Online Patient Conversation Classifier

Indranil Chandra
indranildchandra@gmail.com
@IndranilChandra

Problem Statement

Build an Intelligent pipeline that can segregate patient conversations from the rest of the group given historically tagged patient data. You are expected to build an algorithm where they can ingest the social data and get the patient tags - 1 if patient and 0 if not a patient.

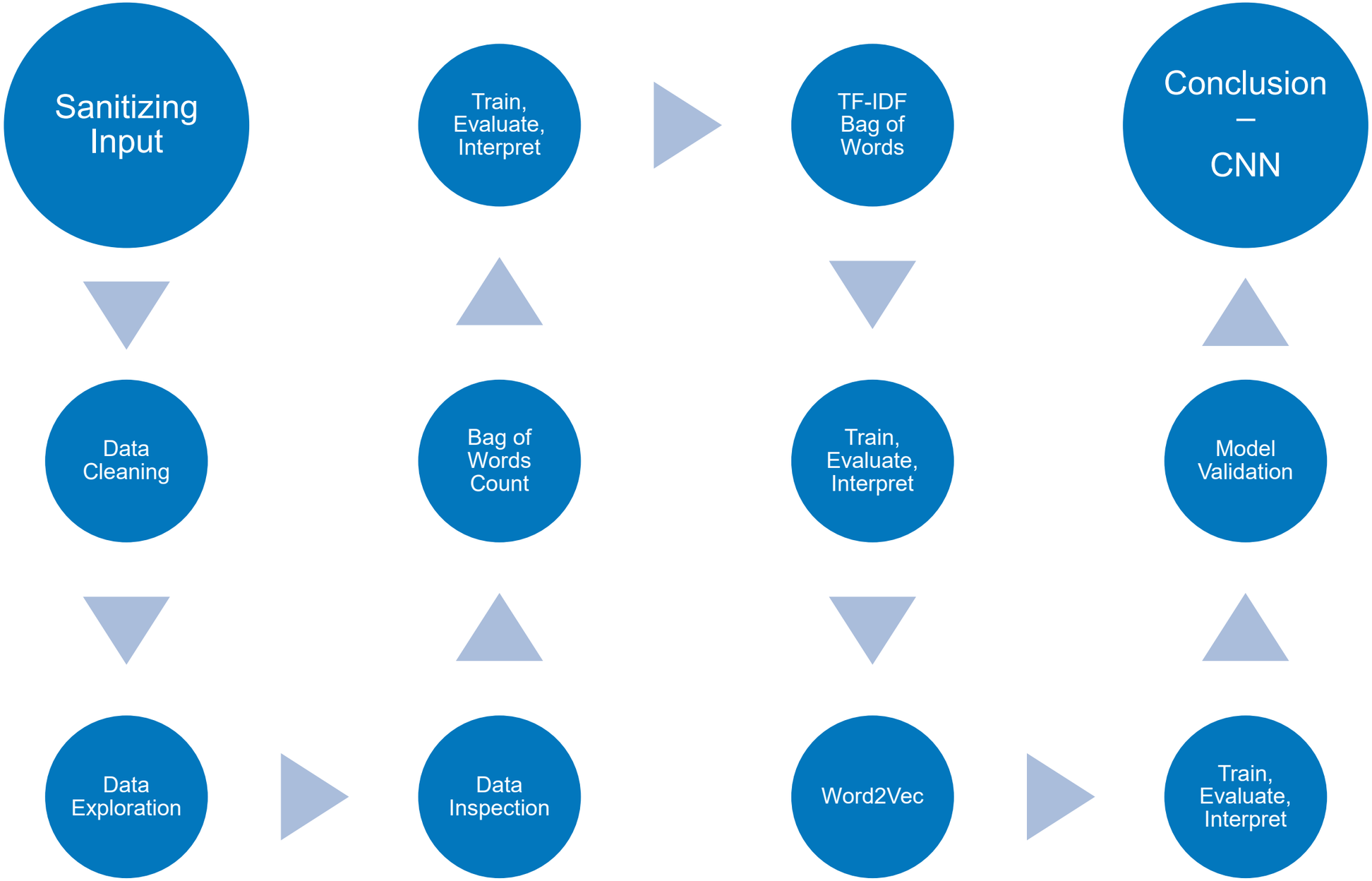
Input Dataset

	A	B	C	D	E	F	G	H	I
1	Source	Host	Link	Date(ET)	Time(ET)	time(GMT)	Title	TRANS_CONV_TEXT	Patient_Tag
2	FORUMS	cafepharma.com	http://cafepharma.com/boar	6/15/2016	13:58:00	6/15/2016 23:28	Epstein	I don't disagree with you in principle. I'm just saying that Entresto has bee	0
3	FORUMS	www.patient.co.uk	http://www.patient.co.uk/fc	05-07-2016	0.820833333	42498.21667	Enlarged Heart.Thr	I am always dizzy I get dizzy standing up so I have made myself get up slov	1
4	BLOG	http://abcnewsrad	http://abcnewsradioonline.c	4/14/2016	15:00:38	4/15/2016 0:30	Queen Latifah Join	Axelle/Bauer-Griffin/FilmMagic(NEW YORK) -- Queen Latifah is taking mat	0
5	FORUMS	www.cancer-forum	http://www.cancer-forums.r	6/18/2016	20:46:00	6/19/2016 6:16	Bulaemia	I am 17 and I have been throwing up for about a year now,almost everyda	1
6	FORUMS	www.diyaudio.con	http://www.diyaudio.com/fc	6/15/2016	03:26:00	6/15/2016 12:56	DIY Silver interconr	Quote: Originally Posted by Boyan Silyavski Wake up my friend, i was talki	0
7	FORUMS	forum.cyclinguk.or	http://forum.cyclinguk.org/v	04-06-2016	05:50:00	04-06-2016 15:20	Personal Question	Theres a discussion about recumbent power on BROL. The extract makes a	0
8	FORUMS	www.reddit.com	https://www.reddit.com/r/t	05-02-2016	0.047916667	42492.44375	TIL that CVS took a	Of course! I just got diagnosed with congestive heart failure and type 2 dia	1
9	BLOG	http://quranfruitz.	http://quranfruitz.blogspot.c	2/28/2016	03:20:00	2/28/2016 13:50	Causes of Low Bloo	Blood pressure is a measurement of the pressure in your arteries during tl	0
10	FORUMS	hmnews.org	http://hmnews.org/health-c	6/15/2016	17:16:00	6/16/2016 2:46	Sleep disorders ma	Other Sleep disorders may predict heart events after angioplasty People v	0
11	FORUMS	www.cafepharma.(	http://www.cafepharma.con	4/29/2016	23:46:00	4/30/2016 9:16	Nominations for w	worse manager ever??? LS out of Richmond. Ugh! What a self centered, I l	0
12	FORUMS	modelmayhem.cor	http://www.modelmayhem.	7/15/2016	15:49:00			Looknsee Photography wrote: When I was living in the Bay Area, Californi	1
13	FORUMS	blueheronhealthn	http://blueheronhealthnews	2/22/2016	18:52:00	2/22/2016 23:52	Comment on This ?	Posted by: Christian Goodman A new study reveals a devastating fact if yo	0
