Sentiment Analysis and User Behavior Prediction in Social Networks

LIU MINGJIE

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

INTRODUCTION

3.1 Introduction

This research chapter unfolds the methodologies applied to studying user sentiment analysis as well as behavior prediction in social networks. The research design, data sources, processing methods, and analytical tools are outlined in detail to provide a structured and reproducible framework for the study. Each step in the data science project lifecycle is discussed to ensure clarity and rigor..

3.2 Research Design

The research design entails the use of two techniques, namely explorative and hypothetical methods. In these models, data is computed and adjusted for both qualitative and quantitative results, and the sentiment analysis and behavior prediction is run simultaneously. Figure 3.1 illustrates the framework for the plan as described in this paper.

The design aims to:

1.Identify the research problem and propose possible solutions.

2. Ensure a choice of appropriate methods and tools for data acquisition and processing.

3. Assess the outcome and its implication for formulating such decision proposals.

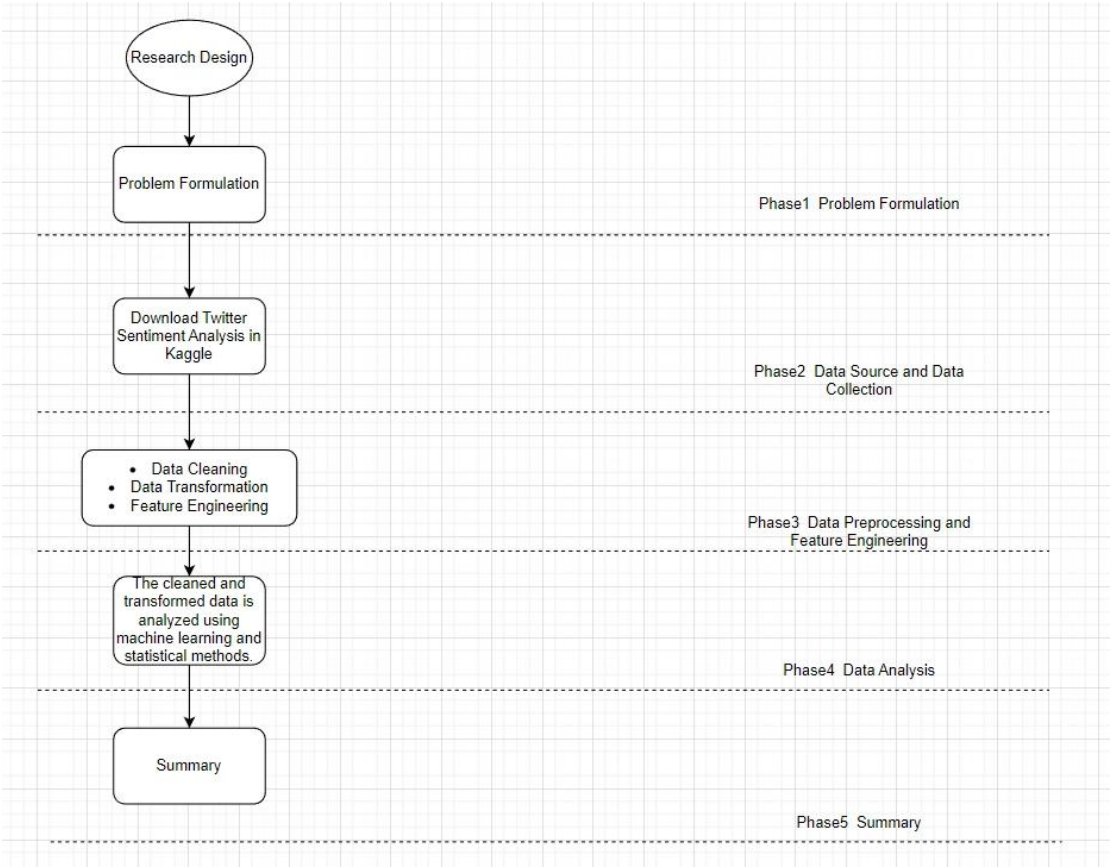

Figure 3.1: Research Framework of Proposal

3.3 Data Science Project Life Cycle

This stage of the study is based on the traditional data science project life cycle, which consists of posing the problem, data collection, preparation, analysis, as well as evaluation.

