Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Collecting and Performing Sentimental Analysis on Tweet

Natural Language Processing

July 8, 2021

Karthik Sundararaj
Kabilan Tamilmani
Vishwas Sharma

Supervised by

Prof. Dr. Paul Plöger, M.Sc. Tim Metzler

Sentiment Analysis

What it is and why it is important

- It identifies and extract subjective information and analye it for sentiments of the data (i.e. Positive, Negative and neutral)
- It is important for the business as it can get customer's sentiment about the business product

Sentiment Analysis

How it became popular?

- The recent advancement in deep learning make us able to do indepth research in the field of sentiment analysis (getting the high value insight from the data).
- This research helps us to get customer's key aspects of products and the service they require.

Dataset

- We used ’Sentiment140’ dataset [1] from Kaggle as our training dataset . It contains 1.6m tweets labeled as either positive or negative.

```
[39] df.head(10)
```

	target	ids	date	flag	user	text
1577700	POSITIVE	2189710623	Tue Jun 16 00:42:33 PDT 2009	NO_QUERY	KayteeNelson	wow, im actually pretty drunk
808010	POSITIVE	1469110749	Tue Apr 07 05:14:42 PDT 2009	NO_QUERY	amysav83	@marleyuk aaaawwwww i taught her well then
104181	NEGATIVE	1822388081	Sat May 16 19:27:48 PDT 2009	NO_QUERY	piticanella	@mistygirlph hey I am still on saturday here....
298724	NEGATIVE	1997774478	Mon Jun 01 17:26:06 PDT 2009	NO_QUERY	LilBit0318	@ShawneyJ : *LOL* Astros fan here! ;-) And dyi...
1322794	POSITIVE	2014834087	Wed Jun 03 02:50:30 PDT 2009	NO_QUERY	DomenicY	is listening to Ghost Man On Third - Taking ba...
396868	NEGATIVE	2056483794	Sat Jun 06 11:13:40 PDT 2009	NO_QUERY	flamingokitty	someone want to tell me why i'm so close to cr...
119174	NEGATIVE	1827991242	Sun May 17 11:52:05 PDT 2009	NO_QUERY	MzWhyteWashed	not going to Greek festival that's what i'm do...
228919	NEGATIVE	1978364208	Sat May 30 23:08:15 PDT 2009	NO_QUERY	SM41890	F**K IT... guess ill go to bed... alone... ...
924249	POSITIVE	1754937439	Sun May 10 07:48:58 PDT 2009	NO_QUERY	roadichose	Focus:Client laundry, cleaning, and towels my ...
248477	NEGATIVE	1982725524	Sun May 31 11:46:45 PDT 2009	NO_QUERY	kbibbs	Getting ready for my niece's 6th birthday part...

Figure 1: Sentiment140 Dataset

Positive and negative

```
df_positive["text"].tolist()[:5]
```

```
["Well, #SWGeekGirls have got to stick together, @lomara! Not many have surfaced that I've been good friends with. ",
 '@wearestereos , YOU GUYS ARE NUMBA ONE ON MOD DAILY ! Congratss  I voted like 500 times a day for you guys ! i la-la-la love you guys!',
 'Happy Birthday to me!  ',
 '@lizziechristine Thank you so much  Miss you toooo!',
 "got back from relay for life, it was pretty good, next year will be better cuz I'll be joining in on the walk "]
```

```
df_negative["text"].tolist()[:5]
```

```
['@enterbelladonna that link requires a password and member name ',
 'I want new Teeth! ',
 'The conference is over, time to study for finals now ',
 "Dammit, I forgot to color my hair. I'm gonna do it now. I have to color my hair back to black again coz it keeps washing off. ",
 'I hate commercials ']
```