14	BLOG	http://medicaldevi	http://medicaldevicemarket	4/26/2016	05:48:00	4/26/2016 15:18	Blood Pressure Mo	Transparency Market Research has published a new market report titled, '	0
15	BLOG	http://www.becon	http://www.becomerichfast	3/18/2016	19:00:28	3/19/2016 4:30	How Much Tax Can	It?s been a big week here in the UK. This week George Osborne dropped t	0
16	FORUMS	www.fark.com	http://www.fark.com/comm	4/21/2016	10:17:00	4/21/2016 19:47	(9116353) Now is w	Russ1642 : So another anti-science anti-medicine person falls victim to a s	1
17	BLOG	http://news.healtl	http://news.health.com/201	3/28/2016	15:37:07	3/29/2016 1:07	Do-It-Yourself Bloo	By Steven Reinberg HealthDay Reporter MONDAY, March 28, 2016 (Healthl	0
18	FORUMS	finance.yahoo.com	http://finance.yahoo.com/m	6/17/2016	05:24:00	6/17/2016 14:54	How big is the ALS (	that's not where the biggest opportunity is - it's with heart failure drug - v	0
19	BLOG	http://wwj.cbsloca	http://detroit.cbslocal.com/2	03-06-2016	12:45:51	03-06-2016 23:15	Former First Lady N	LOS ANGELES (AP) ? Former first lady Nancy Reagan has died at 94 in Bel-A	0
20	FORUMS	guboards.spokesm	http://guboards.spokesmanr	7/25/2016	10:42:00			Originally Posted by Birddog Cause of death was an ""enlarged heart"". I'n	0
21	BLOG	http://mostdiet.bl	http://mostdiet.blogspot.cor	04-03-2016	03:18:00	04-03-2016 12:48	Pot belly ups heart	London: Obesity not only increases the risk of heart failure, but increased	0
22	BLOG	http://eliteathletic	http://eliteathletictraining.b	6/21/2016	16:08:00	6/22/2016 1:38	Is Exercise Better T	According to New York Times columnist, professor and physician Aaron E.	0
23	FORUMS	alkpathway.com	http://alkpathway.com/50-c	4/26/2016	22:51:00	4/27/2016 8:21	50 Moreover, the c	50 Moreover, the cytokines like TNF-??, IL-1?? and IL-6 are also 50 Moreov	0

Description of attributes in dataset

- Source - Type of Social Media Post
- Host - Domain of Social Media Post
- Link - Complete URL of post
- Date - Date of Post
- Time - time Stamp of Post in Eastern Time
- Time(GMT) - time Stamp of Post in GMT
- Title - Title of the Post
- TRANS_CONV_TEXT - Actual Text Conversation of the Post
- Patient_Tag - Patient Flag (1= Patient, 0=Non-Patient)

Exploratory Data Analysis Workflow



Step 1 – Sanitizing Input

- Ensure that text corpus does not contain any unknown or unwanted characters.
- Example – Remove ‘#’ symbols from online data

Step 2 – Data Cleaning

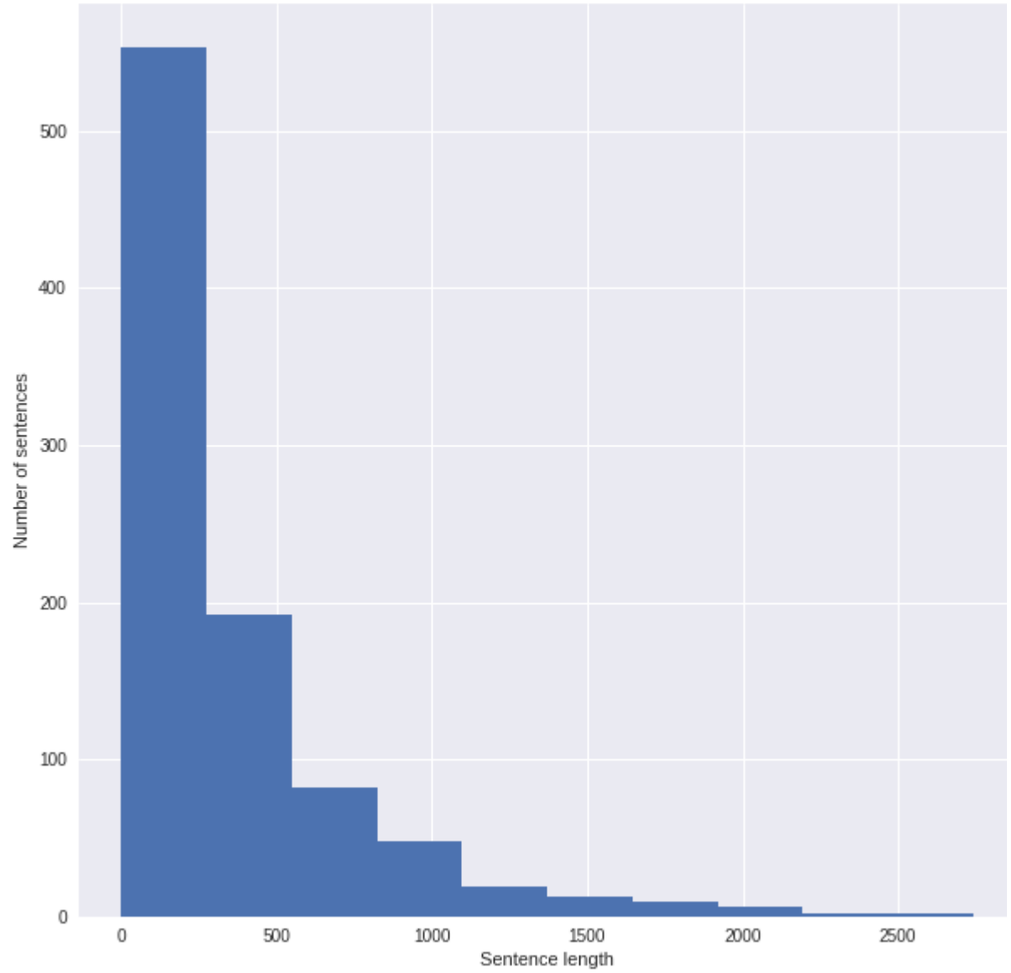
- Data Standardization – convert all characters to lowercase, replace ‘https’ and ‘@’ with ‘’ and ‘at’ respectively
- Drop NA columns

