Chapter 12
Impact of Balancing Techniques for Imbalanced Class Distribution on Twitter Data for Emotion Analysis: A Case Study

Shivani Vasantbhai Vora
CGPIT, Uka Tarsadia University, Bardoli, India

Rupa G. Mehta
Sardar Vallabhbhai National Institute of Technology, Surat, India

Shreyas Kishorkumar Patel
Sardar Vallabhbhai National Institute of Technology, Surat, India

ABSTRACT

Continuously growing technology enhances creativity and simplifies humans' lives and offers the possibility to anticipate and satisfy their unmet needs. Understanding emotions is a crucial part of human behavior. Machines must deeply understand emotions to be able to predict human needs. Most tweets have sentiments of the user. It inherits the imbalanced class distribution. Most machine learning (ML) algorithms are likely to get biased towards the majority classes. The imbalanced distribution of classes gained extensive attention as it has produced many research challenges. It demands efficient approaches to handle the imbalanced data set. Strategies used for balancing the distribution of classes in the case study are handling redundant data, resampling training data, and data augmentation. Six methods related to these techniques have been examined in a case study. Upon conducting experiments on the Twitter dataset, it is seen that merging minority classes and shuffle sentence methods outperform other techniques.

DOI: 10.4018/978-1-7998-7371-6.ch012

INTRODUCTION AND MOTIVATION

Information technology is used in every field of human life and make human's life improved and more accessible. This tool became valued elements of life because it opened many doors to individuals. It firmly entrenched in human lives and facilitated their lives. Continuously growing technology strengthens individual creativity, makes our daily life more accessible, and gives us the facility to predict and cater to our needs. A deep understanding of human behavior is needed in machines and computers to understand our needs. The key part of human behavior is about perceiving and communicating emotions. It also motivates to take actions, influence the quality of decision making, and enhance the ability to empathize and communicate. Machines and computers must deeply understand emotions to anticipate human needs (Chatterjee A et al. (2019)). Emotion recognition and detection are closely related to sentiment analysis. Identification of sentiment intends to detect neutral, negative, or positive feelings from the content (Liu, B. (2012)).

In contrast, Emotion Analysis aims to identify and recognize feelings through text phrases, like joy, happiness, anger, disgust, fear, sadness, surprise, and many more (Picard R. W. (2000)). Recently, an identification of emotion has become a popular application of NLP. It has potential applications in Artificial intelligence (Damani S et al. 2018), Psychology (Druckman J. N. et al. 2008), Human-computer interaction (S. Brave et al.2009), Political science (Valentino N. A. et al. 2011) help in preventing suicide, or measuring the communal well-being (Van der Zanden R. et al. 2014), and Marketing (Bagozzi R. P. et al. 1999) etc.

WhatsApp, Facebook, and Twitter are prominent messaging platforms used by many online users to interact with each other. Statics given by (Statista, 2021) – "by the 3rd quarter of 2020, there are around 187 million daily active users of Twitter worldwide." In varied fields like researchers in marketing, analytics for political parties or social scientists look into twitter data in order to study human behavior in physical world. Tweets are rich sources of textual data containing the emotions of users. These data inherit the imbalanced emotion class distribution. In imbalanced dataset, data samples of one class are higher or lower than that of other group of classes. Figure 1 illustrates an imbalanced data. On encountering a imbalance class distribution problem in the training data, the results of classification task is influenced by majority class (Zhao C. et al. 2020).

Most machine learning classification algorithms are unable to manage imbalanced distribution of classes and are likely to get influenced by majority classes (Kothiya, Y. (2020, July 17)).

In the research literature, various approaches are proposed to cater to the imbalance class distribution issues in the data classification. These approaches are broadly categorized as algorithmic centered approaches and pre-processing methods or data level approaches.

Re-sampling techniques (Kotsiantis S. et al. 2006), reducing redundant data (Y.K. (2019, May 15)), and augmentation of text data are data-level approaches that are included as a solution to handle imbalance distribution of classes. The techniques are utilized to obtain an approximately equal count of samples in the classes. Assumptions created to favor the minority class and change the costs to get the balance classes, is the algorithmic-centered approach. (Kotsiantis S. et al. 2006).

In the machine learning (ML) community, the imbalanced class distribution gained extensive attention as it has produced many research challenges. It demands the experimental comparisons of approaches to take care of the imbalanced data set. A case study focuses on various data-level methods to deal with the imbalance distribution of emotion classes in Twitter data.