Analyzing Customer Support of top companies on Twitter

- **Abstract**

The goal of this project was to aid innovation in natural language understanding and conversational models by using classification model. In this project, the study will focus on modern customer support practices on one of the most popular social media platforms 'Twitter' for Over 3 million tweets and replies from the biggest brands and companies its impact. This will give a wide view to enhance my job as customer service representative in a mall which customer support on twitter is new filed to my company.

- **Design**

The Customer Support on Twitter dataset originates from the "https://www.kaggle.com/thoughtvector/customer-support-on-twitter" offers a large corpus of modern English (mostly) conversations between consumers and customer support agents on Twitter, and has three important advantages over other conversational text datasets:
Focused - Consumers contact customer support to have a specific problem solved.
Natural - Consumers in this dataset come from a much broader segment than other corpuses.
Succinct - Twitter's brevity causes more natural responses from support agents (rather than scripted), and to-the-point descriptions of problems and solutions. Also, it's convenient in allowing for a relatively low message limit size.

- **Data**

This dataset contains 2811774 row, where each row is a tweet and there is 7 columns. Every conversation included has at least one request from a consumer and at least one response from a company. Which user IDs are company user IDs can be calculated using the inbound field which is the target. Classifying tweets accurately via machine learning models would enable the companies to improve customer support.

- **Algorithms**

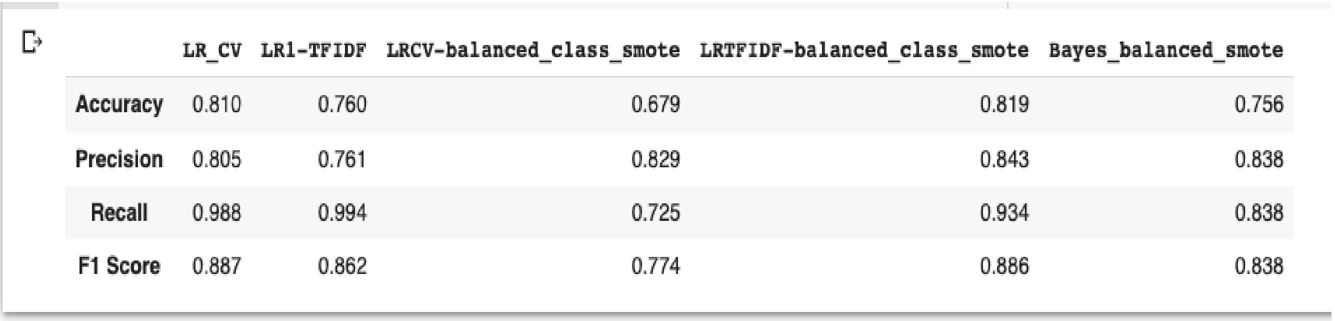
1. Various text preprocessing / cleaning steps for tweets.
2. Modify the dataframe to get query - response pairs in every row.
3. Making sure the dataframe contains only the needed columns.
4. Statistical Insight mean, max, min.
5. Adding new column the difference to the customer tweets and companies respond.
6. Doing a sentiment analyzing of subjectivity and polarity.

- **Models**

Logistic regression and naive bayes classifiers were used before settling on logistic regression as the model with strongest.

- **Model Evaluation and Selection**

I chose Logistic regression with TFIDF and balanced with SMOTE.



	LR_CV	LR1-TFIDF	LRCV-balanced_class_smote	LRTFIDF-balanced_class_smote	Bayes_balanced_smote
Accuracy	0.810	0.760	0.679	0.819	0.756
Precision	0.805	0.761	0.829	0.843	0.838
Recall	0.988	0.994	0.725	0.934	0.838
F1 Score	0.887	0.862	0.774	0.886	0.838

- **Tools**

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Visime for interactive visualizations

- Communication

Through presentation on zoom.