Step 3 – Data Exploration

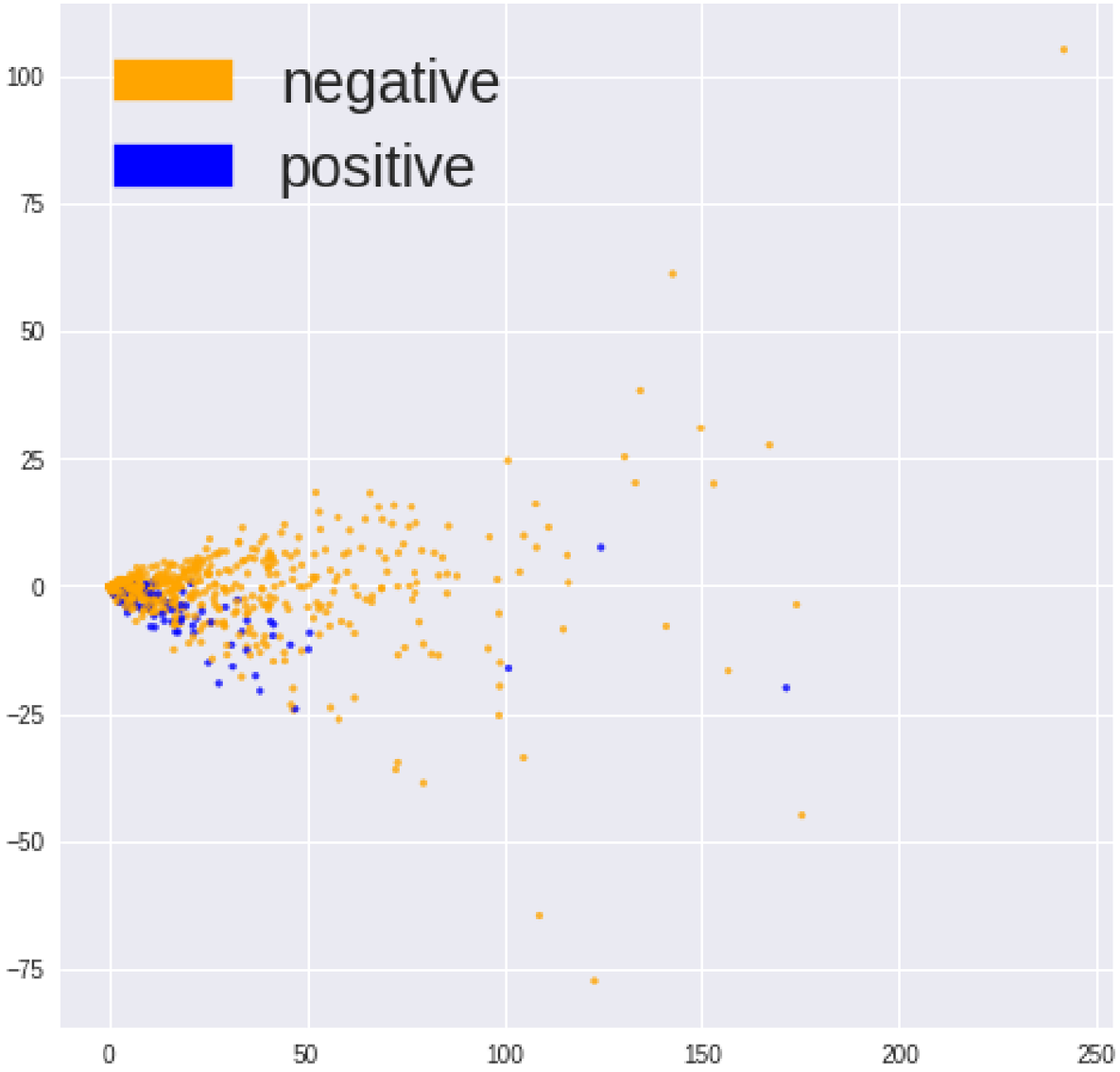
- 801 – Negative Examples, 125 – Positive Examples
- Ratio < 10:1 so undersampling / oversampling not mandatory

Step 3 – Data Exploration

- 313481 words total, with a vocabulary size of 22185
- Max sentence length is 2741



