the context of word  $j$ , and  $f$  is a weighting function that assigns relatively lower weight to rare and frequent co-occurrences.

Pennington et al. (2014)

Powerful feature  
representation:

BERT (Bidirectional  
Encoder  
Representation  
transformer)

- A powerful pre-trained deep learning model for word-embedding



Powerful feature  
representation:

BERT (Bidirectional  
Encoder  
Representation  
transformer)

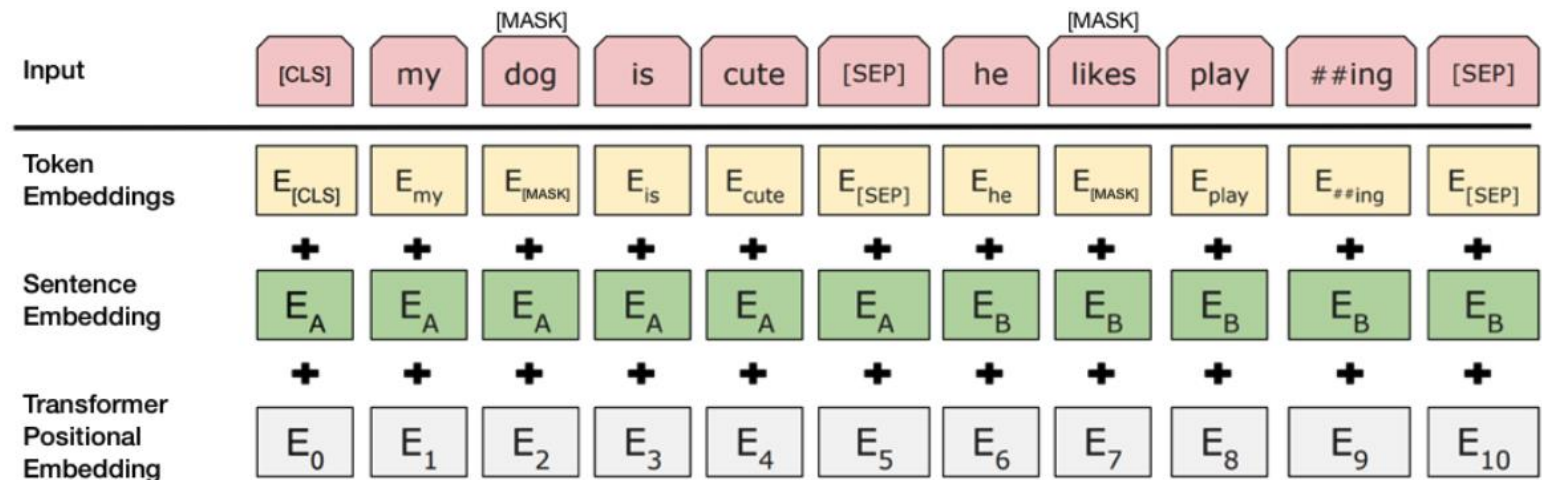
- Remember the position of the word

Positive: Actual sentence sequences

- [CLS] the man went to [MASK] store [SEP]
- he bought a gallon [MASK] milk [SEP]
- Label: IsNext

Negative: Randomly chosen second sentence

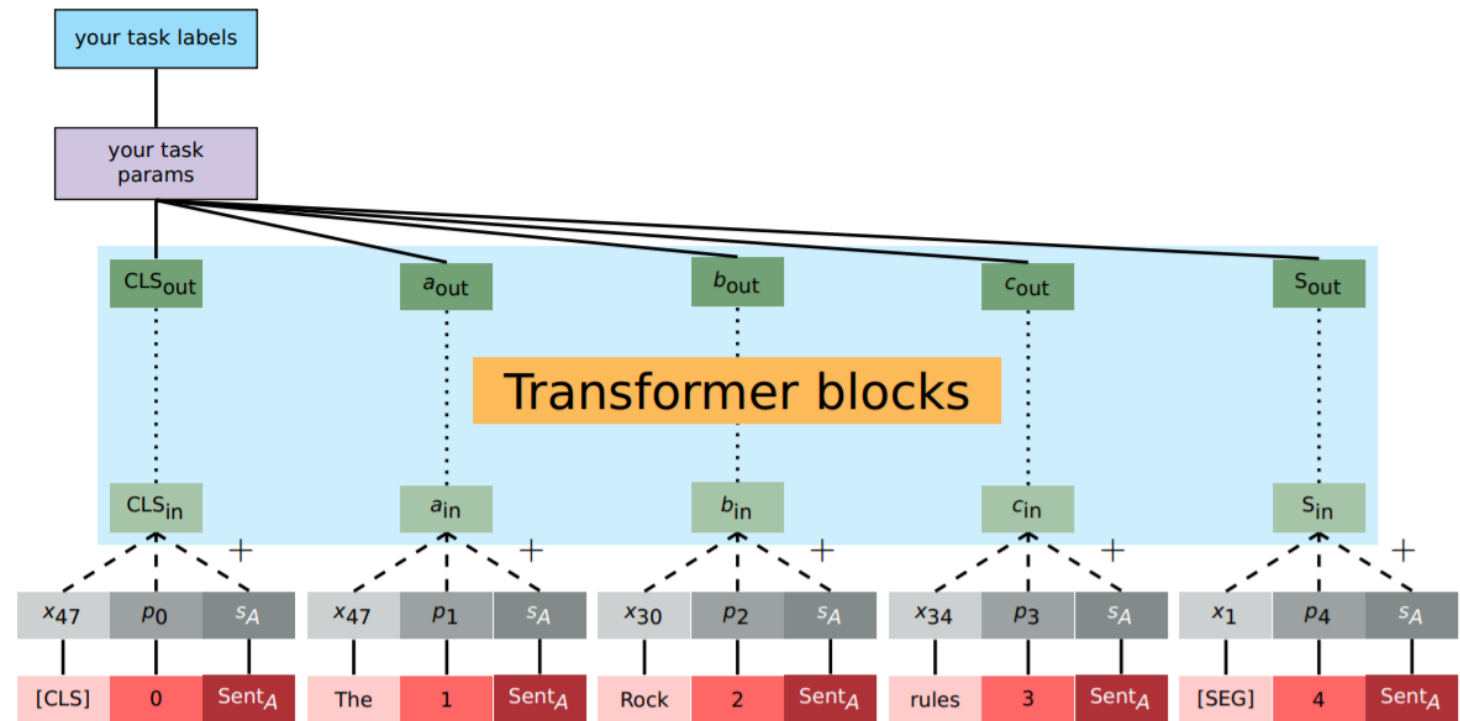
- [CLS] the man went to [MASK] store [SEP]
- penguin [MASK] are flight ##less birds [SEP]
- Label: NotNext



