Great American Insurance Group

- **Team Members**
- Srivastava,Rajat
- Koganti, Sahit
- Garodia, Saket
- Guntaka, Praveen
- Li, Heng

Outline

- **Project Background**
  - Business Context
  - Motivation
  - Goal
- **Technology and Methodology**
- **Final Deliverable**
  - Model Result
  - Project Pipeline
- **Future Scope**

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Background – Business Context

An insurance company has **3 primary goals**

1. What products to insure?
2. What premium to keep?
3. How to assess risk?

A detailed **sentiment analysis** can aid in with these 3 goals.

Background – Motivation

- **Data:** Social media comes with a lot of real time data that can be analyzed to understand the sentiment.

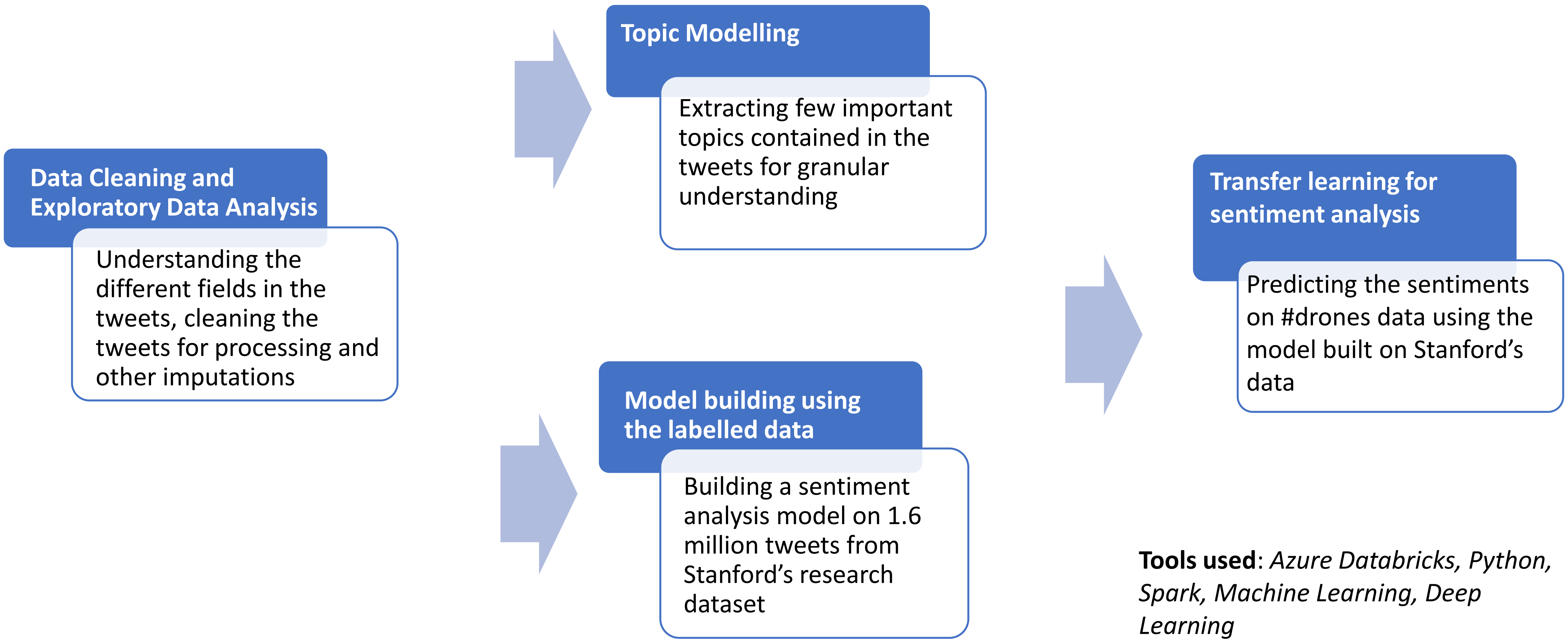
- **Tools:** Last 2 decades have been the most evergreen years in NLP research getting us access to more advanced tools.

Background — Goal

Goal: To build a reproducible pipeline that takes in tweets and details out various topics and sentiment associated with each tweet.

Scope and Deliverable: To show a demo with 500k tweets containing ‘#drones’ scraped from Twitter.