Step 4 – Data Inspection

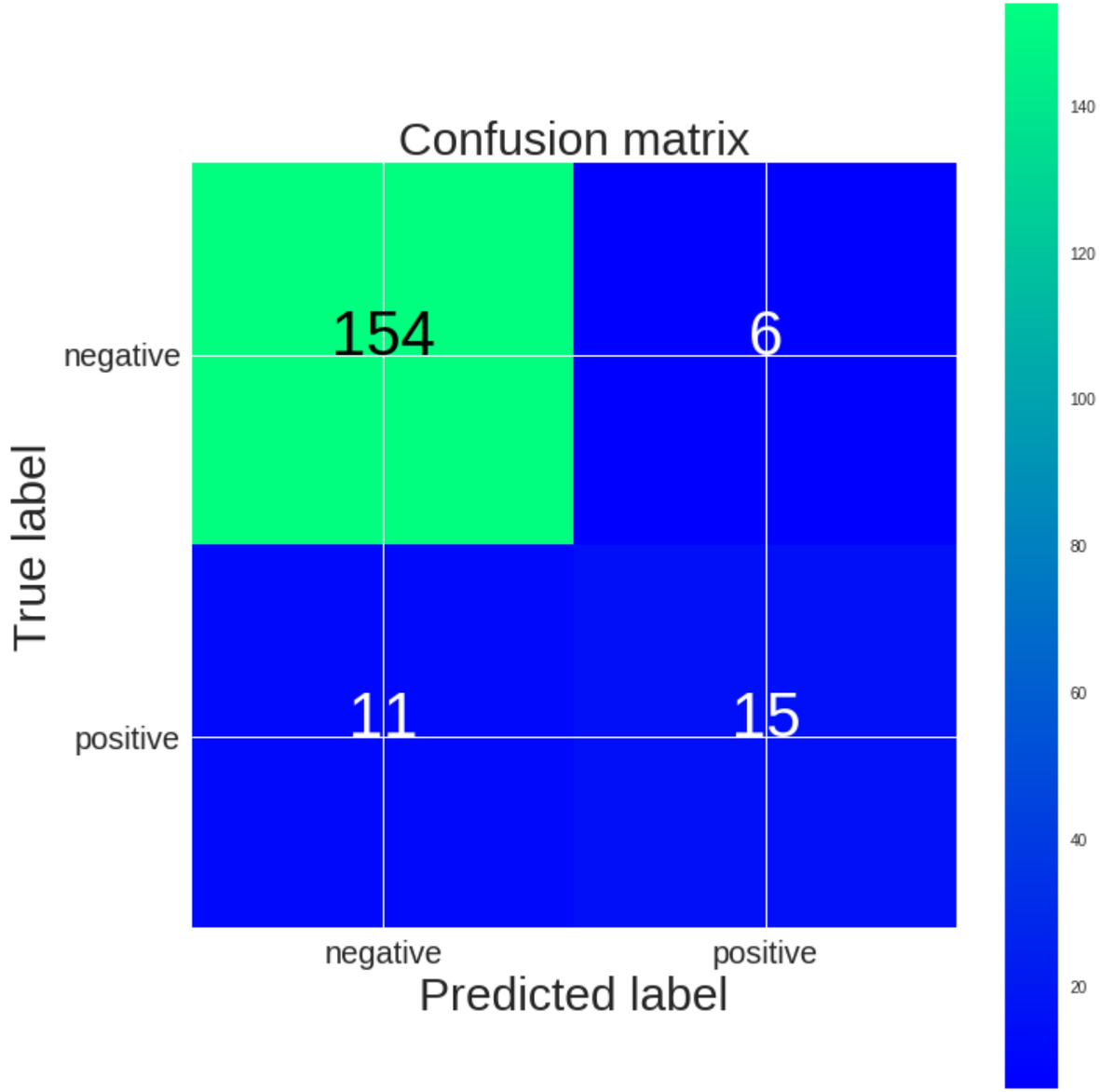


Step 5 – Bag of Words Count

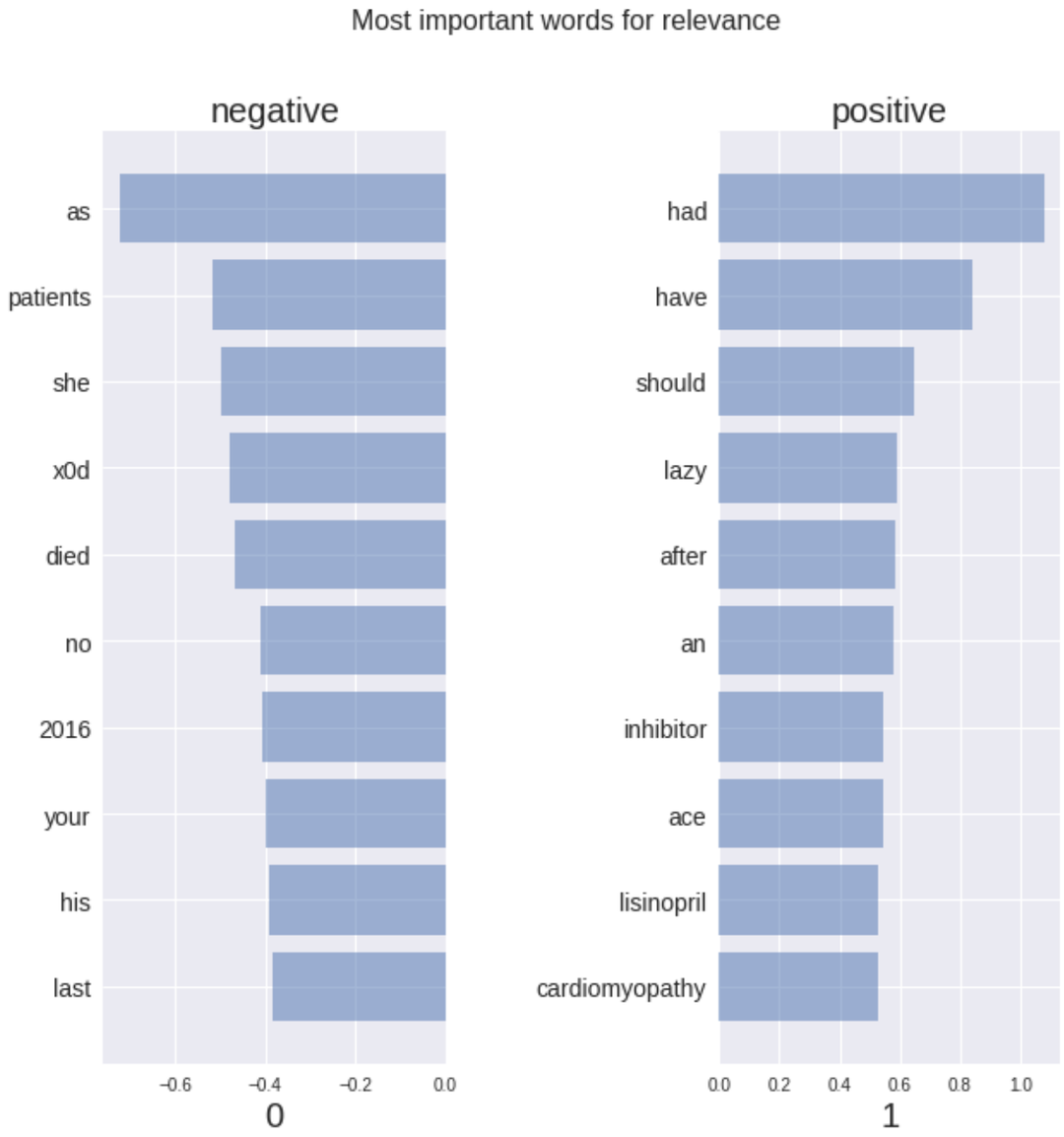
- Represent text for computers -> encode each character individually
- Create a useful embedding for each sentence, then use these embeddings to accurately predict the relevant category.
- Use a bag of words model, and apply a logistic regression on top.
- A bag of words just associates an index to each word in our vocabulary, and embeds each sentence as a list of 0s, with a 1 at each index corresponding to a word present in the sentence.

Step 6 – Train, Evaluate & Interpret

- Vanilla Logistic Regression algorithm
- Evaluation Metrics –
    1. accuracy = 0.909
    2. precision = 0.903
    3. recall = 0.909
    4. f1 = 0.904



Step 6 – Train, Evaluate & Interpret...



