Newspaper Articles using Sentiment Analysis

Roshan Srivastava (MBA Tech. EXTC J047)

Avanti Bhandarkar (BTech. EXTC C008)

Chetan Popli (BTech. EXTC C032)

Data Extracted

Sources: NDTV (Left Wing), Indian Express (Neutral), New Indian Express (Right Wing)

Data Extracted: Headline and article text

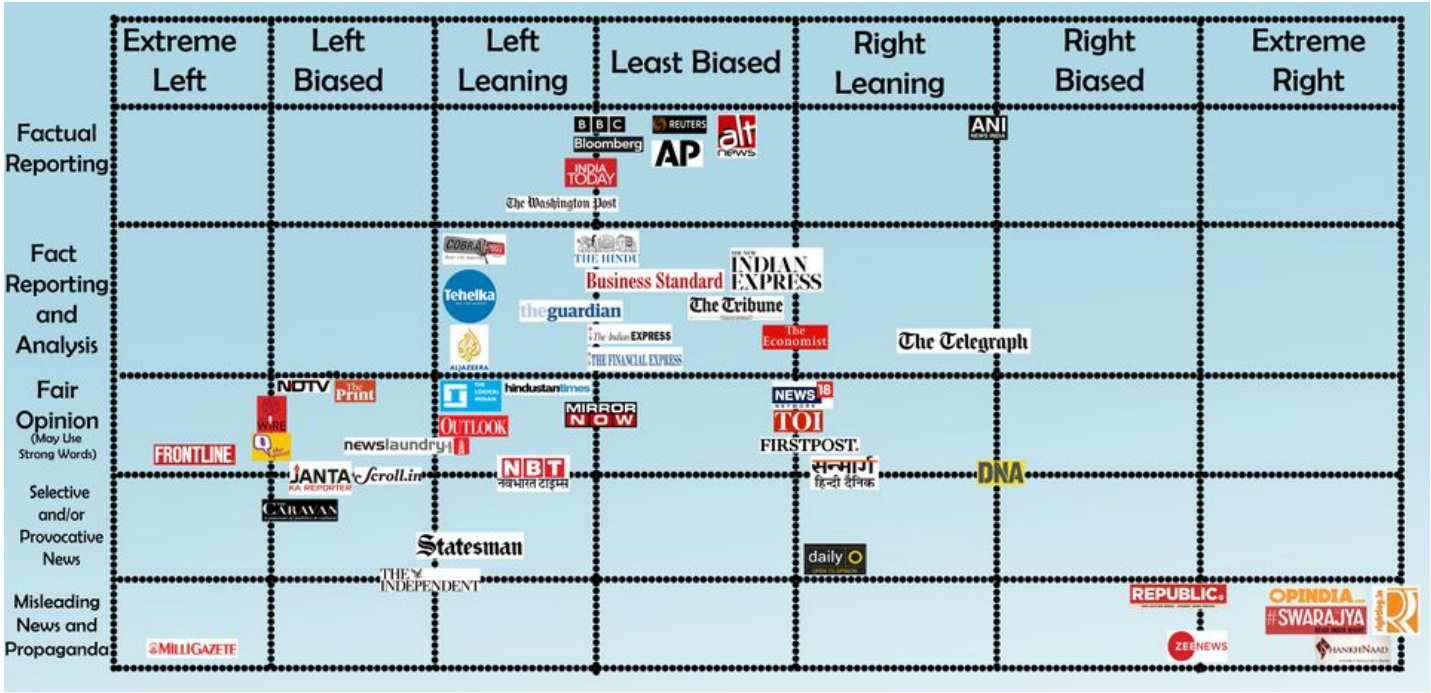


Figure 1: Chart of inherent biases and journalistic integrity for Indian media houses

```
# creating folders for data storage and clearing previous data
for folder in ["indianexpress","ndtv","newsexp"]:
    os.system(f"rm -rf {folder}")
    os.mkdir(folder)

# asking for a keyword to search for at news sites
keys = input("Enter a one-word keyword about the topic you want to search: ")
```

```
#NEWSEXP - right wing
#initializing selenium instance
global_url = "https://www.newindianexpress.com/topic?per_page="
headlines_text=[]
articles_text = []
options = Options()
options.headless = True
options.add_argument("--log-level=3")
driver = webdriver.Chrome(ChromeDriverManager().install(), chrome_options = options)
driver2 = webdriver.Chrome(ChromeDriverManager().install(), chrome_options = options)
print("\nExtracting data from Right wing source")

for i in range(4):
    driver.get(f"{global_url}{i}0&term={keys}&request=ALL&search=short")
    page = bs4.BeautifulSoup(driver.page_source, 'html.parser')
    headlines = page.find_all('div', {'class':'search-row_type'})
    for headline in headlines:
        if headline.h4.text not in headlines_text:
            headlines_text.append(headline.h4.a.text)
            page= driver2.get(headline.h4.a['href'])
            page = bs4.BeautifulSoup(driver2.page_source,'html.parser')
            article = page.find('div',{'class':'articlestorycontent'}).find_all('p')
            articles_text.append(" ".join(i.text for i in article))

for i in range(min(50,len(headlines_text))):
    with open(f"newsexp/newsexp_{i}.txt",'w') as file:
        file.write(headlines_text[i])
        file.write("\n")
        file.write(articles_text[i])