3.3.1 Problem Formulation

The book is to walk you through the most common problem in building interpretable and accurate models for sentiment analysis and user behavior forecasting based on social network data. Sub-questions include:

1. The relationship between image modality and affective polarity:
In the anote.xlsx dataset, do Tweets in the image modality show a different distribution of sentiment polarity than other modalities (e.g., plain text)?.
2. Relationship of Emotional Polarity to a Specific Brand:
Do Tweets from a particular brand (e.g., Amazon, Microsoft) show a specific trend of sentiment polarity in the twitter_validation.csv dataset?

3.3.2 Data Source and Data Collection

The data was found from Twitter and the website.

Many of these procedures include:

1. Data collection. Information in the dataset was obtained through automated scripts from Twitter. These data include tweet text, author information, and posting dates, among other details.
2. Data anonymization. Personal information of authors has been de-identified to protect their privacy. No specific user identification information is included.
3. Text cleaning. Text within the data has been processed and cleaned to remove sensitive or personal information while retaining information about topics and sentiments.

This twitter_company1 file contains Twitter data related to a particular company or brand, primarily used for sentiment analysis. The main contents and columns included are:

Twitter ID: A unique identifier for the tweet.

Content Theme: The theme of the content, which appears to be related to "Borderlands."

Tweet content: The specific text content of the tweet.

Modality: The media type of the tweet, here it is "text," indicating textual content.

Polarity: The sentiment polarity of the tweet, including "Positive" (positive sentiment).

The twitter_company dataset that has been collected is 74681 rows of data with 5 columns as shown in the Figure 3.1

[9]:
```
import pandas as pd

# Load the CSV file
csv_file_path = 'twitter_company1.csv'
csv_data = pd.read_csv(csv_file_path)
csv_data
```

[9]:

	Twitter ID	Content Theme	Tweet content	modality	polarity
0	2401	Borderlands	I am coming to the borders and I will kill you...	text	Positive
1	2401	Borderlands	im getting on borderlands and i will kill you ...	text	Positive
2	2401	Borderlands	im coming on borderlands and i will murder you...	text	Positive
3	2401	Borderlands	im getting on borderlands 2 and i will murder ...	text	Positive
4	2401	Borderlands	im getting into borderlands and i can murder y...	text	Positive
...	...	...	...	...	...
74676	9200	Nvidia	Just realized that the Windows partition of my...	text	Positive
74677	9200	Nvidia	Just realized that my Mac window partition is ...	text	Positive
74678	9200	Nvidia	Just realized the windows partition of my Mac ...	text	Positive
74679	9200	Nvidia	Just realized between the windows partition of...	text	Positive
74680	9200	Nvidia	Just like the windows partition of my Mac is l...	text	Positive

74681 rows × 5 columns

Figure 3.1 The twitter_company Dataset Preview

This anote file contains sentiment analysis of Twitter data that includes images and text. The main contents and columns included are:

Twitter ID: A unique identifier for the tweet.

Tweet content: The specific text content of the tweet.

Image: A link to the image attached to the tweet.

Label the user ID: Labeling categories are divided into two categories according to polarity and emotion. The polarity is marked as positive, neutral, and negative. Choose 1 from 3, and the mood is marked as happy Joy, sad Sad, afraid of fear, angry, surprised, disgusted, and confused Confused's 16 multiple choices.Both datasets involve sentiment analysis, but the CSV file focuses mainly on textual content, while the Excel file.

Modality: The media type of the tweet, shown as "image," indicating the tweet includes an image.

Polarity: The sentiment polarity of the tweet, including "negative" (negative sentiment), "positive" (positive sentiment), and "neutral" (neutral sentiment).

Emotion: Emotion, containing specific emotional tags such as "anger" (anger), "joy" (joy), "anyway" (anyway), and "confused" (confused).