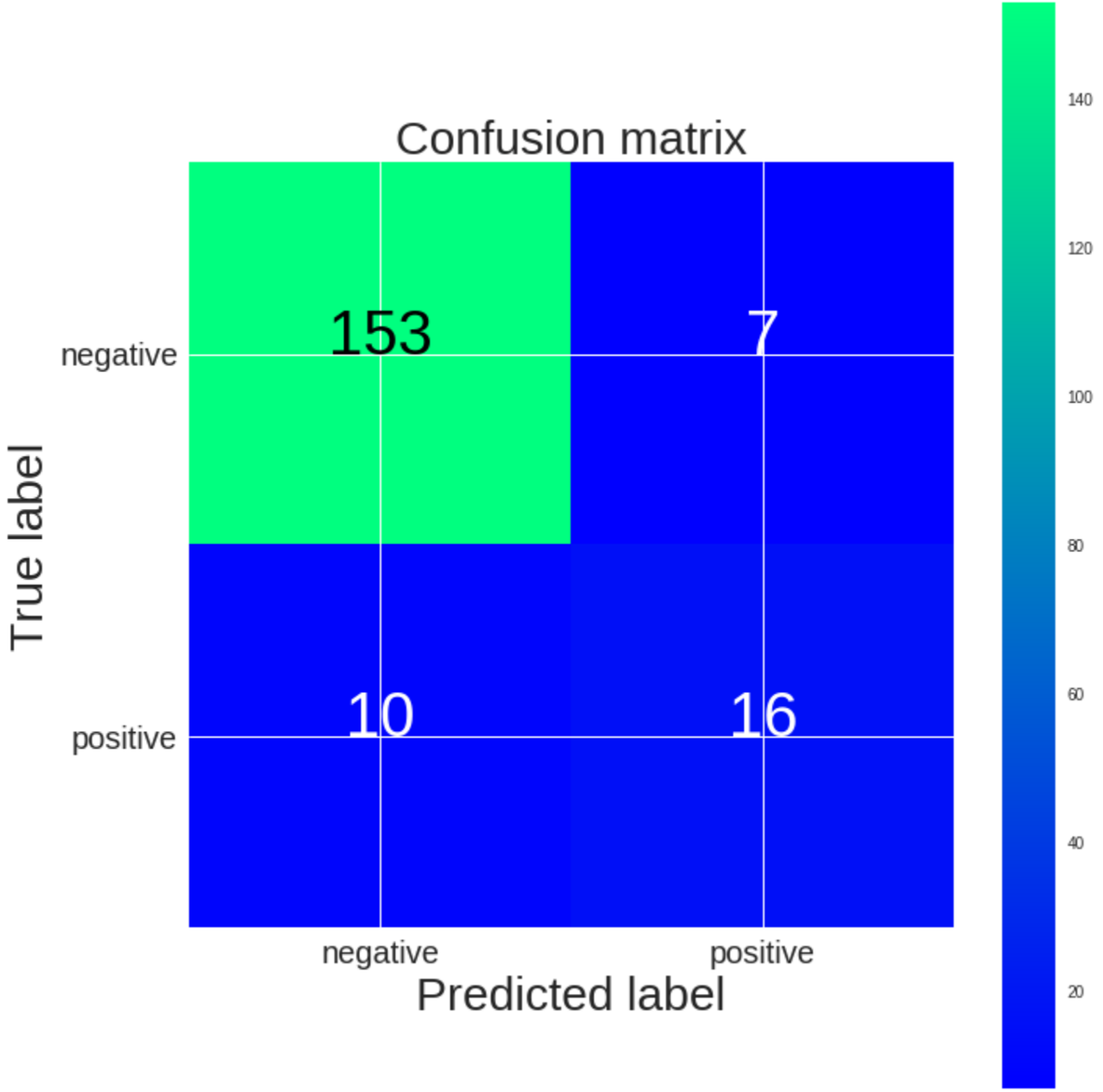
Step 7 – TF-IDF

- Term Frequency – Inverse Document Frequency.

- means weighing words by how frequent they are in our dataset, discounting words that are too frequent, as they just add to the noise.

Step 8 – Train, Evaluate & Interpret

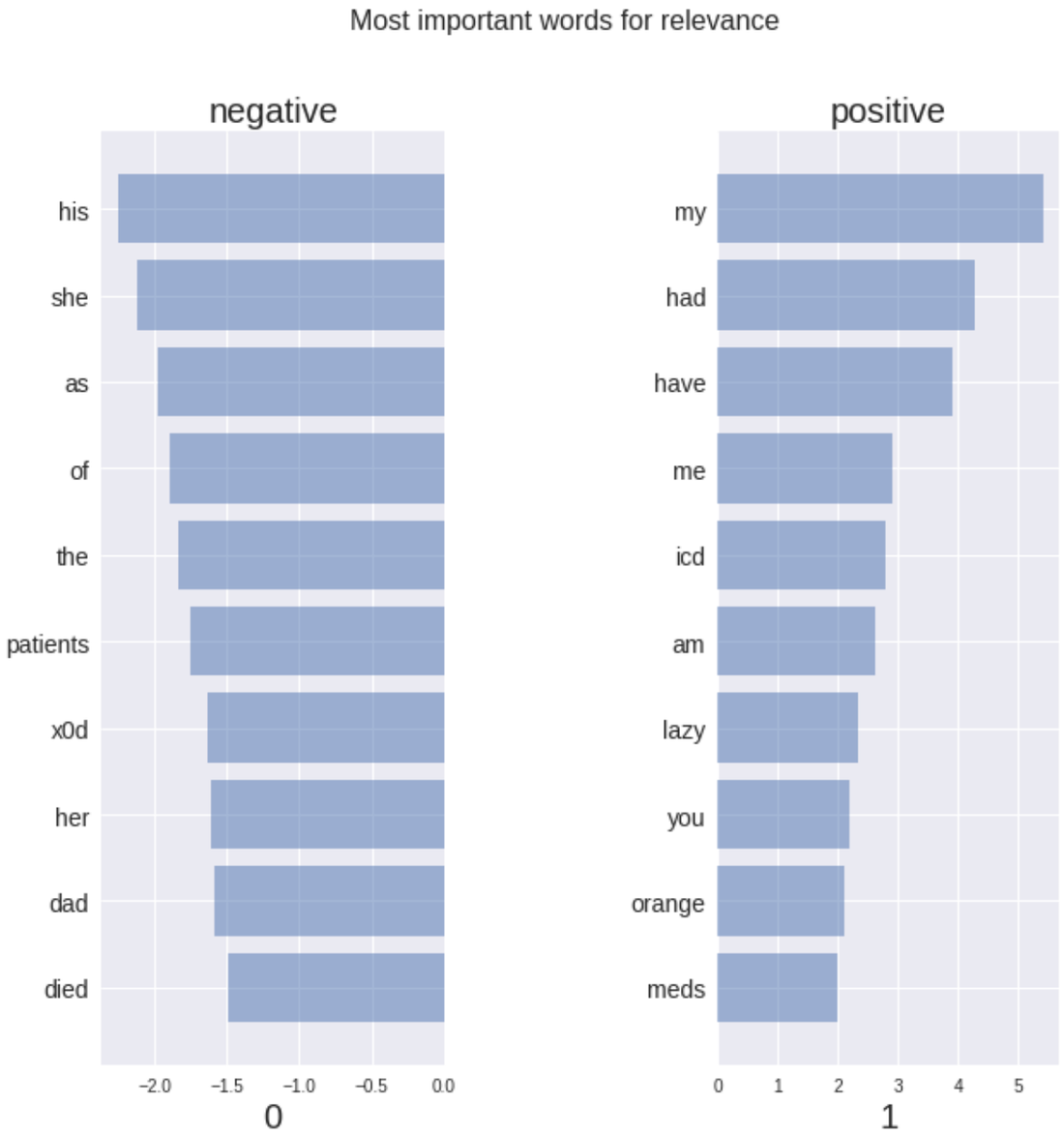
- Vanilla Logistic Regression algorithm
- Evaluation Metrics –
    1. accuracy = 0.909
    2. precision = 0.905
    3. recall = 0.909
    4. f1 = 0.906
- TFIDF confusion matrix [[153 7]
  [10 16]]
- BoW confusion matrix [[154 6]
  [11 15]]