```

```python
#INDIANEXPRESS Neutral
print("\nExtracting data from Neutral source")
global_url = "https://indianexpress.com/page/"
headlines_text=[]
articles_text = []

for _ in range(1,3):
    driver.get(global_url + f"{i}/?s=" + f'+{keys}')
    page = bs4.BeautifulSoup(driver.page_source, 'html.parser')
    headlines = page.find_all('div', {'class':'details'})
    for headline in headlines:
        if headline.h3.a.text not in headlines_text:
            headlines_text.append(headline.h3.a.text)
            page = requests.get(headline.h3.a["href"])
            page = bs4.BeautifulSoup(page.text,'html.parser')
            article = page.find_all('p')
            articles_text.append(" ".join(i.text for i in article))

for i in range(len(headlines_text)):
    with open(f"indianexpress/indianexpress_{i}.txt",'w') as file:
        file.write(headlines_text[i])
        file.write("\n")
        file.write(articles_text[i])
```

```python
#NDTV left wing
print("\nExtracting data from Left wing source")
global_url = "https://www.ndtv.com/search?searchtext="
headlines_text=[]
articles_text = []
driver.get(global_url + keys)

for _ in range(2):
    driver.execute_script("allloadNews();")
    time.sleep(1)

page = bs4.BeautifulSoup(driver.page_source, 'html.parser')
headlines = page.find_all('div', {'class':'src_itm-ttl'})
for headline in headlines:
    if headline.a['title'] not in headlines_text:
        headlines_text.append(headline.a['title'])
        page = requests.get(headline.a["href"])
        page = bs4.BeautifulSoup(page.text,'html.parser')
        article = page.find_all('div','sp-cn ins_storybody')
        articles_text.append(" ".join(i.text for i in article))

for i in range(len(headlines_text)):
    with open(f"ndtv/ndtv_{i}.txt",'w') as file:
        file.write(headlines_text[i])
        file.write("\n")
        file.write(articles_text[i])

driver.close()
```

Dataset Generation

The scraped data is converted into a Pandas dataframe with the following columns:

- Headline: first line of the extracted data
- Article_text: remaining lines
- Source: extracted from .txt file name
- Bias: based on inherent bias of the source
- Cleaned_text: stores result of data preprocessing

```
df = pd.DataFrame(columns=['headline', 'article_text', 'source','bias',
'cleaned_text', 'compound', 'positive', 'neutral','negative'])

headline = []
article_text = []
source = []
bias = []
names = ['ndtv','indianexpress','newsexp']
```

```
for j in names:
  for i in range(30): #add range here based on number of scraped articles
    try:
      with open(f'{j}/{j}_{i}.txt', 'r') as f:
        # read first line as headline
        head = f.readline()
        head = head.strip()
        headline.append(head)

        # read all other lines as the article body, join to convert list to string
        lines = f.readlines()
        lines = ''.join(lines[0::1])
        article_text.append(lines)

        #append source and bias to the list
        source.append(j)

        if str(j) == 'ndtv':
          bias.append('L')
        elif str(j) == 'indianexpress':
          bias.append('N')
        elif str(j) == 'newsexp':
          bias.append('R')
    except:
      continue

# add all lists to the dataframe
df["headline"] = headline
df['article_text'] = article_text
df['source'] = source
df['bias'] = bias
```

Data Preprocessing and Cleaning

1. Converting string from mixed case to entirely lower case
2. Replacing all numbers, punctuations, emojis and whitespaces with a blank space
3. Removing all the stopwords as per NLTK's ENGLISH_STOP_WORDS corpus
4. Lemmatizing the data using the WordNet Lemmatizer model

```
# function for cleaning the data, remove tweet text, ads etc.
def clean(text):
    text = str(text)
    text = text.lower()
    text = re.sub('[^a-zA-Z]', ' ', text)
    text = re.sub(r"\s+[a-zA-Z]\s+", ' ', text)
    text = re.sub(r'\s+', ' ', text)
    stops = ENGLISH_STOP_WORDS
    text = [w for w in text.split() if not w in stops]
    lemmatizer = WordNetLemmatizer()
    text = [lemmatizer.lemmatize(word) for word in text]
    text = [ word for word in text if len(word) > 1 ]
    text = " ".join(text)
    return text

df_formed['cleaned_text'] = df_formed['article_text'].apply(clean)
```

Raw Data - 573 words

Home Minister Amit Shah discussed the security situation in Kashmir at a high-level meeting.Srinagar: Delimitation in Jammu and Kashmir will be followed by elections and restoration of statehood, Union Home Minister Amit Shah said on Saturday as he visited the Valley for the first time since Article 370 was revoked two years ago."Why should we stop delimitation? Nothing is going to stop. After delimitation, there will be elections and then restoration of statehood," Mr Shah said in an address to members of youth clubs in Srinagar.Earlier this year, Prime Minister Narendra Modi and Home Minister Shah met representatives of Kashmir's political parties in Delhi. Following the meeting, both Mr Modi and Mr Shah had stressed that delimitation should be completed at the earliest so that steps to restore statehood can be taken.The Union Home Minister's visit to the Valley comes at a time when the centre faces a major security challenge in the wake of targeted civilian killings that triggered an exodus of migrant labourers and Kashmiri Pandits. The visit also comes amid the ongoing anti-terror operation in Poonch that has claimed the lives of nine soldiers, including two officers.Earlier on Saturday, Mr Shah discussed the security situation at a high-level meeting.In his address to members of youth clubs, Mr Shah said the revocation of Article 370 is "irreversible".Defending the communication blockade and curfew following the August 5, 2019 move to revoke Article 370, Mr Shah said they were a "bitter pill" meant to save lives.Kashmir was under curfew for months and witnessed the world's longest internet shutdown after the revocation of Article 370. "There was a lot of criticism about why there is curfew, why there is internet shutdown. I will answer. First I want to ask a question. For 70 years, three families ruled here. Why were 40,000 people killed in Kashmir? Do you have an answer?""During that time [when Article 370 was revoked], they tried to incite people. A conspiracy was hatched, some foreign powers were part of it. How many fathers would have shouldered coffins of their young sons if we had not imposed a curfew. Who was saved by imposing the curfew? The youth of Kashmir was saved."The Home Minister said the abrogation of special status changed Kashmir's narrative from terrorism to development."Two years ago, news from Kashmir was about terrorism and stone-pelting. Today it's development, education, skill development, youth engagement," he said.Mr Shah said he is visiting Kashmir to "forge a friendship with the youth of Kashmir". "Join hands with Modi ji and the Government of India and become partners in the journey to take Kashmir forward," he said, adding that the youth in the Valley must take advantage of various opportunities being created by the administration for their progress."Make democracy stronger here, let the youth respond to the elements who try to make people go astray," he said. Extensive security arrangements have been made in view of the Home Minister's three-day visit.Instead prior to his visit 700 civilians were detained, booked under PSA & many shifted to jails outside Kashmir. Such oppressive steps further vitiate an already tense atmosphere. 'Normalcy acrobatics' are in full swing while reality is denied & obfuscated.— Mehbooba Mufti (@MehboobaMufti) October 23, 2021PDP chief and former Chief Minister Mehbooba Mufti alleged that 700 civilians had been detained and charged under the Public Security Act ahead of the Home Minister's visit. "'Normalcy acrobatics' are in full swing while reality is denied & obfuscated," she
tweeted.