Figure 2: Positive and Negative Tweet

Distribution of dataset

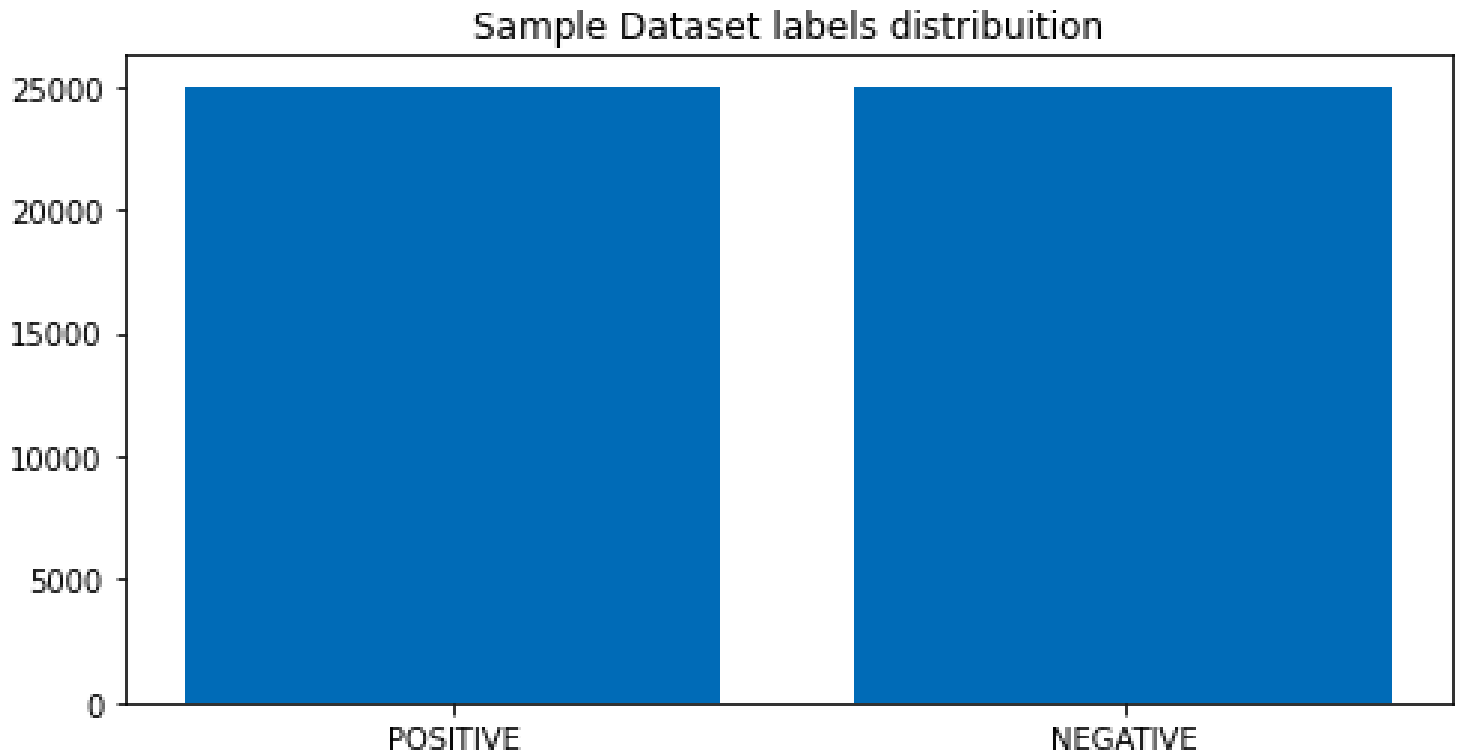


Figure 3: Data distribution

Tokenization and Denoising

```
["I'm cold  my feet 'specially. They might fall off. I kind of need them for things.",
 '@munziii: Definetly not ',
 'Whomp whomp... No more puppet commercials  Thanks GOD for youtube',
 "@EliseInChicago @Soulati This|close confirming return to 2nd hometown but in Aug.  Go ahead, and I'll catch up on all missed beers then.",
 'i am kind of mad right now cause outside is like rainy and cloudy  what happened to the sun!?!?!',
 'im bored. gonna go on tj and watch mitchell davis vids ',
 'I am soo burnt from playing volleyball and working...im going to be soo sore tomorrow! ',
 '@gdmpwml whys that? work reasons? ',
 'Early bed time tonight. My head hurts. Night all in the land of twitter. ',
 "@insinglefile i hate what you've done to me "]
```

