Characteristics of viral news on Mashable

Albert Lee, Yueh-Chang Kuo, Conor Walsh, Jaeger Wells
MSDS 422
Winter 2024
Group Project MidPoint Presentation 02/04/2024

Executive Summary

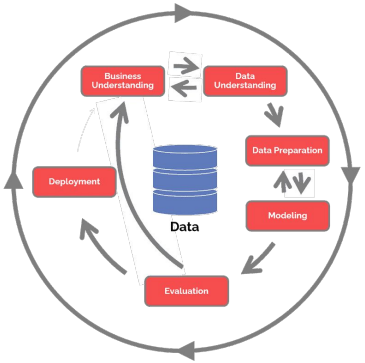
Online news and social media have become a dominant force in our society today. "Going viral" or shareability is how people learn about world events. While news feels like it is unbalanced and incredible negative most of the time, there is a need to explore what are the attributes of articles that drive shareability. By utilizing data from over 2 years from the website Mashable, we will look to understand what are the characteristics of a news article going viral.

Data Source: *K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal.*

Problem Statement & Process Model

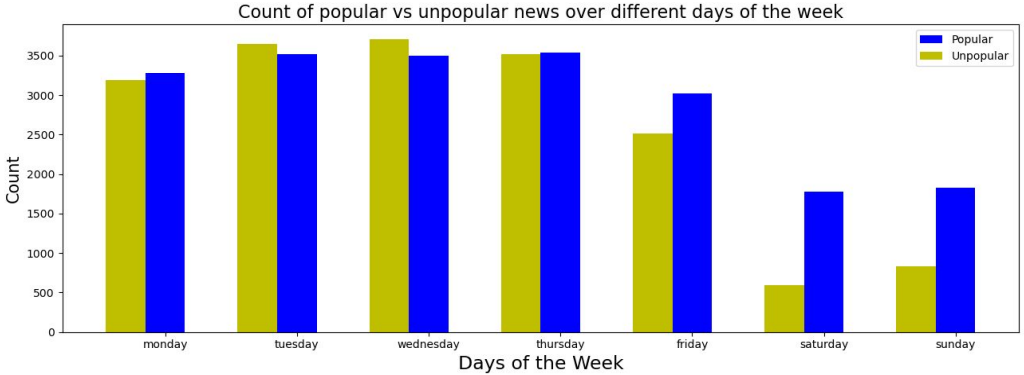
Problem Statement: What are the characteristics that make news go viral? We will look to understand if subjectivity, overall polarity, channel news is being displayed, and type of file format for news impacts the total number of shares on Mashable.

Process Model: We have implemented a Cross-Industry Standard Process for Data Mining (CRISP-DM) as our approach to answering the problem statement. This allows us the flexibility to approach the problem either by vertical or horizontal slicing of our features.



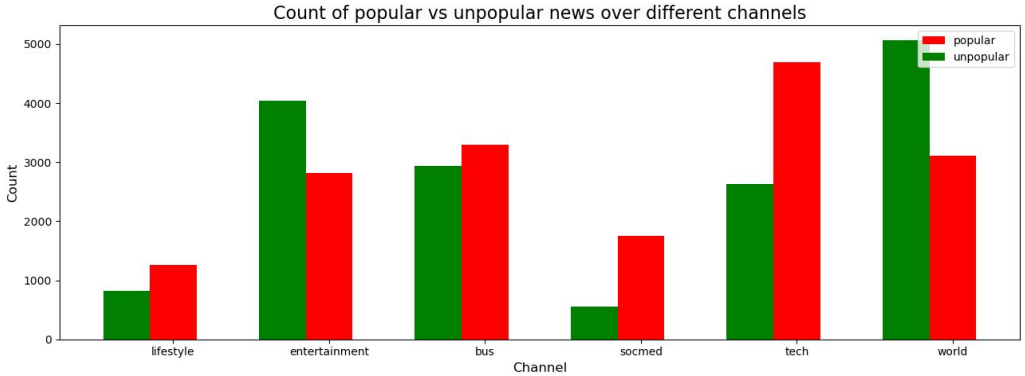
Data Source: *https://www.datascience-pm.com/crisp-dm-2/*

Exploratory Data Analysis (EDA)



Using EDA, we're able to determine the count of popular vs. unpopular news over different days of the week. News is more popular Monday, Thursday, Friday, Saturday, and Sunday. Whereas the news is more unpopular on Tuesday and Wednesday.

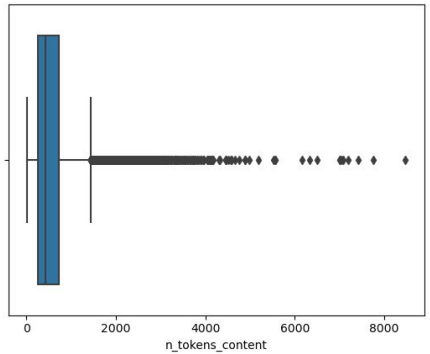
Using EDA, we're also able to determine the count of popular vs. unpopular over different channels. Lifestyle, Business, Social Media, and Tech are more popular vs. Entertainment and World are more unpopular.

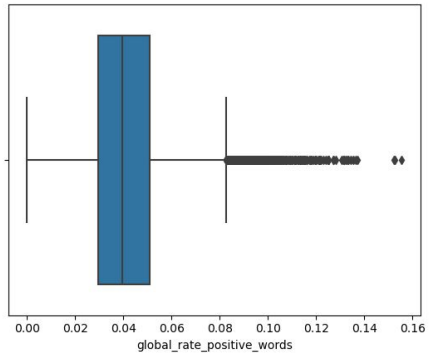


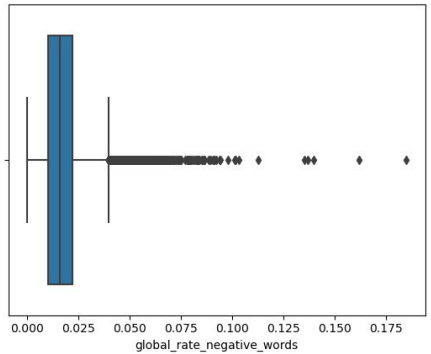
Feature Engineering

While performing EDA, we realized that there were no strong correlations with Shares and also the dataset had a large number of outliers across the variables (see below examples). Were were able to treat the outliers using “Yeo-Johnson” methodology.

Before Feature







After Feature