Processed Data - 323 words

home minister amit shah discussed security situation kashmir high level meeting srinagar delimitation jammu kashmir followed election restoration statehood union home minister amit shah said saturday visited valley time article revoked year ago stop delimitation going stop delimitation election restoration statehood mr shah said address member youth club srinagar earlier year prime minister narendra modi home minister shah met representative kashmir political party delhi following meeting mr modi mr shah stressed delimitation completed earliest step restore statehood taken union home minister visit valley come time centre face major security challenge wake targeted civilian killing triggered exodus migrant labourer kashmiri pandits visit come amid ongoing anti terror operation poonch claimed life soldier including officer earlier saturday mr shah discussed security situation high level meeting address member youth club mr shah said revocation article irreversible defending communication blockade curfew following august revoke article mr shah said bitter pill meant save life kashmir curfew month witnessed world longest internet shutdown revocation article lot criticism curfew internet shutdown answer want ask question year family ruled people killed kashmir answer time article revoked tried incite people conspiracy hatched foreign power father shouldered coffin young son imposed curfew saved imposing curfew youth kashmir saved home minister said abrogation special status changed kashmir narrative terrorism development year ago news kashmir terrorism stone pelting today development education skill development youth engagement said mr shah said visiting kashmir forge friendship youth kashmir join hand modi ji government india partner journey kashmir forward said adding youth valley advantage various opportunity created administration progress make democracy stronger let youth respond element try make people astray said extensive security arrangement view home minister day visit instead prior visit civilian detained booked psa shifted jail outside kashmir oppressive step vitiate tense atmosphere normalcy acrobatics swing reality denied obfuscated mehbooba mufti mehboobamufti october pdp chief chief minister mehbooba mufti alleged civilian detained charged public security act ahead home minister visit normalcy acrobatics swing reality denied obfuscated

Sentiment Analysis

Algorithm Used: VADER (Valence Aware Dictionary and sEntiment Reasoner)
Returns document-level polarity and degree of said polarity