Project Methodology:



Data Cleaning and Exploratory Data Analysis

Understanding the different fields in the tweets, cleaning the tweets for processing and other imputations

Topic Modelling

Extracting few important topics contained in the tweets for granular understanding

Model building using the labelled data

Building a sentiment analysis model on 1.6 million tweets from Stanford's research dataset

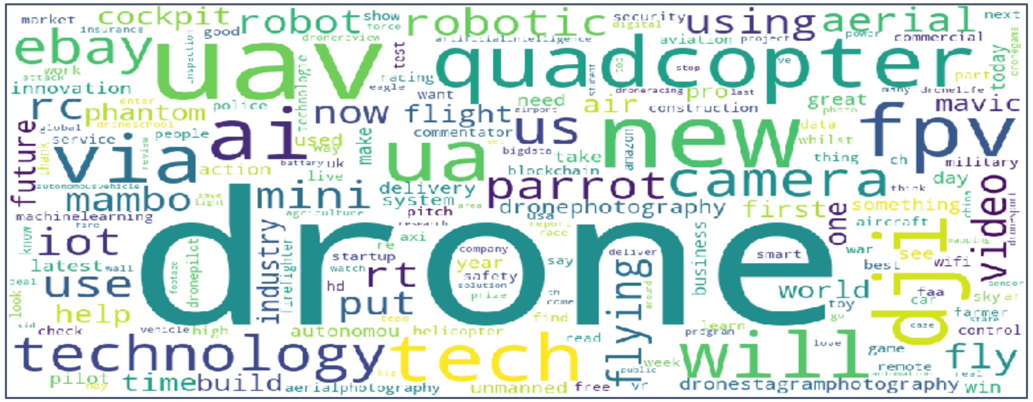
Transfer learning for sentiment analysis

Predicting the sentiments on #drones data using the model built on Stanford's data

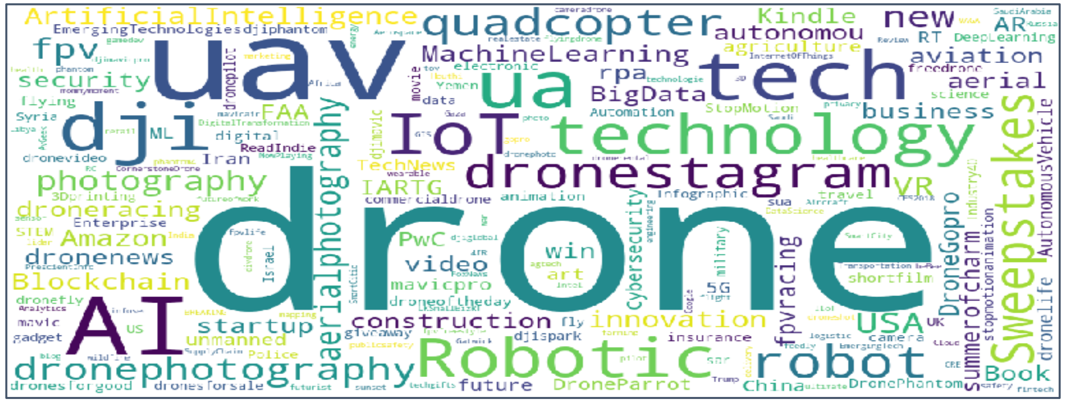
Tools used: *Azure Databricks, Python, Spark, Machine Learning, Deep Learning*

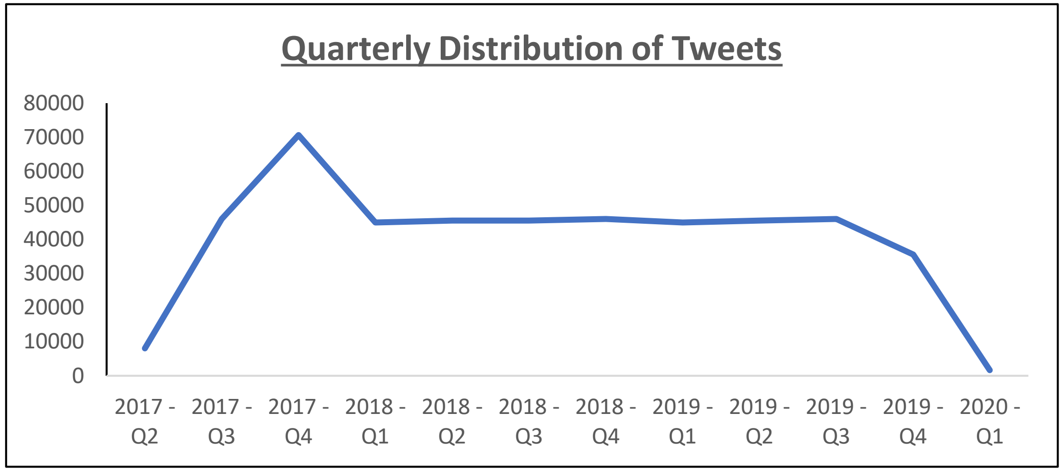
Exploratory Analysis Results:

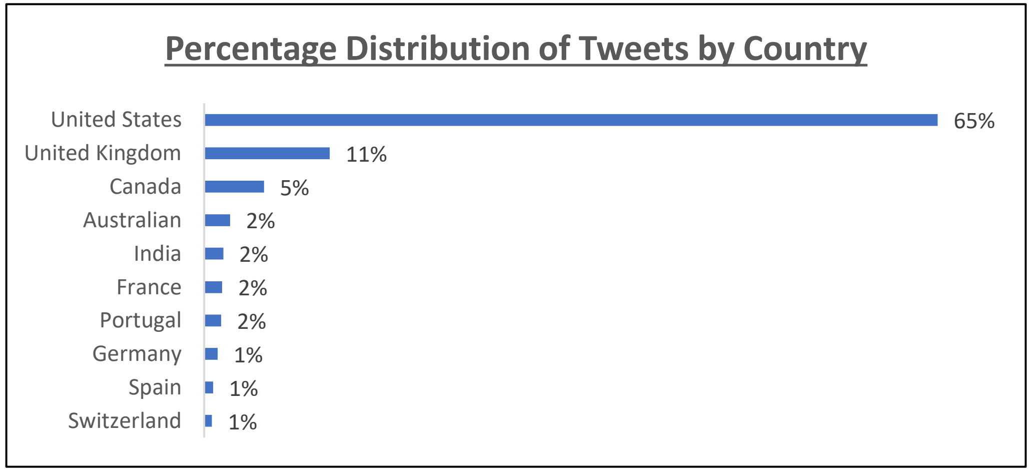
Tweet Text



Hashtag Text




Quarterly Distribution of Tweets

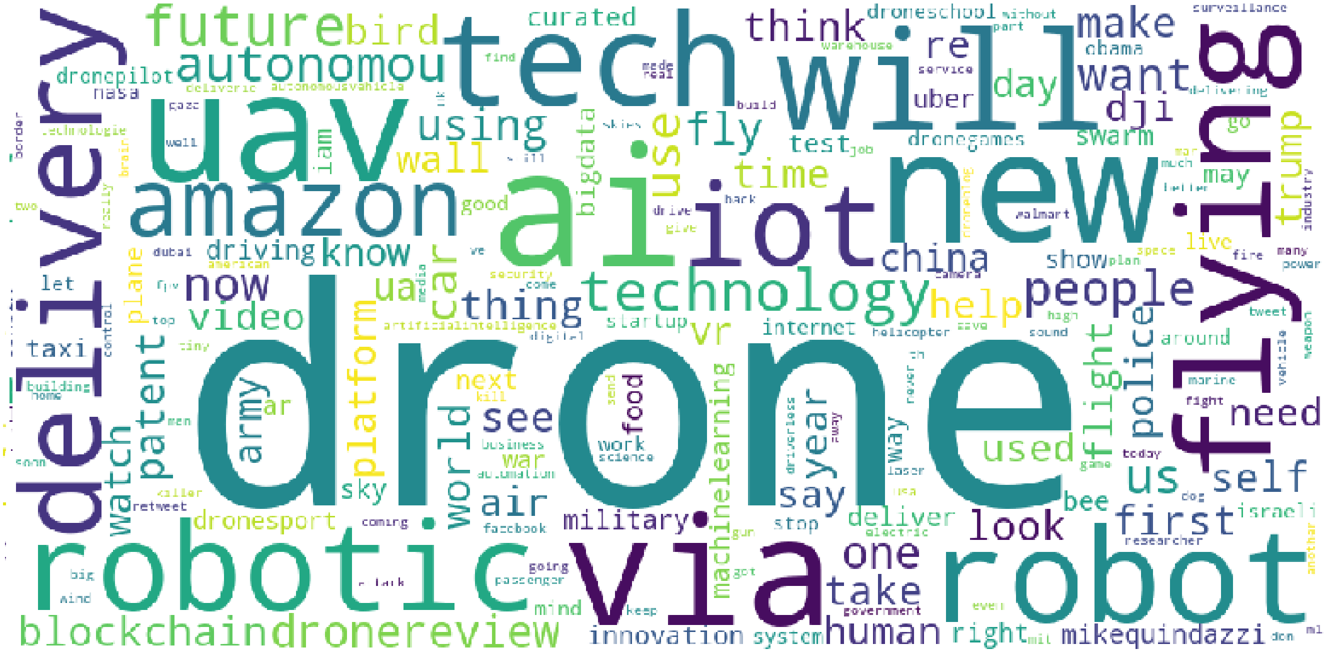

Percentage Distribution of Tweets by Country

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

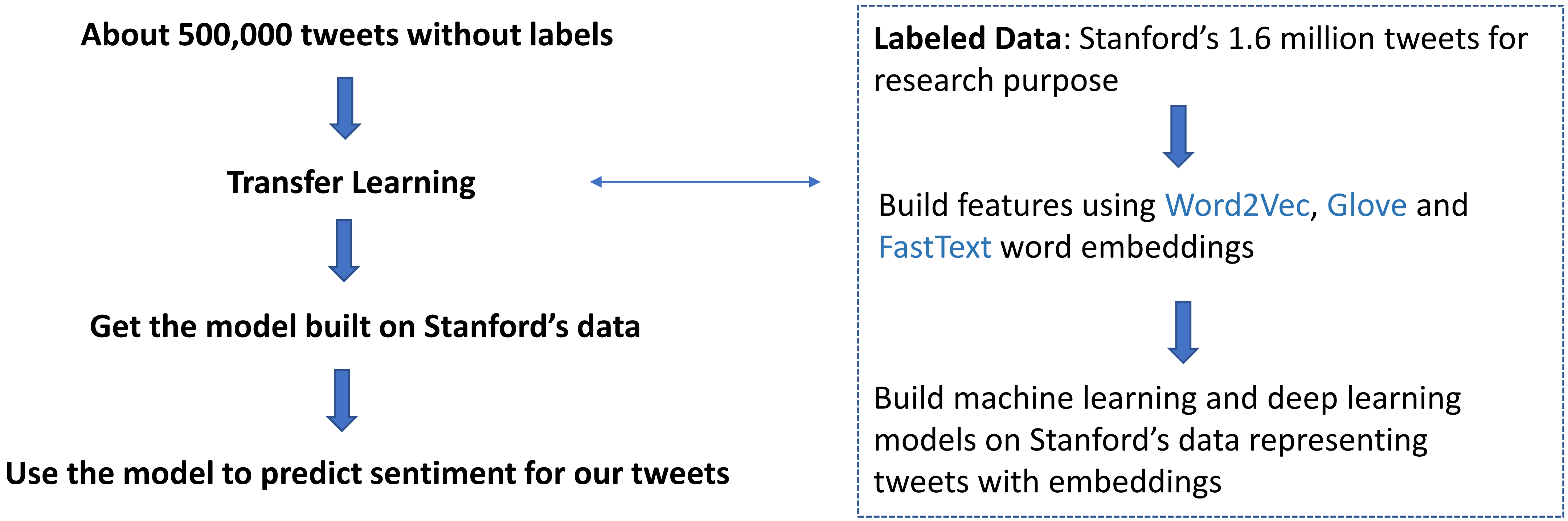
Topic Modelling Results(#Drones Data):

Here's the list of words belonging to different topics that we fed into the LDA algorithm:

- ❑ Industry Applications
- ❑ Drone Accessories
- ❑ Photography
- ❑ Geopolitical
- ❑ AI and Future



Sentiment Analysis(Methodology)



About 500,000 tweets without labels

Transfer Learning

Get the model built on Stanford's data

Use the model to predict sentiment for our tweets

Labeled Data: Stanford's 1.6 million tweets for research purpose

Build features using Word2Vec, Glove and FastText word embeddings

Build machine learning and deep learning models on Stanford's data representing tweets with embeddings

***Reference for the dataset**: http://help.sentiment140.com/for-students*

Sentiment Analysis Results:

	Model Description	Accuracy	F1-Score
1	Term Frequency – Inverse Document Frequency with Naïve Bayes	75%	0.76
2	GloVe with Gradient Boosting Classifier	76%	0.77
3	**CNN+LSTM Deep Neural Networks**	**83%**	**0.85**