Sanh et al. (2019)

Powerful feature  
representation:

BERT (Bidirectional  
Encoder  
Representation  
transformer)



Sanh et al. (2019)

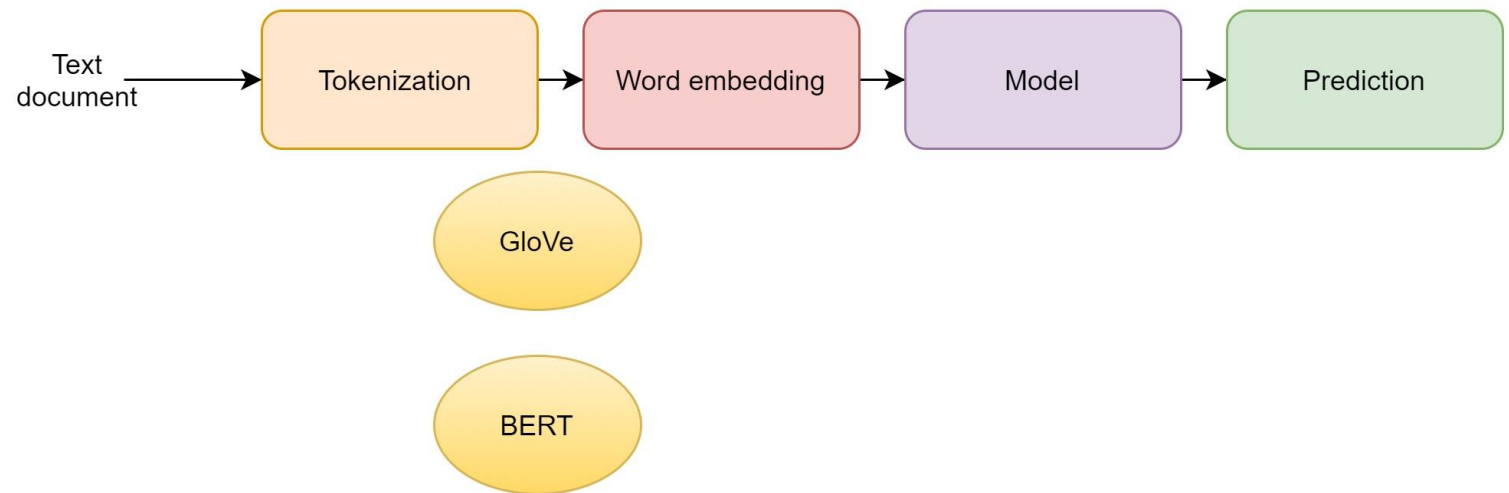
# Focus

Camera is defective  
and I don't have an  
option to exchange  
it with seller,

[Camera, is , defective,  
and, I, don't, have, an,  
option, to, exchange, it,  
with, seller,

[1.75, 2,1,2,0.12]