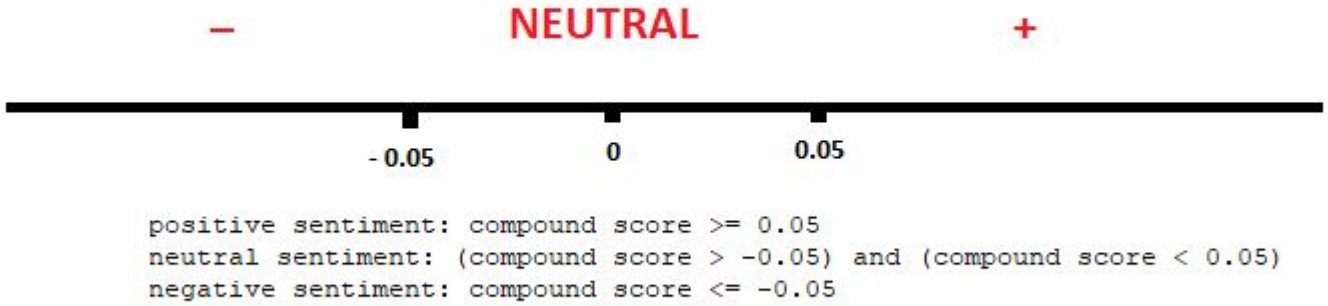


Figure 2: Polarity assignment thresholds for VADER

Dataframe columns for polarity added after sentiment analysis:

- Compound: normalised polarity score between -1 and +1
- Positive
- Neutral
- Negative

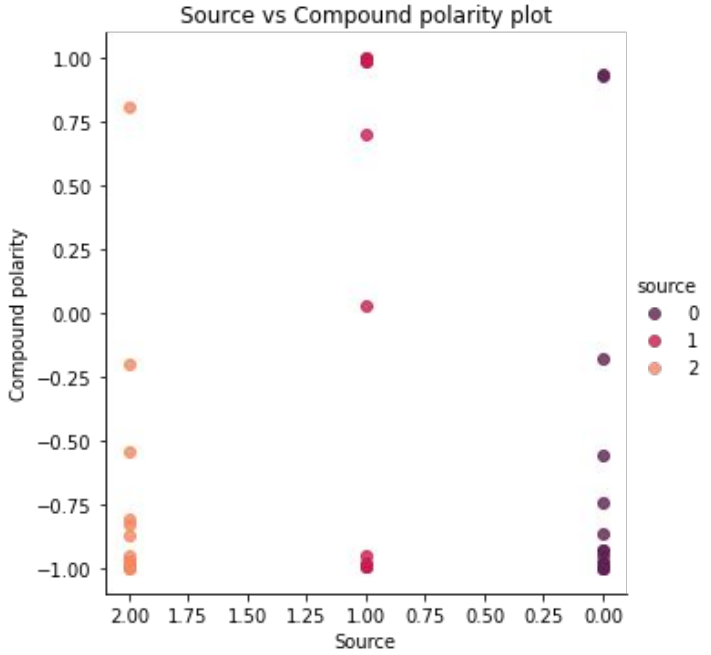
```
analyzer = SentimentIntensityAnalyzer()

# use VADER for finding compound, positive, negative, neutral sentiments
def compound(doc):
    analyzer= SentimentIntensityAnalyzer()
    score_compound = analyzer.polarity_scores(doc)['compound']
    return score_compound

def pos(doc):
    analyzer= SentimentIntensityAnalyzer()
    score_pos = analyzer.polarity_scores(doc)['pos']
    return score_pos

def neu(doc):
    analyzer= SentimentIntensityAnalyzer()
    score_neu = analyzer.polarity_scores(doc)['neu']
    return score_neu

def neg(doc):
    analyzer= SentimentIntensityAnalyzer()
    score_neg = analyzer.polarity_scores(doc)['neg']

    return score_neg
```

```
# adding to the dataframe

df_formed['compound'] = df_formed['cleaned_text'].apply(compound)
df_formed['positive'] = df_formed['cleaned_text'].apply(pos)
df_formed['neutral'] = df_formed['cleaned_text'].apply(neu)
df_formed['negative'] = df_formed['cleaned_text'].apply(neg)

# deleting redundant column
try:
    df_formed.drop(['Unnamed: 0'], axis = 1)
except:
    pass
```