Step 8 – Train, Evaluate & Interpret...

- Even though the accuracy and f1 score decreased, from the important words for relevance list and PCA projects we can say that using TF-IDF is yielding results which is more reasonable to deploy in production.
- Also note that True Positive Count increased after using TF-IDF which should be the most important metric in our use case, hence we conclude that TF-IDF is yielding better results.

Step 8 – Train, Evaluate & Interpret...

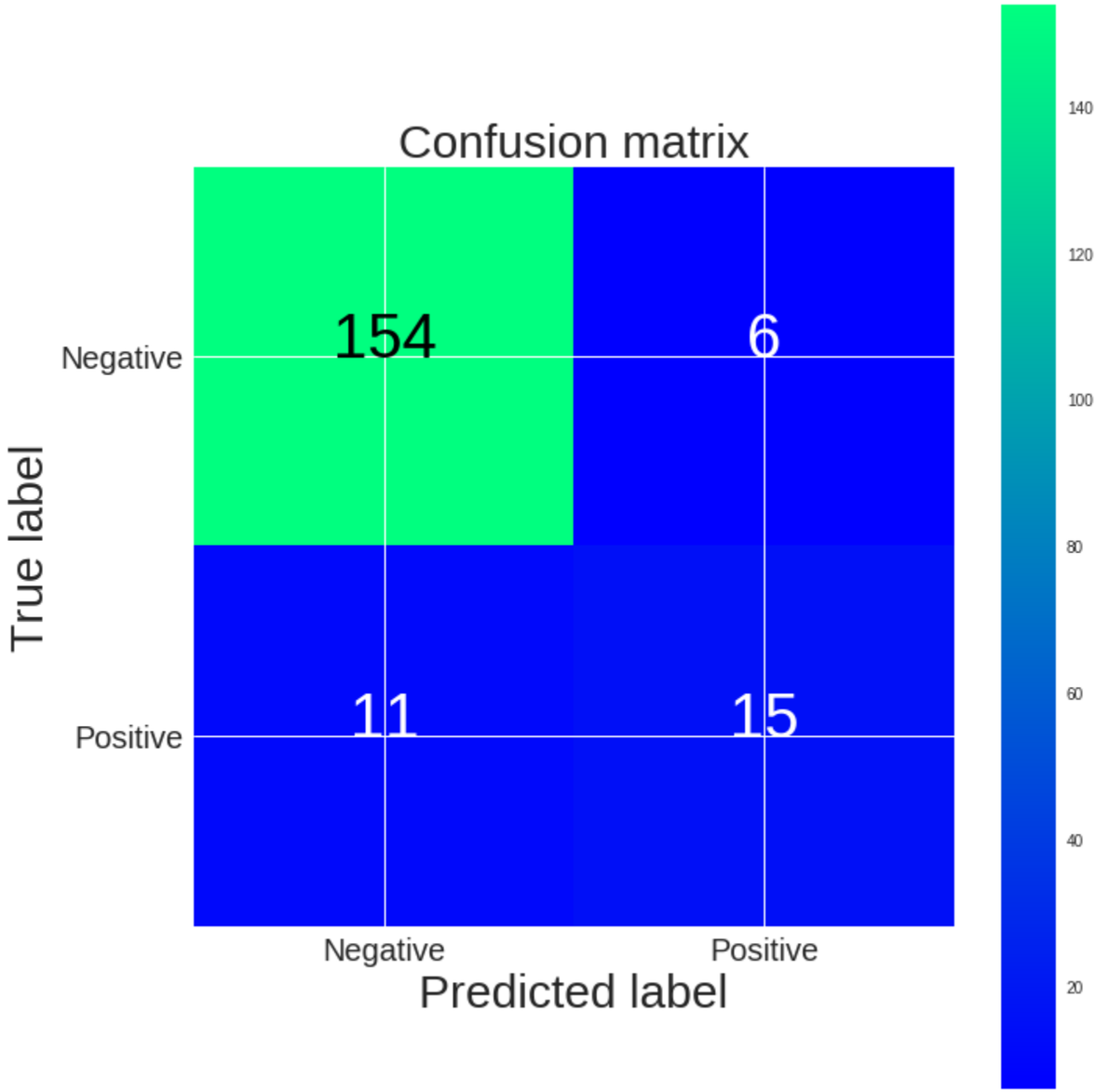


Step 9 – Word2Vec

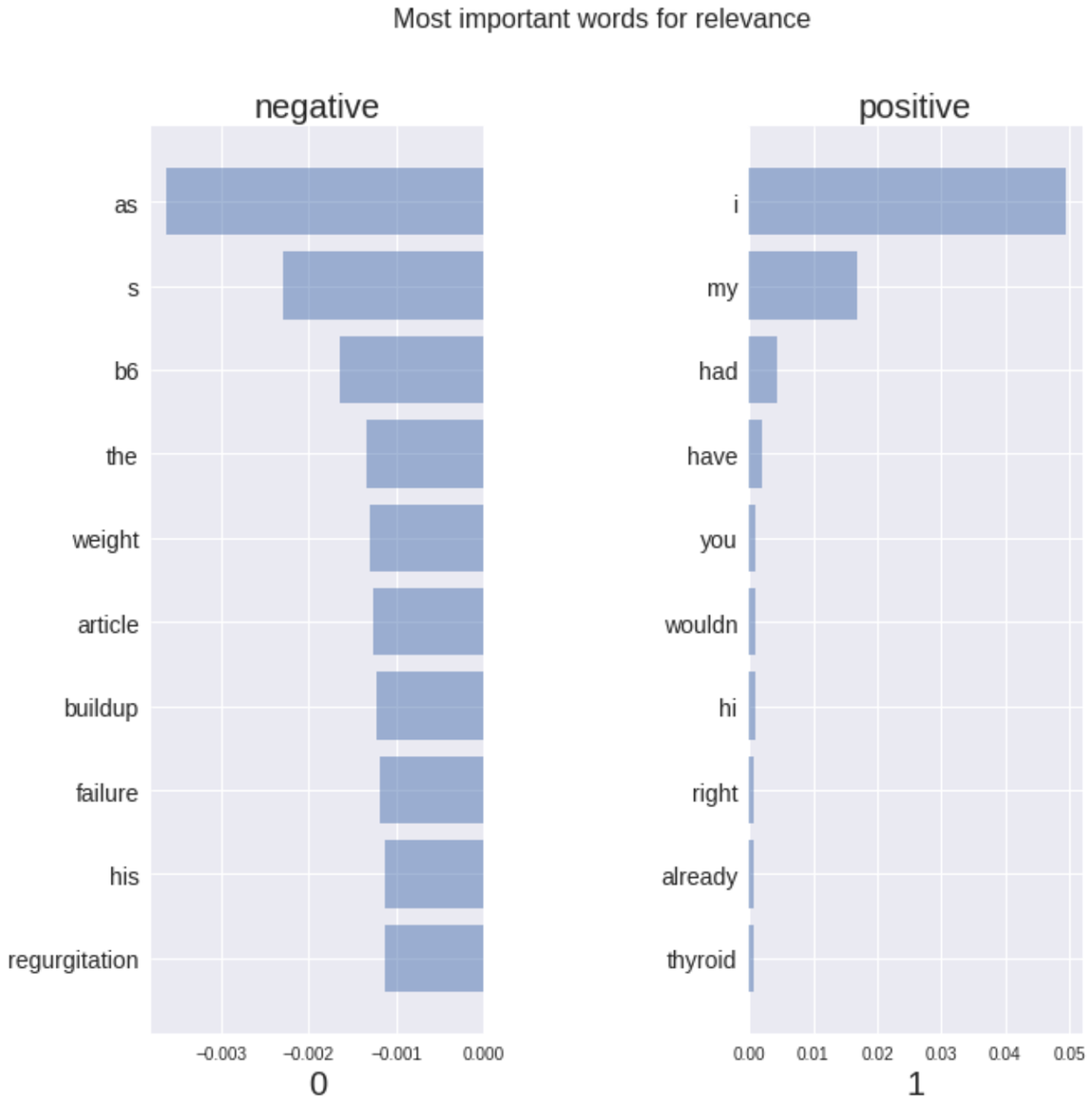
- Word2vec is a model that was pre-trained on a very large corpus, and provides embeddings that map words that are similar close to each other. A quick way to get a sentence embedding for our classifier, is to average word2vec scores of all words in our sentence.

Step 10 – Train, Evaluate & Interpret

- Vanilla Logistic Regression algorithm
- Evaluation Metrics –
    1. accuracy = 0.898
    2. precision = 0.916
    3. recall = 0.898
    4. f1 = 0.904
- Word2Vec confusion matrix [[146 14]
  [5 21]]
- TFIDF confusion matrix [[153 7]
  [10 16]]
- BoW confusion matrix [[154 6]
  [11 15]]

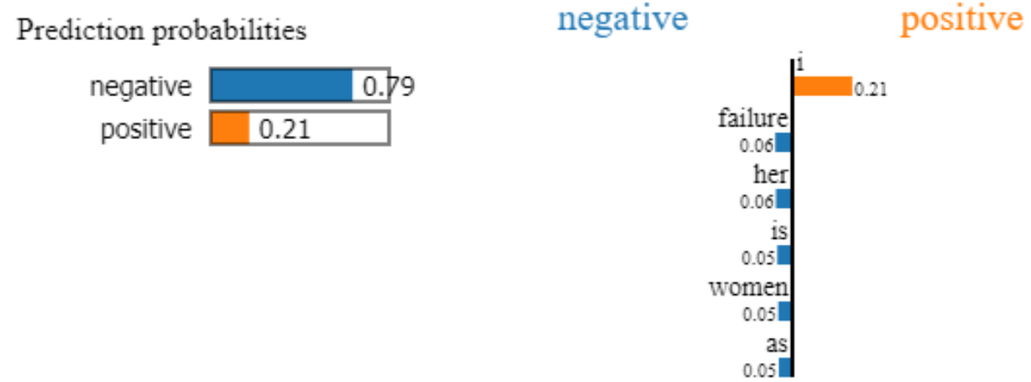


Step 10 – Train, Evaluate & Interpret...



Step 11 – Model Validation

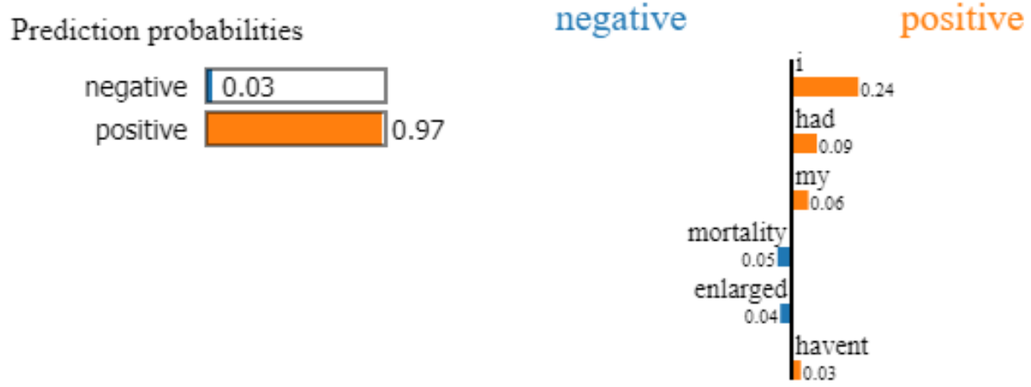
```
Index: 2
True class: negative
```



Text with highlighted words

?thank god i got here when i did,? 32 year old stacy rollins of napoleon, ohio, said during a recent checkup at the heart and vascular center at the university of toledo medical center a month ago, ut health cardiothoracic surgeon dr mark bonnell saved her life by implanting a battery powered blood pump inside her chest to take over for her failing heart stacy rollins talked with ut health cardiothoracic surgeon dr mark bonnell during a recent checkup she is sharing her story during american heart month february is american heart month, and rollins is sharing her story to encourage other younger women to take care of themselves and pay attention to early warning signs of heart failure ?i was in pretty good shape, but i had been under a lot of stress,? rollins said ?i started to feel terrible i couldn?t breathe at night i couldn?t go up the stairs i was coughing i thought it was pneumonia ? turned out her heart was barely pumping the cough wasn?t a cold the fatigue and breathlessness were symptoms of heart failure, which can become rapidly fatal in rollins? case, she had familial idiopathic cardiomyopathy ? a weakening of the heart muscle that is inherited with unknown cause her

```
Index: 135
True class: positive
```