Negative



# Data

## Amazon camera review

market place	customer _id	review _id	product _id	product _parent	product _title	star _rating	helpful _votes	total _votes	vine	verified _purchase	review _headline	review _body	review _date
US	12491786	R1C0E VYBV 2Z18Z	B000U9 2DLA	8E+08	Kastar Camera & Camcorder Battery Home Travel ...	5	6	6	N	Y	A Must For On The Go	Purchased the charger from Nextop2100 , arrived...	1/22/2008
US	37280009	R3Q83 7V78N S6B9	B0008G CYNW	6E+08	Olympu s Camed a C755 4MP Digital Camera with 1...	3	15	17	N	Y	The Good and the Not so Good	I was attracted to this camera for three reaso...	9/8/2005
US	16532896	R22CZ CJH5Y NX1O	B003CZ 7L84	9E+08	Intova CP-9 Compac t Digital Camera with 130 fe...	4	1	1	N	Y	Great for what we paid!	We took it with us to Cancun. My husband took ...	11/2/2011

Review 1-3 : negative  
Review 4,5 : positive

Input

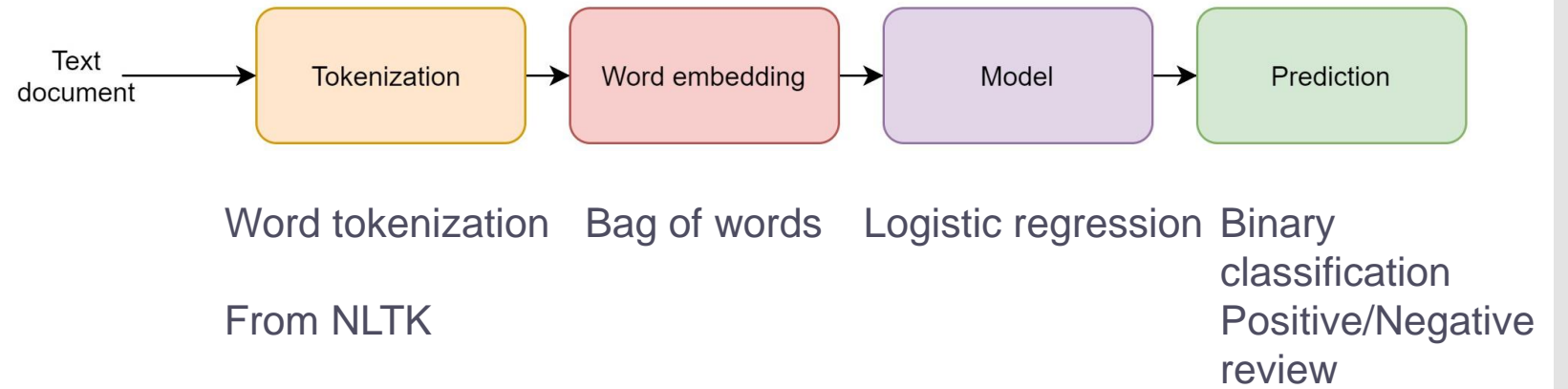
Imbalanced data : 13925 positives and 4083 negative reviews

Use only 10% of data

70% training , 30 % test set



# Baseline model



## Evaluation Metric

- Use Macro-average F1 to compare the effect of different word embedding method to the model
- Due to the imbalanced → using accuracy as metric is not ideal since it fails to control for size imbalances in the classes and just assume all the dataset with equal distribution.
- We used F1 score which combines precision and recall and give equal weight to precision and recall. Because the F score does not have normalization of the size of the data sets with K category. Hence, we also used **macro-average F1 score**, which is the mean of F1 score for each category (positive and negative reviews).

# Results

## Baseline Model

Data	Review	Accuracy	Precision	Recall	F_score	Macro_F_score
Training	Positive	0.849	0.875	0.452	0.875	0.860
Training	Negative	0.849	0.845	0.979	0.845	0.860
Test	Positive	0.685	0.500	0.06	0.500	0.596
Test	Negative	0.685	0.692	0.973	0.692	0.596

## GloVe

Data	Review	Accuracy	Precision	Recall	F_score	Macro_F_score
Training	Positive	0.952	1.0	0.806	1.0	0.970
Training	Negative	0.952	0.941	1.0	0.941	0.970
Test	Positive	0.704	0.571	0.235	0.571	0.647
Test	Negative	0.704	0.723	0.919	0.723	0.647

## BERT

Data	Review	Accuracy	Precision	Recall	F_score	Macro_F_score
Training	Positive	1.0	1.0	1.0	1.0	1.0
Training	Negative	1.0	1.0	1.0	1.0	1.0
Test	Positive	0.852	0.909	0.589	0.909	0.873
Test	Negative	0.852	0.837	0.973	0.837	0.873

Model Result		
	Training (Macro_F_score)	Test (Macro_F_score)
Baseline Model	0.860	0.596
GloVe	0.970	0.647
BERT	1.0	0.873

## Future Work

- We only used 10% of data and will try to run all the data in the AWS
- We can combine other amazon review data (cell phone, games,..etc) to see if the results are generalizable for all the amazon data
- Maybe combining some information of review title and helpful votes may help the model prediction
- Maybe use sampling method may help the model prediction due to the imbalanced datasets.
- The limitations for BERT is the sentence is limited to 512 tokens (some review context got truncated during the tokenization), there are new modified method - Roberta)