Figure 4: Before Prepocessing

Tokenization and Denoising

```
['cold foot specially might fall kind need thing',
 'definetly',
 'whomp whomp puppet commercial thanks god youtube',
 'soulati close confirm return 2nd hometown aug go ahead catch missed beer',
 'kind mad right cause outside like rainy cloudy happen sun',
 'im bore gonna go tj watch mitchell davis vids',
 'soo burnt play volleyball work im go soo sore tomorrow',
 'whys work reason',
 'early bed time tonight head hurt night land twitter',
 'hate do']
```

Figure 5: After Prepocessing

Data visualization

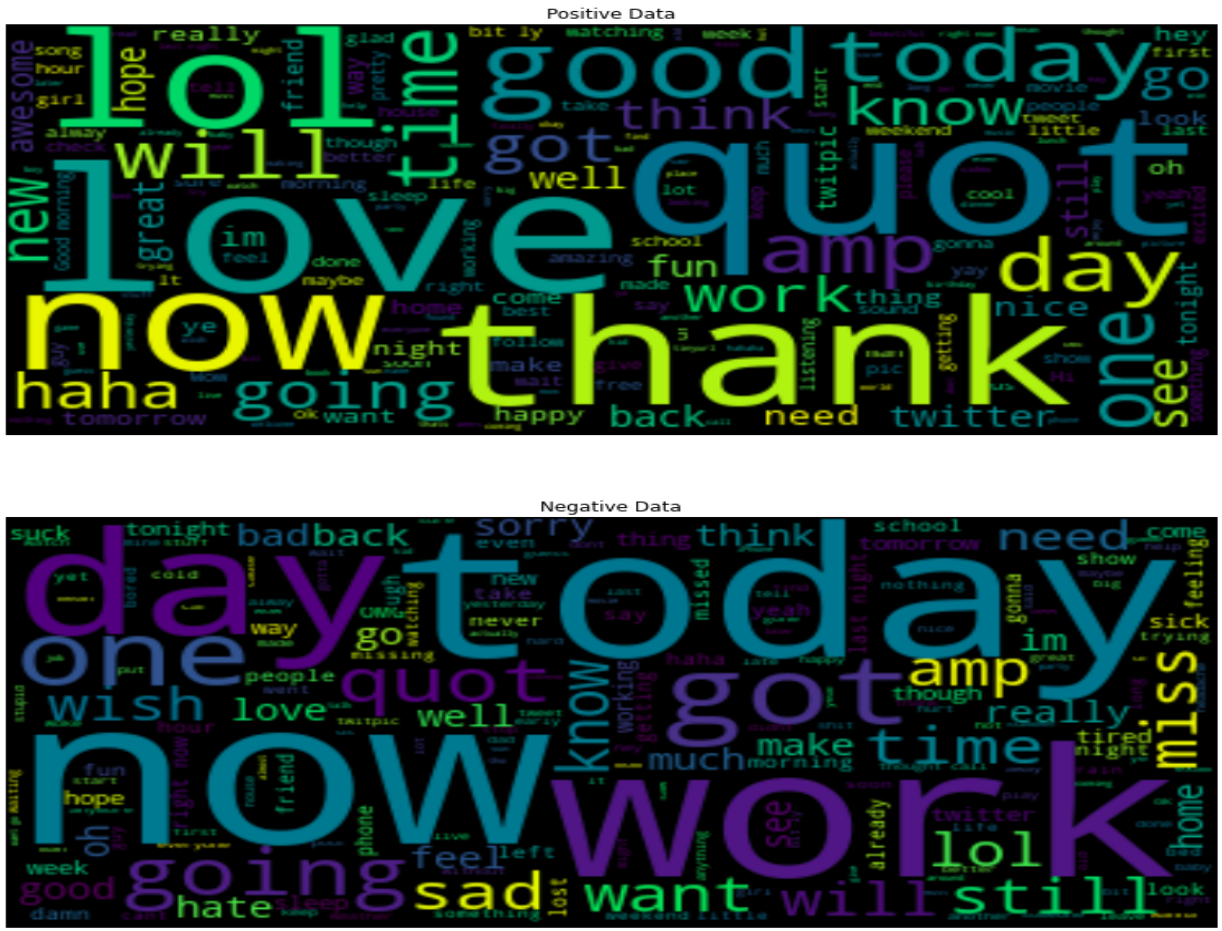


Figure 6: Positive and Negative word cloud

RandomForest Model

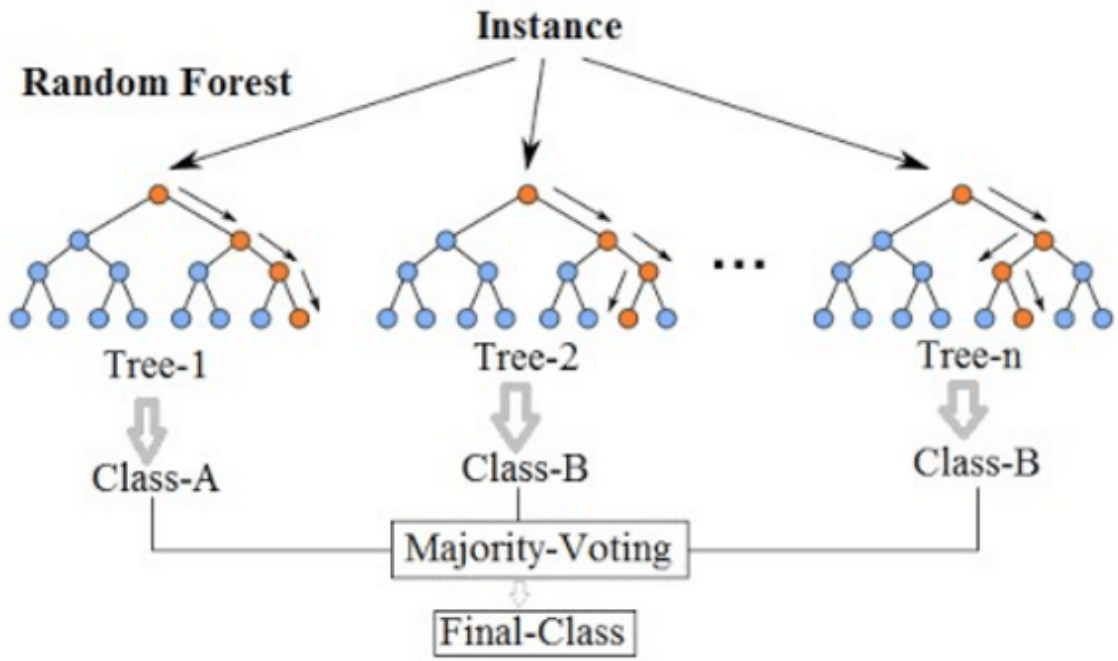


Figure 7: RandomForest Model

Model analysis

```
[[632 387]
 [504 477]]
              precision    recall  f1-score   support

           0       0.56      0.62      0.59      1019
           1       0.55      0.49      0.52       981

    accuracy                           0.55      2000
   macro avg       0.55      0.55      0.55      2000
weighted avg       0.55      0.55      0.55      2000

0.5545
```

Figure 8: RandomForest Model Summary

Word2Vec vs FastText Embedding Model

- **Word2Vec model**
  - In Word2vec model, the word with similar "meaning or relation" are mapped into a common vector.
  - We want to use the surrounding words to represent the target words with a Neural Network whose hidden layer encodes the word representation.
- **FastText model**
  - Model is a extension of Word2Vec model developed by Facebook AI Research center
  - It breaks a word into several n-grams and the embedding vector is the sum of all these n-grams.

Sentiment Model

```
Model: "SentimentModel"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
EmbeddingLayer (Embedding)   (None, 300, 300)         9933900
_________________________________________________________________
dropout_1 (Dropout)          (None, 300, 300)          0
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)               160400
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 101
=================================================================
Total params: 10,094,401
Trainable params: 160,501
Non-trainable params: 9,933,900
_________________________________________________________________
```