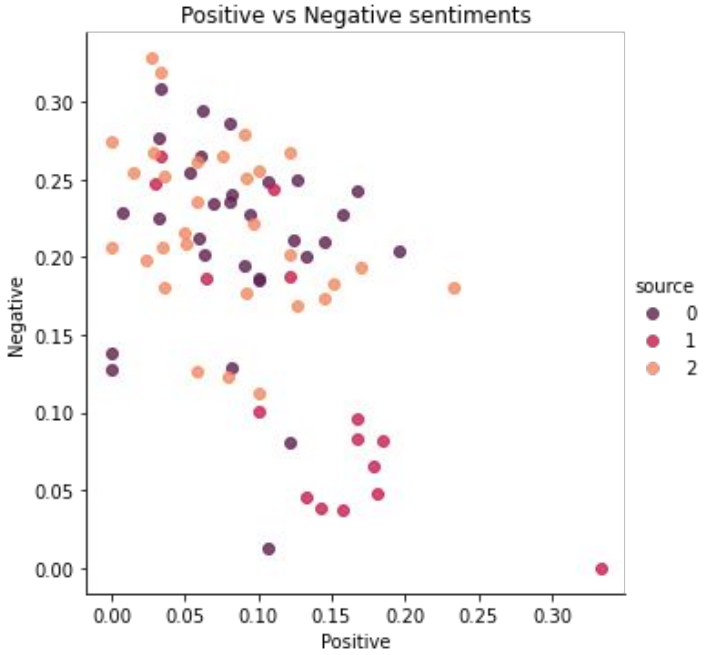
```
negative = df_formed[df_formed['compound']<0]

positive = df_formed[df_formed['compound']>0]

neutral = df_formed[df_formed['compound']==0]

df_formed.to_csv('nlp_sentiments.csv')

df = pd.read_csv('nlp_sentiments.csv')

cleanup_nums = {"source":     {'ndtv':0, 'indianexpress':1, 'newsexp':2},
                "bias":     {'L':0, 'N':1, 'R':2}}
df = df.replace(cleanup_nums)
```

Detection of Media Bias (using Anomaly Detection)

```
plt.figure(figsize=(30,30))
sns.lmplot("source", "compound",data=df, hue='source', fit_reg=False, palette = 'rocket');
plt.title("Source vs Compound polarity plot")
plt.ylabel("Compound polarity")
plt.xlabel("Source")
plt.gca().invert_xaxis()
plt.show()
```




```
plt.figure(figsize=(30,30))
sns.lmplot("positive", "negative",data=df, hue='source', fit_reg=False, palette = 'rocket');
plt.title("Positive vs Negative sentiments")
plt.ylabel("Negative")
plt.xlabel("Positive")
plt.show()
```



Method used: Isolation Forest, which uses decision trees, with partitions created by:

- randomly selecting a feature
- selecting a random split value between min and max values of the selected feature

Outliers are less frequent than regular observations + are different in terms of values
∴ they should ideally be identified close to the tree root with fewer splits.

