ML PROJECT

Presented by:

ARYAN GUPTA-20UCS034

ARYAN ADITYA-20UCS031

RESARCH PAPER TOPIC

TURKISH SENTIMENT ANALYSIS FOR OPEN AND DISTANCE EDUCATION SYSTEMS

INTRODUCTION

- Sentiment analysis (SA)-primary purpose is to analyze people’s sentiments or opinions toward entities such as topics, events, individuals, issues, services, products, organizations, and their attributes.

- Open and distance education (ODE) represents approaches that focus on opening access to education and training provision, freeing learners from time and place constraints, and offering flexible learning opportunities to individuals and groups of learners.

RELEVANCE OF SENTIMENT ANALYSIS

- in several fields like healthcare, political events, finance, marketing, and education
- analyze consumer behavior
- predict revenue
- institutional decision/policy making
- intelligent information/learning systems
- improvements via feedback

Existing System

Opinion Surveys , Blogs and Eductational portals

very tedious, slow, and time-consuming

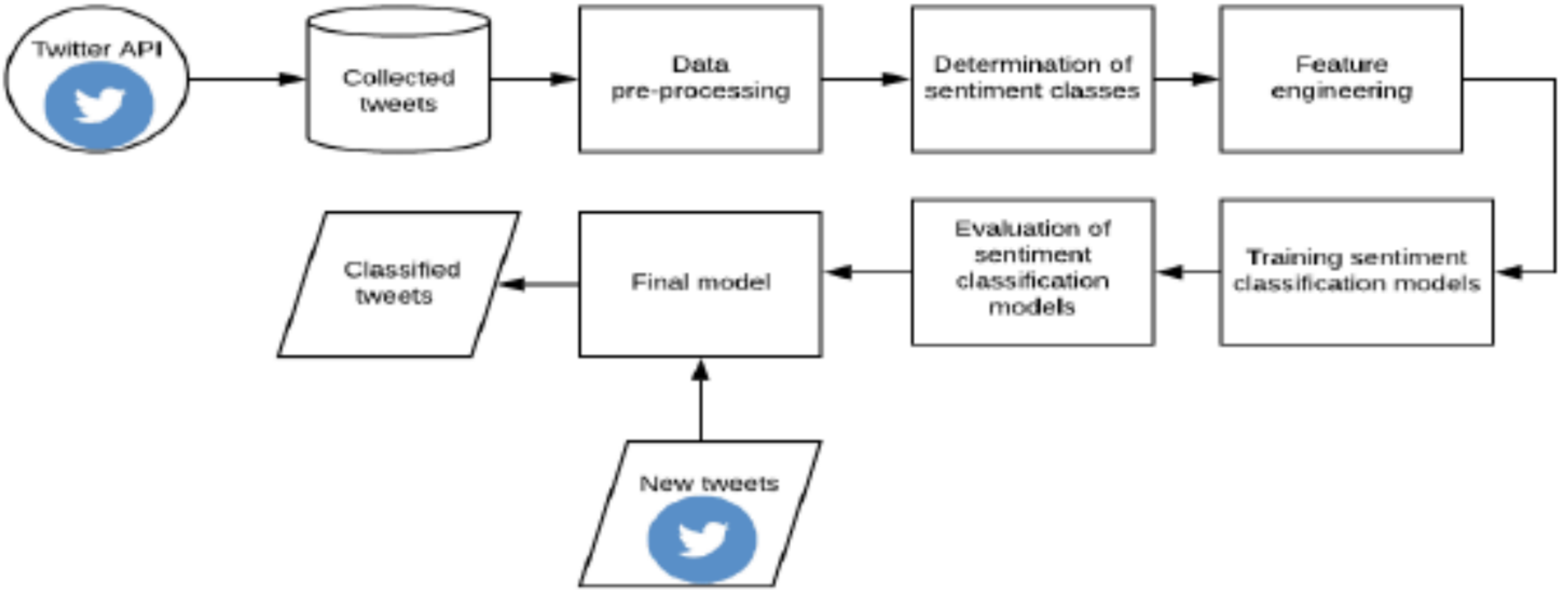
MACHINE LEARNIG APPROACH

- One of the main goals of SA is to exercise “Sentiment Polarity Classification” which is used to obtain the semantic polarity (positive, negative, or neutral) of a text.

- There are three main approaches for this goal in SA; Lexicon-based Approach, Machine-learning-based Approach, and Hybrid Approach.

- Machine learning-based methods are chosen in our study because there is no comprehensive polarity lexicon in Turkish, nor is there a narrow polarity lexicon containing terms related to ODE.