Figure 9: Embedding Model Summary

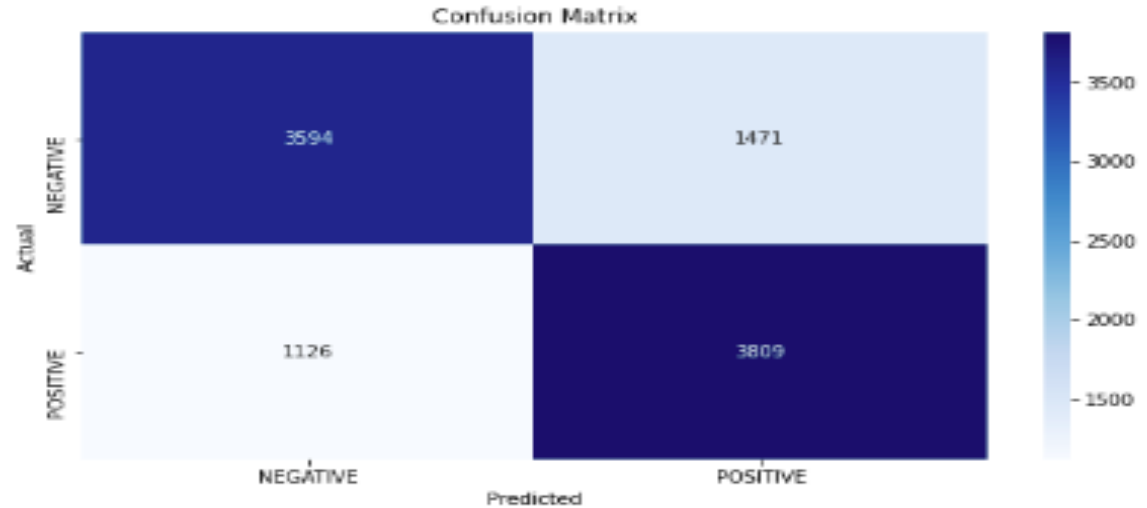
Model analysis

```
Classification Report
              precision    recall  f1-score   support

    NEGATIVE       0.76      0.71      0.73      5065
    POSITIVE       0.72      0.77      0.75      4935

    accuracy                           0.74     10000
   macro avg       0.74      0.74      0.74     10000
weighted avg       0.74      0.74      0.74     10000

Accuracy :74.03%
```




```
CPU times: user 45.8 s, sys: 1.92 s, total: 47.8 s
Wall time: 41.4 s
```

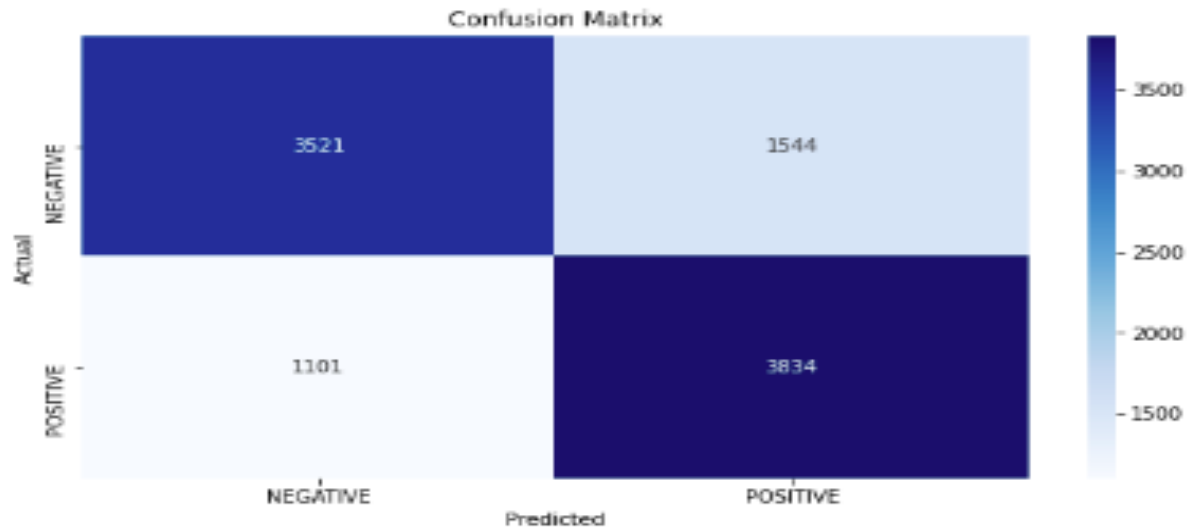
Figure 10: Word2Vector Summary

Model analysis

Classification Report

	precision	recall	f1-score	support
NEGATIVE	0.76	0.70	0.73	5065
POSITIVE	0.71	0.78	0.74	4935
accuracy			0.74	10000
macro avg	0.74	0.74	0.74	10000
weighted avg	0.74	0.74	0.74	10000

