Chapter 11
Fake News and Imbalanced Data Perspective

Isha Y. Agarwal
Sardar Vallabhbhai National Institute of Technology, Surat, India

Dipti P. Rana
https://orcid.org/0000-0002-5058-1355
Sardar Vallabhbhai National Institute of Technology, Surat, India

ABSTRACT

Fake news has grabbed attention lately. In this chapter, the issue is tackled from the point of view of collection of quality data (i.e., instances of fake and real news articles on a balanced distribution of subjects). It is predicted that in the near future, fake news will supersede true news. In the media ecosystem this will create a natural imbalance of data. Due to the unbounded scale and imbalance existence of data, detection of fake news is challenging. The class imbalance problem in fake news is yet to be explored. The problem of imbalance exists as fake news instances increase in some cases more than real news. The goal of this chapter is to demonstrate the effect of class imbalance of real and fake news instances on detection using classification models. This work aims to assist researchers to better resolve the problem by illustrating the precise existence of the relationship between the imbalance and the resulting impact on the output of the classifier. In particular, the authors determine that data imbalance and accuracy are inversely proportional to each other.

INTRODUCTION

"Fake news" has gained a significant research attention worldwide since 2017 (Allcott & Gentzkow, 2017). Although fake news is in the public realization, it is important to note that digitally generated text and content can reach far then expected more than the true news. News or media articles online are less authenticated than that of original news media sources such as magazines or newspapers (Xiao, 2018; Zhang et al., 2019). Massive data is getting produced, either manually or by AI, to have a political or financial gains (Vosoughi et al., 2018; Berinsky, 2015). Fake news is those news stories that claim to

DOI: 10.4018/978-1-7998-7371-6.ch011

be accurate, but contain factual misrepresentations with the intention of arousing emotions, attracting viewership, or deceiving (Mihaylov et al., 2018).

By 2022, more fake information than real information would be processed by most people in developed nations also (Gartner, 2018). This will create imbalance in data in online media ecosystem making people exposed to more of Fake News then Real News.

In recent years, there has been significant contribution in the fake news detection area. Major contributions in this field focus on the detection methodology i.e. feature engineering, model construction and so on. However, one area is yet to be explored in the perspective of fake news i.e. nature of the data at disposal for classification. A severe class disparity among these groups causes one of the major problems faced by current machine learning models used for fake news detection. Therefore, most models struggle to distinguish instances that fall into minority groups correctly. The detection of fake news is difficult because of the infinite magnitude and imbalance nature of news data information. The class imbalance problem in data mining is the biggest problem. The problem of imbalance exists where one of the two groups has more samples than other classes. The algorithm mostly focuses more on classifying major samples while ignoring minority samples or misclassifying them. Minority samples are those that occur occasionally, but are very important. There are various ways to deal with imbalance. In this paper we present evaluation of data oriented methods for imbalance handling i.e. over sampling and under sampling. The goal is to recognize the shortcomings of current approaches to class imbalance; more precisely, the techniques of machine learning.

Many contributions have been dedicated to the issue of classification of imbalance data. In an extensive bulk of literature (Zhi-Hua Zhou & Xu-Ying Liu, 2006; Kaur & Gosain, 2018; Wasikowski & Chen, 2010), many solutions have been proposed that are based on machine learning and data mining algorithms. Class imbalance, however, has remained an unresolved problem (Richhariya & K Singh, 2014; Shuo Wang & Xin Yao, 2012).

In this research, comparative analysis of performance of classifier for imbalance data using is made. Focus is more specifically on data oriented methods with various machine learning algorithms. A research on state-of-the-art technologies currently used in applications was conducted to resolve the issue of imbalance classification. The area for imbalance data chosen here for experimentation is Fake News. Contribution of the work in this book chapter are highlighted in the next section.

Contribution

In this research, comparative analysis of approaches to deal with data imbalance is presented. Precisely the data oriented approaches. The list of unique contributions made by this research work is given below:

- Study the impact of data imbalance on Fake News
- Analysis of existing datasets for imbalance
- Comparative analysis of various techniques of dealing with imbalance
- Analysis and validation of oversampling approach for fake news

In the following background section, summary of research literature related to publicly available datasets for fake news research and methodologies existing for imbalanced classification is reviewed. The proceeding section has detailed description of proposed framework. The next section explains in