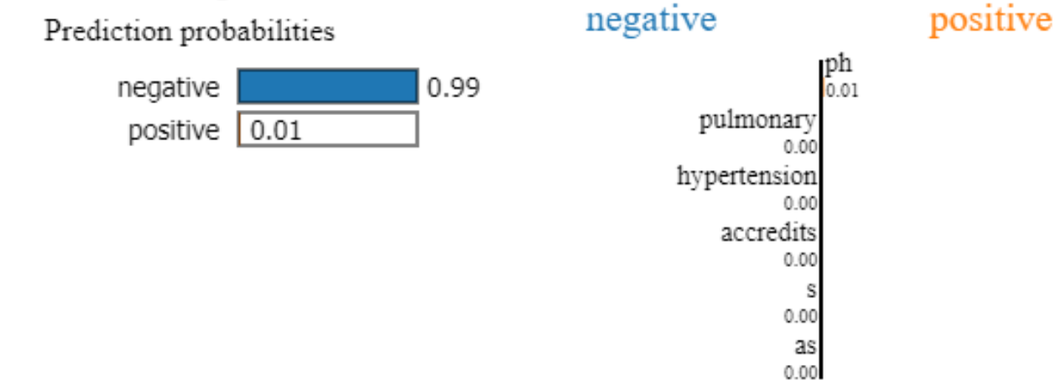


Text with highlighted words

i have had heart problems for 18 months now, i havent been diagnosed with a specific cause yet but the cardiologist is looking at myocarditis or heart failure my symptoms are worse so i went the docs and he had a xray done the results said i had an enlarged heart so he now wants me to have a echo again to see what is happening i cant seem to find much about enlarged hearts and the mortality so i was hoping some one here could answer my question, how long do people survive with enlarged hearts

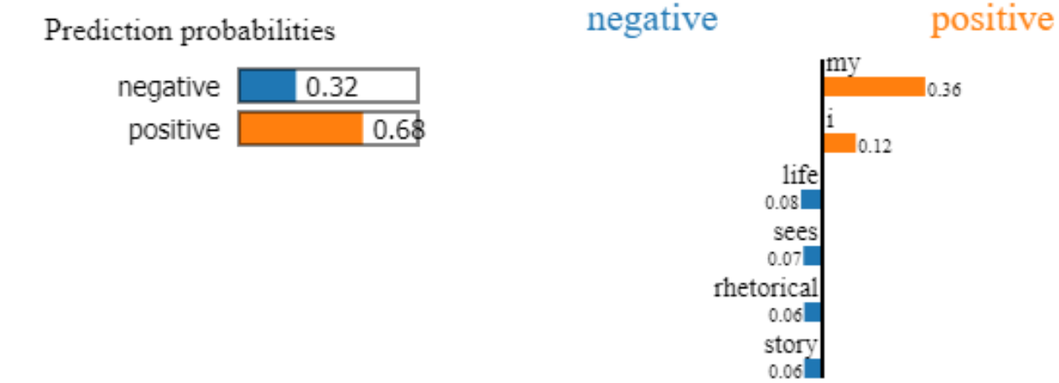
Step 11 – Model Validation

```
Index: 4
True class: negative
```



recognized for excellence in care for patients with ph since the launch of the pha accredited ph care center (phcc) program, a total of 34 center of comprehensive care (ccc) programs have been accredited, including 30 adult centers and four pediatric centers as centers are accredited, they also agree to contribute to the pulmonary hypertension association registry (phar) this patient registry will collect data used to evaluate outcomes for people living with ph and help researchers learn more about the rare disease, leading to the possibility for new breakthroughs in treatment and quality of care for more information on phcc, go to or email phcc about the pulmonary hypertension association headquartered in silver spring, md , with a growing list of chapters across the country, the pulmonary hypertension association (pha) is the country?s leading pulmonary hypertension organization its mission is to extend and improve the lives of those affected by ph its vision is a world without ph, empowered by hope pha achieves this by connecting and working together with the entire ph community of patients, families and medical professionals among its programs, pha facilitates more than 245 support groups around the country and delivers continuing education for medical professionals through pha online university now celebrating its 25th anniversary, pha has provided more than 17 million in ph research commitments for more information, please go to , on twitter or for the

```
Index: 28
True class: positive
```



Text with highlighted words

lindalouwho revrendjim the local paper here recently ran a story on the highest paid medical people in the area most were hospital administrators, but my cardiologist was on the list at just under a million a year this man literally saved my life i would not still be here wasting my time reading the bullshiat you people post, and occasionally posting my own stupid bullshiat, if not for his management of my heart failure so i'm okay with this raises hand also ok with this, have had my life saved a few times the shame (and problem) is no one wants to be a general practitioner family doctor anymore, too many areas are underserved also about money isn't everything? rhetorical my wife is in family medicine and was at the top of her class in med school, so she could have gone in other directions but she loves it and the patients she sees the 5x more paycheck sure would be nice though

Step 12 – CNN for Text Classification

- Embeddings Generation - Using Tensoflow's learn.preprocessing.VocabularyProcessor module to transform each word in the text corpus into a vector space (Word2Vec).

- Patient Conversation Classification - Convolutional Neural Network (CNN)

1. used an embedding layer, followed by a convolutional, max-pooling and softmax layer.
2. used 50% dropout and L2 regularization in Cross Entropy loss in to avoid over-fitting.

a. Model Hyperparameters –

-> Dimensionality of character embedding – 128

-> Filter sizes - (3,4,5)

-> Number of filters per filter size - 128 ,

-> Dropout keep probability - 0.5 (50%) ,

-> L2 regularization lambda - 0.0 ,

b. Training Parameters -

-> Batch Size - 64 ,

-> # of Epochs - 200 ,

-> Evaluate model on Dev Set - every 100 steps ,

-> Save Model checkpoint - every 100 steps ,

-> Number of checkpoints - 5

Codebase

- Exploratory Data Analysis Notebook - https://github.com/indranildchandra/online-patient-conversation-classifier/blob/master/src/Research_Online_Patient_Conversation_Classifier.ipynb

- Final Submission Notebook – https://github.com/indranildchandra/online-patient-conversation-classifier/blob/master/src/Online_Patient_Conversation_Classifier.ipynb

Contact Details

https://about.me/indranilchandra

https://github.com/indranildchandra

https://in.linkedin.com/in/indranildchandra

Indranil Chandra
indranildchandra@gmail.com
@IndranilChandra