Accuracy :73.55%



CPU times: user 53.9 s, sys: 3.58 s, total: 57.4 s
Wall time: 42.3 s

Figure 11: FastText Summary

Tweepy

- It is an open source Python package that gives you a very convenient way to access the Twitter API with Python.
- Authentication is required to access this API
- We have to give topic, counts of tweet we want to scrap and language of the tweet
- While scrapping we can neglecting retweets

Pfizer tweet

```
[65] tweet_texts = [tweet.text for tweet in tweets]
     tweet_texts[:10]
```

```
['Double Vaxxed. #AstraZeneca x #Pfizer',
 'DOES ANYONE KNOW what vaccine is being given at the #Downsview #Hangar Sports and Event Centre? #vaccine #pfizer #moderna #toronto #ontario',
 'Me after my second Pfizer vaccine #FullyVaccinated #pfizer #perfect https://t.co/UIqgt8CZD8',
 'i love my #Pfizer booty',
 '#Pfizer #Vaccine Causes #HeartAttack and #Myocarditis In Healthy Young Man: Christopher Boeckman's Story https://t.co/UvvJx2hf9p',
 'Check out my Gig on Fiverr:Do awesome jewelry background remove and retouching https://t.co/7qyTkEUeEA\n#DOOMETERNAL… https://t.co/yL5iCEOPQK',
 'Double dosed! 💉#AstraZeneca 💉#Pfizer',
 'Check out my Gig on Fiverr: amazon product photo background remove, retouching https://t.co/Y35c03FjYW\n#DOOMETERNAL… https://t.co/tk6xuwtKK2',
 'Double dosed! #Pfizer #Pfiderna #Covid19AB #GetVaccinated #seconddose #yeg https://t.co/yT43UlLQko',
 'Had to do some scrolling to find this, got my shot #2 today, same beat me up shirt, Go Habs Go.  #Pfizer… https://t.co/H2Z0sLpPln']
```

Analyzing tweet

	A	B	C	D
		tweets	label	score
0		The first #Pfizer shot was put in a nurse's arm in Long Island, NY, on Dec 14. That is two hundred days ago. Two h,Ä¶ https://t.co/EOKON5nEWw	NEGATIVE	0.455356806516647
1		#COVID19: #Moderna fourth vaccine to get nod in India; #Pfizer next https://t.co/PgkpsbWrme	POSITIVE	0.544526696205139
2		Officially vaccinated (x2). I can confidently state it,Äôs been a full 28 minute car ride with no side effects whatsoever. #Pfizer	NEGATIVE	0.479852795600891
3		@CityofJax @UFHealthJax @FeedingNEFL @AgapehealthJax Dear @UFHealthJax @CDCgov @JNJNews Any info on,Ä¶ https://t.co/GsL67Vnv0k	POSITIVE	0.716200232505798
4		#Coronavirus: #Bennet visited the #vaccination center and insisted on young people to get vaccinated as soon as pos,Ä¶ https://t.co/leKQi9Jlrk	POSITIVE	0.594818830490112
5		#Pfizer Reaps Hundreds of Millions in Profits From #Covid #Vaccine yet the company can not be held liable for any i,Ä¶ https://t.co/Can85aBFcu	POSITIVE	0.580791354179382
6		$MFST MEDICAL LASERS MANUFACTURER , FDA CLEARANCE, MEDICAL TRIALS $hpil $pbya $cybl $eeenf $phil $cowi $hcmc $apyp,Ä¶ https://t.co/KE7jz4F2Sa	POSITIVE	0.560133934020996
7		Able to watch part of the #ENGGER match in the vaccine queue in Rio de Janeiro t©üè¥Û†ÅßÛ†ÅcÛ†Å•Û†ÅÆÛ†ÅßÛ†Åø©üáß©üá∑ #ItsComingHome #Pfizer,Ä¶ https://t.co/dYouVl0rO0	POSITIVE	0.568073093891144
8		#Moderna gets nod as 4th #CovidVaccine in India, #Pfizer next https://t.co/4vY0KjlGG3	NEGATIVE	0.453714400529862
9		@gyan_chakshu @WTF_1234567890a @ndtv #Cipla knew exactly which palms to grease in Govt of India. #Pfizer did not.,Ä¶ https://t.co/9zhz7fkXcs	NEGATIVE	0.32682204246521
10		Strange vaccine story! When you take the shot you don,Äôt feel any pain but after an hour you will feel so heavy! Can,Ä¶ https://t.co/LTYh7xSsob	NEGATIVE	0.052794724702835