FLOW CHART


Twitter API
Collected tweets
Data pre-processing
Determination of sentiment classes
Feature engineering
Training sentiment classification models
Evaluation of sentiment classification models
Final model
Classified tweets
New tweets

METHODOLOGY

Data Pre-processing

- Data pre-processing involves four steps in our study; cleaning, normalization, tokenization, and stop word removal.
- In the cleaning step; duplicated tweets, unrelated links, URLs, advertisements, and news were removed from the dataset using regular expression.
- In the normalization step, we eliminated punctuations in the dataset and converted the dataset to a lower case.
- In the tokenization step, tweets were split into words as a token by whitespaces.
- In the stop words removing step, we removed the stop words, which in the stop words list and the hashtags used in the data collection step from the tweets.

Determination of Sentiment Classes

- To determine a sentiment class of each tweet, the tweets in the dataset were manually labeled as positive, negative, or neutral.

Feature Engineering

- Bag of words (BoW), term frequency-inverse document frequency (TF-IDF), is used as vector space models in our study.
- (TF) of a particular term (t) is calculated as the number of times a term occurs in a document to the total number of words in the text . IDF is the log of the inverse probability of a term being in the text.
- The term frequency-inverse document frequency tf-idf(t,d) is calculated as

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t,d)$$

Sentiment Classification

- In this study, we used Support Vector Machine (SVM), Logistic Regression (LR), K-nearest neighbor (KNN), and Artificial Neural Network (ANN) to classify the tweets into positive, negative, or neutral sentiments.

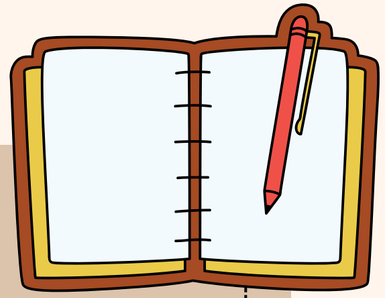
Parameter Setting

- Parameter setting is a crucial issue to improve the performance of classifiers. Grid search technique is used for finding the optimal combination of parameter values of these classifiers.

Evaluation Measures

- If the dataset is imbalanced, which means the class distribution is not uniform among the classes, F-score can be used as an evaluation measure. .

DISCUSSION



While SVM, KNN, and ANN gave better classification success with the BoW vector space model than the TF-IDF vector space model, LR gave better classification success with the TF-IDF vector space model than the BoW vector space.

FUTURE WORK in paper

1. We can diversify our data sources by collecting data from different social media platforms such as Facebook and Instagram.

2. we can develop a comprehensive Turkish lexicon for performing lexicon-based SA.

Conclusion

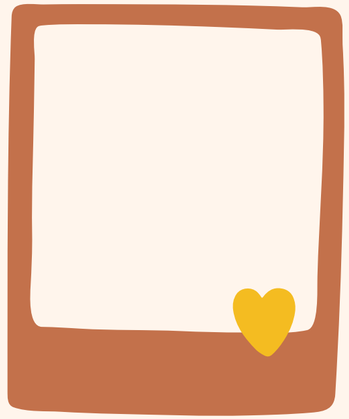
As a result of the presented SA models, ODE managers can develop strategies that will increase student satisfaction, based on the tweets whose feelings for the Open and Distance Education System are negative and positive. Based on the positive sentiments, strategies can be developed to reinforce ODE's positive transactions and ensure its sustainability. Besides, negative sentiments can be used in decision making, allowing an understanding of student dissatisfaction and how the institutions adopt actions to improve the quality of teaching and examination processes used.

LIMITATIONS IN RESEARCH

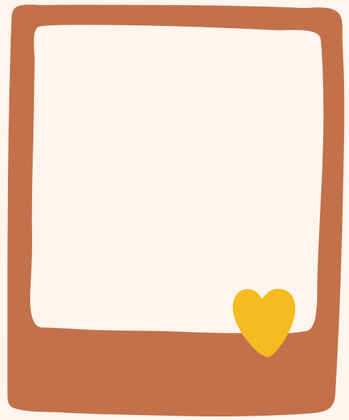
1. **Lack of Exploratory Data Analysis :** In our research paper, there was absence of analysis of quantitative aspect of the tweets like what was the overall topology of our collected data.
As an improvement, we performed Exploratory data analysis by deriving wordClouds, word counts and lenth counts etc.

2. **Imbalanced Data :** Our data is imbalanced with very little neutral and negative values compared to positive sentiments. t is indicated how the model is having difficulty in training positive classes because of imbalance. As an improvment, **We will balance our dataset before going into modelling process**

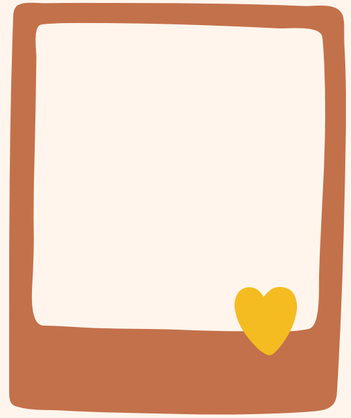
OUR FUTURE IMPROVEMENTS



Fine-grained sentiment analysis



Real-time sentiment analysis



Incorporating deep learning techniques

THANK YOU SO MUCH!