Contamination = number of anomalous observations / number of normal observations

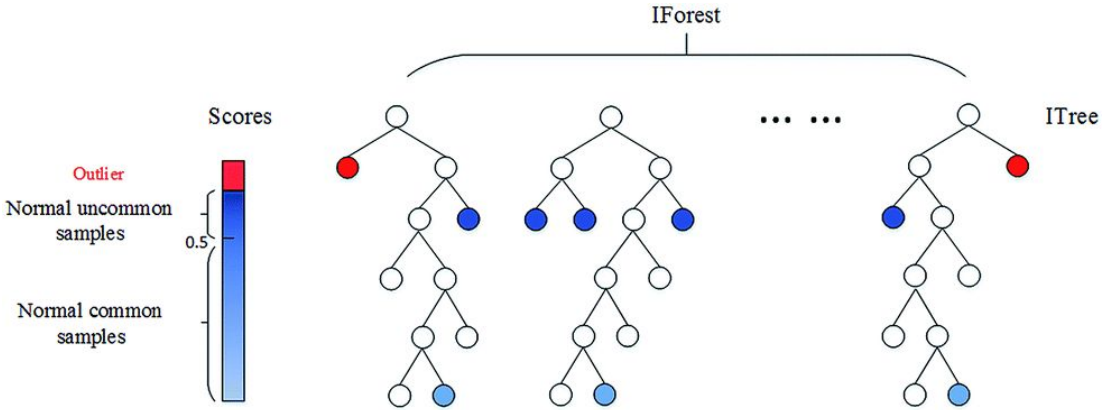
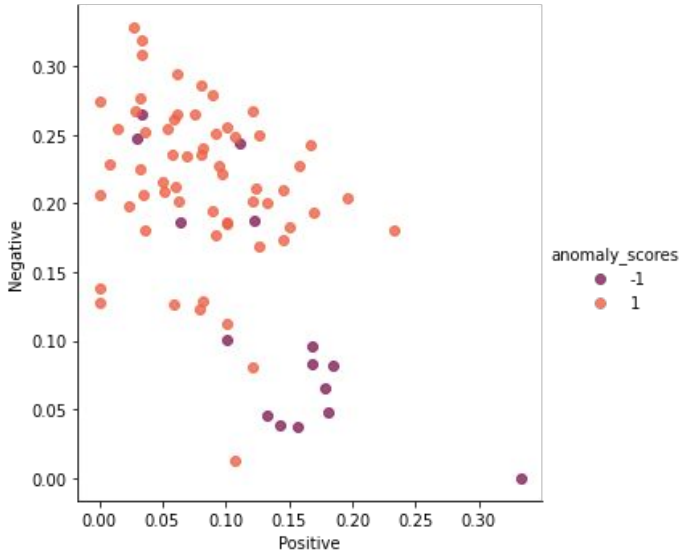


Figure 3: Isolation forest algorithm

```
anomoly_fraction = len(positive)/float(len(negative+neutral))

model = IsolationForest(random_state=1, contamination=anomoly_fraction)

cleanup_nums = {"source":     {'ndtv':0, 'indianexpress':1, 'newsexp':2},
                "bias":     {'L':0, 'N':1, 'R':2}}
df_formed = df_formed.replace(cleanup_nums)

model.fit(df_formed[['source']])

df_formed['scores'] = model.decision_function(df_formed[['source']])
df_formed['anomaly_scores'] = model.predict(df_formed[['source']])
anomaly_count = df_formed[df_formed['anomaly_scores']<=0]
anomaly_count = anomaly_count.shape[0]
print('\nNumber of anomalies in the dataset: ',anomaly_count)

X_train = df_formed[['compound']]
model.fit(X_train)
y_train = model.predict(X_train)
plt.figure(figsize=(5,5))
# sns.scatterplot(data=df, x="source", y="compound" , hue = y_train, palette = 'flare')
sns.lmplot("positive", "negative",data=df_formed,
hue='anomaly_scores', fit_reg=False, palette = 'rocket');
plt.xlabel('Positive');
plt.ylabel('Negative');
```



Advantages and Disadvantages

Advantages

- With our project, it is possible to detect outlying biases that media centers usually don't hold. This helps us to know whether a single reporter is trying to either defame the situation or make it look good.

- On a larger scale, this project can help determine news sources' type of bias.

Disadvantages

- Due to constraints of news sources, only one keyword is allowed. This makes narrowing down the event much difficult.

- Due to frequent activity, some website block the IP address for a limited time, rendering the system useless for that particular 'wing'. This has only been seen for "OPIndia" website.

Future Scope

- Accommodation for more than one keyword for the news extraction; this would help better narrow the event hence giving more accurate results.

- Addition of more news sources which would more inclusively represent all biases; this gives us much more data to work with

- Researching and finding better sentiment analysis methods than VADER; getting more parameters will help better clarify the anomalies.

- Performing topic modelling using methods such as Latent Dirichlet Allocation to correlate the topics found in the article with the article being an anomaly.

Thank You!