Pfizer tweet Analysis

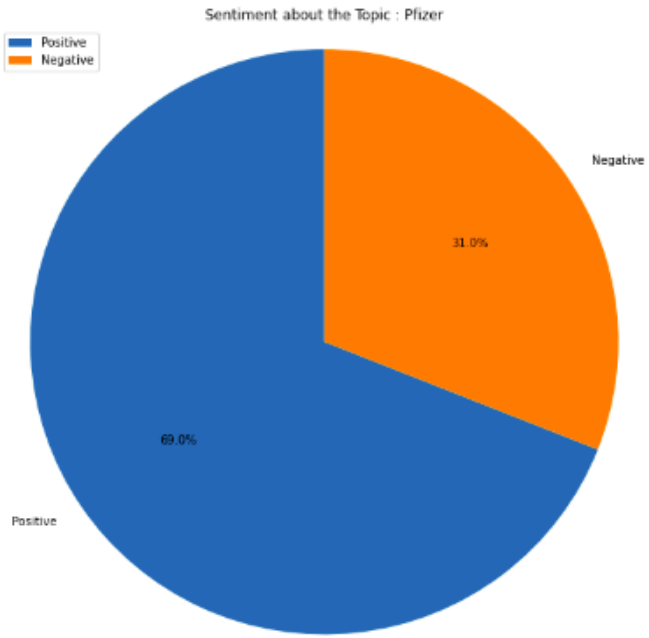


Figure 12: Pfizer Tweet analysis

Astrazeneca tweet Analysis

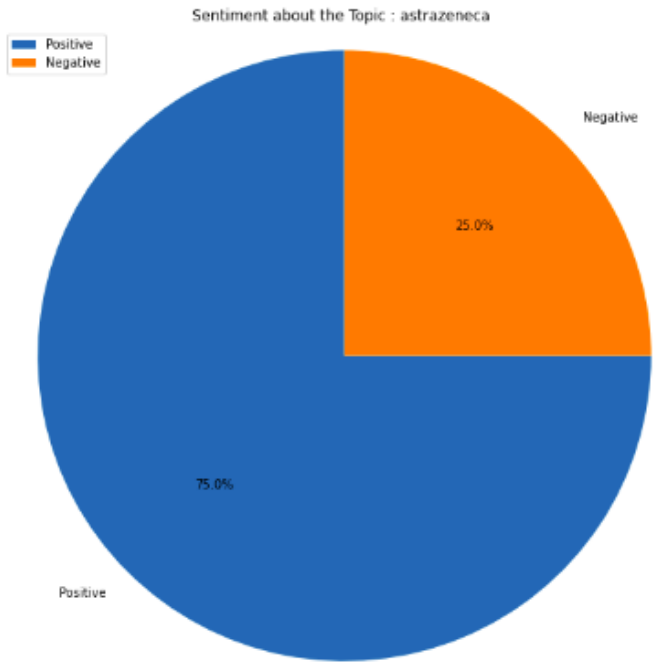


Figure 13: Astrazeneca Tweet analysis

Sputnik tweet Analysis

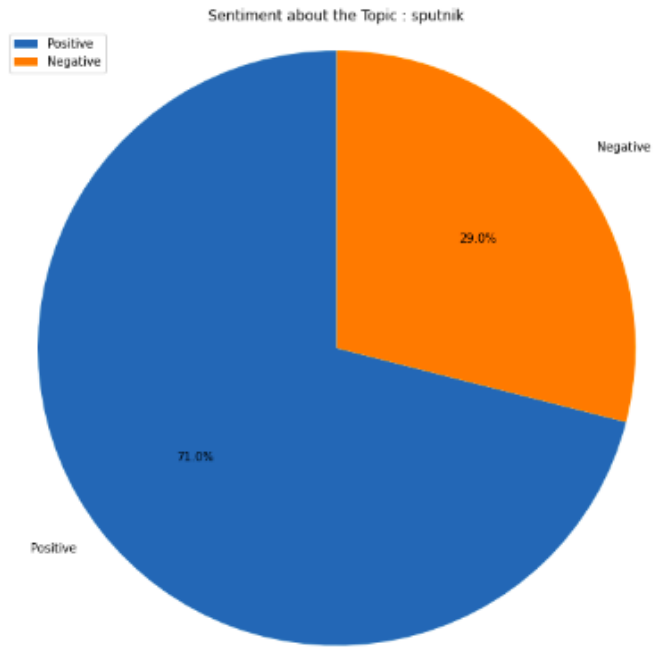


Figure 14: Sputnik Tweet analysis

CovieShield tweet Analysis

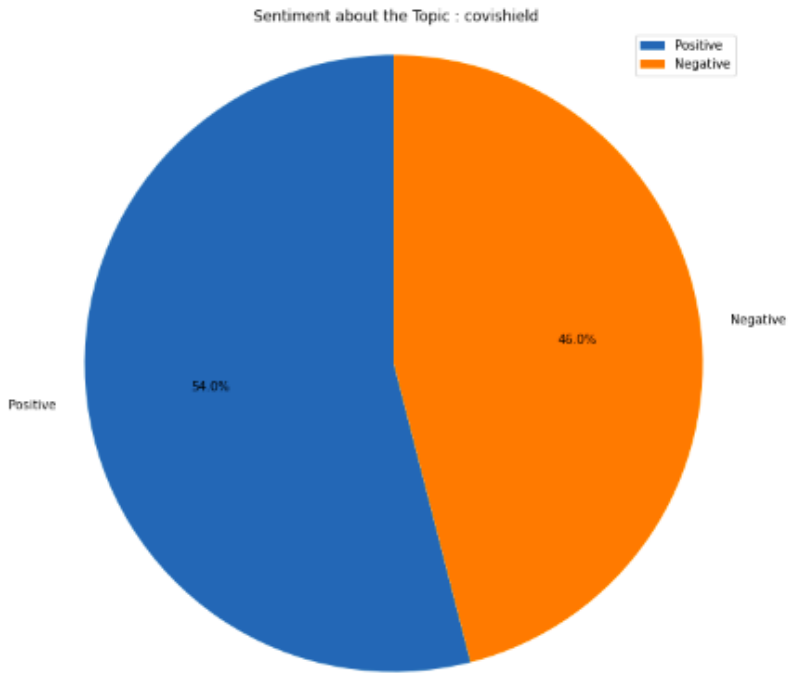


Figure 15: CovieShield Tweet analysis

1. Introduction

2. Dataset

3. Preprocessing

4. Models

5. Analyzing tweets

6. Conclusion and Take away message

Conclusion

Take away message

- Tweepy makes life easier if we have to scrap tweets to make dataset. We just have to give topic and language we want.
- Neglecting retweets is very easy through Tweepy. Our dataset will have uniqe data.
- Lemmatizing and pos tagging helps in preproceesing the dataset
- Word2Vec and FastText produced a good accuracy rate with Sentimental140 datatset

Reference

- https://www.kaggle.com/kazanova/sentiment140.
- https://www.kaggle.com/maxjon/complete-tweet-sentiment-extraction-data.
- https://machinelearningmastery.com/use-word-embedding-layers-deep-learning-keras/
- Agarwal, Apoorv, et al. "Sentiment analysis of twitter data."Proceedings of the Workshop on Language in Social Media (LSM 2011). 2011.
- https://docs.tweepy.org/en/latest/
- https://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/.YJJ3hKIzaco