Results on 1.6 million Stanford's tweets: The best metric was achieved by a deep learning hybrid model (Long Short-Term Memory(LSTM) in combination with Convolutional Neural Network (CNN))

CNN + LSTM (Metrics)

F1 score -- 0.8468

Precision – 0.8566

Recall – 0.8330

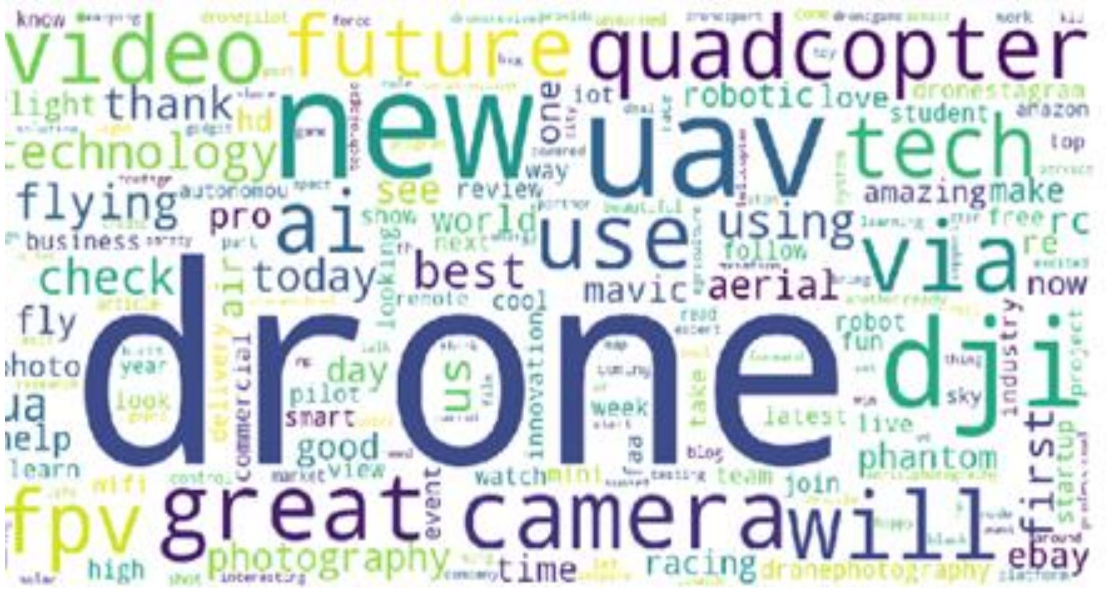
Precision is about 86% and is greater than recall which is 83.30%.

__Reference paper:__ https://www.academia.edu/35947062/Twitter_Sentiment_Analysis_using_combined_LSTM-CNN_Models

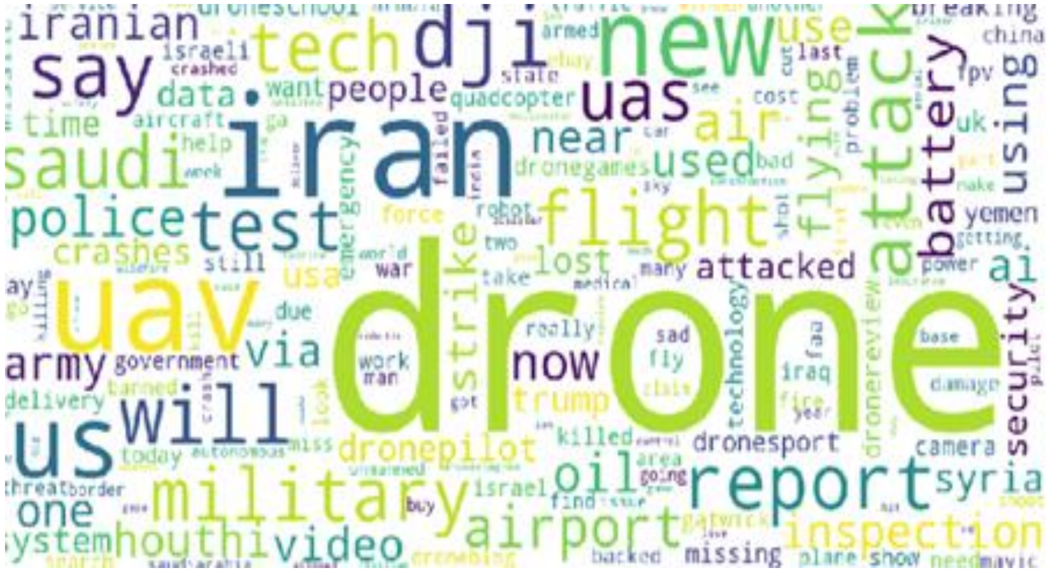
University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Sentiment Analysis Results:

Top words used in tweets with positive sentiment

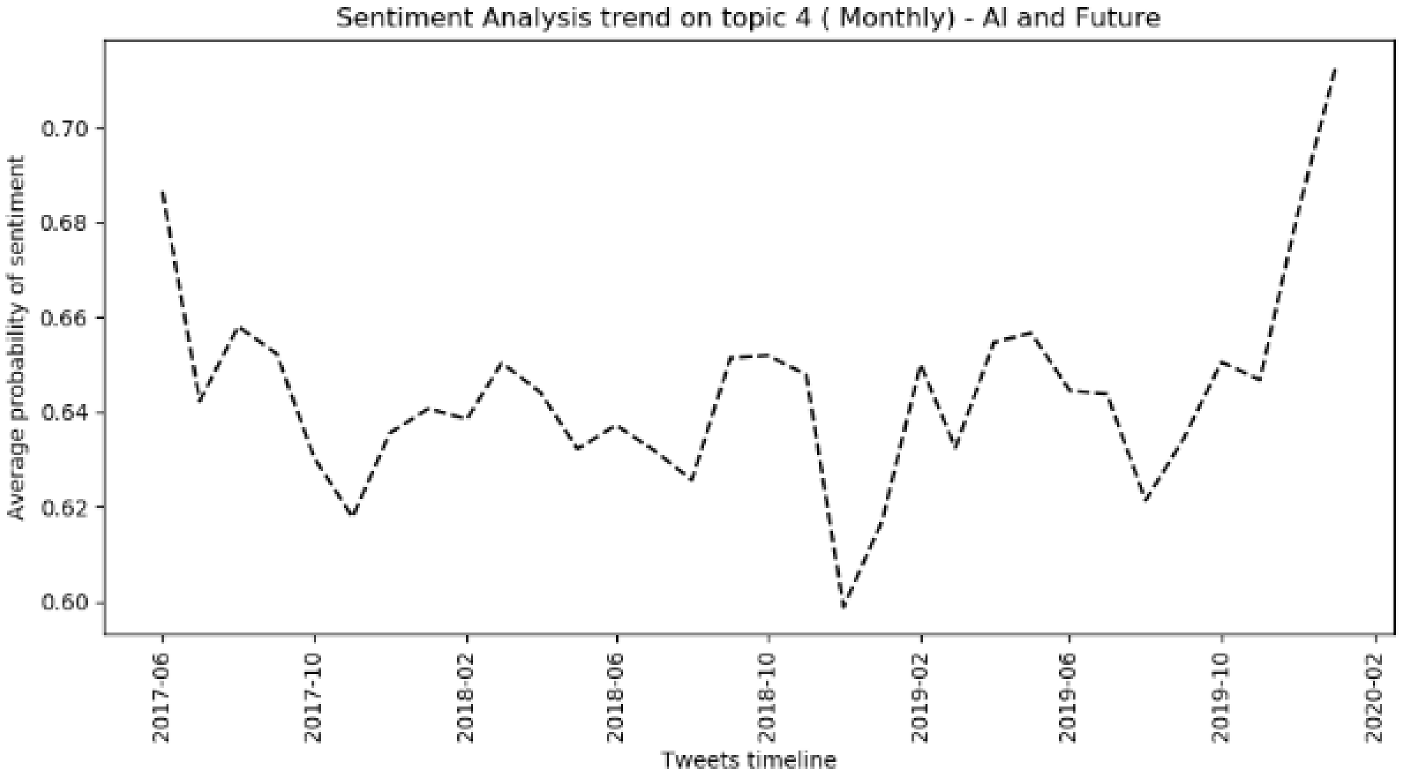


Top words used in tweets with negative sentiment



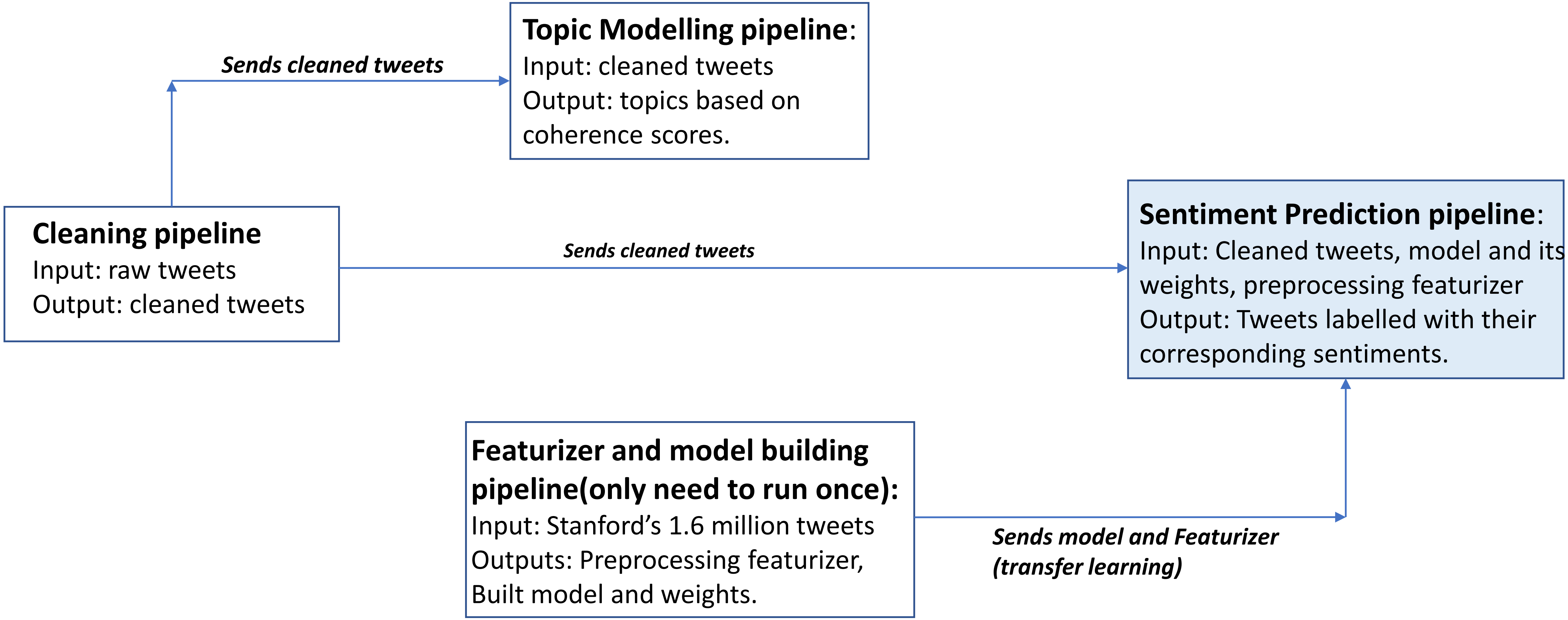
Sentiment Analysis Results:

- ❑ **Industry Applications**
- ❑ **Drone Accessories**
- ❑ **Photography**
- ❑ **Geopolitical**
- ❑ **AI and Future**



University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Pipeline(Overview):



Topic Modelling pipeline:
Input: cleaned tweets
Output: topics based on coherence scores.

Sends cleaned tweets

Cleaning pipeline
Input: raw tweets
Output: cleaned tweets

Sends cleaned tweets

Sentiment Prediction pipeline:
Input: Cleaned tweets, model and its weights, preprocessing featurizer
Output: Tweets labelled with their corresponding sentiments.

Featurizer and model building pipeline(only need to run once):
Input: Stanford’s 1.6 million tweets
Outputs: Preprocessing featurizer, Built model and weights.

Sends model and Featurizer (transfer learning)

University of CINCINNATI | CARL H. LINDNER COLLEGE OF BUSINESS

Future Scope:

- **Using Bert for Sentiment Analysis:** Bert can increase accuracy by 2-3%.

- **Using more training data for transfer learning**: Scaling 1.6 million tweets to 10-15 million tweets can certainly provide us more confidence in sentiment prediction.

- **Integration with Tableau:** Integrating Azure with Tableau can help in getting interesting visualization as part of the pipeline itself.

- **Named Entity Recognition**: Given the scope, named entity recognition can be performed which can help in useful analysis of such topics as important people, places and organizations.