This dataset provides sentiment analysis of tweets with images, helping to understand the emotional impact and expression of the tweet content and its accompanying images.The twitter_company dataset that has been collected is 9662 rows of data with 7 columns as shown in the Figure 3.2

```
*[14]: import pandas as pd
       # Load the CSV file
       csv_file_path = 'anote.csv'
       csv_data = pd.read_csv(csv_file_path, encoding='utf-8', encoding_errors='ignore')
       csv_data
```

[14]:

	Twitter ID	Tweet content	Image	Label the user ID	modality	polarity	emotion
0	1483637809449779200	kate and toby's smoker heating up to destroy t...	http://10.112.67.227:8666/img/download/1483637...	3	image	negative	["anger"]
1	1498163459749715973	@VIVIZ_staff All the best sinb, eunha, umji ...	http://10.112.67.227:8666/img/download/1498163...	3	image	positive	["joy"]
2	1499194611511791617	does your heart ever go :\n\n ...	http://10.112.67.227:8666/img/download/1499194...	3	image	neutral	["anyway"]
3	1499196380400787460	@MrNiceGuy513 @Seedalicious @Great_Katzby @POC...	http://10.112.67.227:8666/img/download/1499196...	3	image	neutral	["confused"]
4	1499212575128670209	"Dont stare at your dmen after every goal agai...	http://10.112.67.227:8666/img/download/1499212...	3	image	negative	["anger","surprise"]
...	...	...	...	...	...	...	...
9657	1506815587749613571	They already got my blood pressure up #Snowfal...	http://10.112.67.227:8666/img/download/1506815...	1	synthesis	negative	["fear"]
9658	1506815644934848513	#SnowfallFX This literally went from bad to wo...	http://10.112.67.227:8666/img/download/1506815...	1	synthesis	negative	["disgust","anyway"]
9659	1506815765319766018	They got out the cage and now they are in anot...	http://10.112.67.227:8666/img/download/1506815...	1	synthesis	negative	["anyway"]
9660	1506815793039921156	Hear me out - if Heather was to be honest and ...	http://10.112.67.227:8666/img/download/1506815...	1	synthesis	positive	["proud"]
9661	1506815851781050373	This tiger shit ... #SnowfallFX https://t.co/swl...	http://10.112.67.227:8666/img/download/1506815...	1	synthesis	negative	["disgust","anyway"]

9662 rows × 7 columns

Figure 3.2 Anote Dataset Preview

3.3.3 Data Pre-Processin

1 Data Cleaning:

Cleaning up the missing and tinted data.

Several data types like types, punctuation, and some symbols can introduce noise. Filters can help remove such noise.

Text Normalization, which includes the transformation of text into some format, e.g. Case Normalization, Lemmatization.

2 Data Transformation:

Word tokens and phrases can be used as a text.

Categorical variables are usually encoded with two possible methods (verbose description goes here). Word embeddings, such as Word2Vec or TF-IDF, can be used for this conversion.

3 Feature Engineering:

Extract of linguistic features that gives the idea about sentiment score as well as n-gram.

The second thing is the identification of features that depend on the behavior of users, e.g. interaction patterns and levels of activity.

The last thing is developing some domain-specific features because of the research context which we have been talking about.

3.4 Data Analysis

The data is pre-processed and analyzed by using statistical and machine learning techniques. Key steps include:

Model Training: The application of models for classification, regression, and sequence projections.

Model Evaluation: Considers as performance metrics accuracy, precision, recall, F1, Mean Squared.

Cross-Validation: The establishment of a train/test dataset splits helps in ensuring a generalizability of models.

Exploratory Data Analysis (EDA): Generating data plots that illustrate the way data is distributed, correlate between each other, as well as trends, with the purpose of accumulating knowledge.

3.5 Summary

his chapter explicitly explained in detail the research approach and methodology regarding the problem statement, the primary data collection, and preprocessing the data. It indicated the application of a clear process of getting ready the data for sentiment analysis and human user behavior prediction. The detailed data processing steps, consequently, provide a thorough and accurate investigation, the end product being new information and novel perspectives emerging from the subsequent chapters.