Rodrigo Sasse David

Using User-generated Content from Online Social Media to Forecast Movie Success

Goiânia

2017

Rodrigo Sasse David

Using User-generated Content from Online Social Media to Forecast Movie Success

Final project, in order to obtain degree, presented to Prof. Dr. Deborah Alves Silva Fernandes, from the Institute of Computing, Federal University of Goiás.

Universidade Federal de Goiás - UFG

Escola de Engenharia Elétrica, Mecânica e de Computação - EMC

Projeto Final de Curso 2

Supervisor: Prof. Dr. Deborah Silva Alves Fernandes

Goiânia

2017

Rodrigo Sasse David

Using User-generated Content from Online Social Media to Forecast Movie Success/ Rodrigo Sasse David. – Goiânia, 2017-

45 p. : il. (algumas color.) ; 30 cm.

Supervisor: Prof. Dr. Deborah Silva Alves Fernandes

Thesis (Graduation) – Universidade Federal de Goiás - UFG
Escola de Engenharia Elétrica, Mecânica e de Computação - EMC
Projeto Final de Curso 2, 2017.

1. Online Social Media. 2. Online Social Network. 3. Sentiment Analysis. 5. Forecast. 6. Box office.

CDU 02:141:005.7

Rodrigo Sasse David

Using User-generated Content from Online Social Media to Forecast Movie Success

Final project, in order to obtain degree, presented to Prof. Dr. Deborah Alves Silva Fernandes, from the Institute of Computing, Federal University of Goiás.

Paper approved. Goiânia, December 13, 2017:

Prof. Dr. Deborah Silva Alves Fernandes
Supervisor

Prof. Dr. Carlos Galvão Pinheiro Júnior
Invited 1

Prof. Ms. Márcio Giovane Cunha Fernandes
Invited 2

Goiânia
2017

To V. and S. They can not read this, but their sacrifices gave me tools to write it.
To Y. For being there for me for half of our lives, always.

Abstract

In a widely connected world, people are sharing their opinions and thoughts like never before. Social media plays an important role in the communication process nowadays, establishing trends and topics ranging from the environment and politics to technology and entertainment. For companies, this information is a resource to be harnessed just like the traditional word-of-mouth. Past studies have successfully correlated social media content to day-to-day events, ranging from the stock market to epidemics. This paper employees machine learning on sentiment analysis, combining it with the attention and popularity from Twitter data to forecast the opening box office revenue of movies. We obtained correlation ranging from 85% to 90%.

Key-words: Online social media, online social networks, Twitter, sentiment analysis, forecast, box office.

Resumo

Em um mundo vastamente conectado, pessoas estão compartilhando suas opiniões e pensamentos como nunca antes. Midias sociais *online* têm um papel importante no processo de comunicação atualmente e vêm moldando tópicos desde o meio-ambiente e política à tecnologia e entretenimento. Para empresas, esse é um recurso a ser investigado assim como o tradicional "boca-a-boca". Estudos anteriores correlacionaram, com sucesso, conteúdo de mídias sociais a eventos do dia-a-dia, do mercado financeiro a epidemias. Este estudo utiliza aprendizado de máquina em análise de sentimentos, combinando-a com a atenção e popularidade obtidos do Twitter a fim de prever a arrecadação de filmes. Obtivemos correlações entre 85% e 90%.

Palavras-chave: Mídias sociais *online*, redes sociais *online*, Twitter, análise de sentimentos, previsão, arrecadação, *box office*.

List of Figures

List of Tables

List of abbreviations and acronyms

API	Application Programming Interface
BOR	BOR
DJIA	Dow Jones Industrial Average
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
JSON	JavaScript Object Notation
OSM	Online Social Media
NLP	Natural Language Processing
REST	Representational State Transfer
SA	Sentiment Analysis
SVM	Support Machine Vector
URL	Uniform Resource Locator

Contents

1 Introduction

Before the Web 2.0[1], whenever someone needed an opinion regarding a product or service, it probably meant reaching out to family or friends. For a company, getting to know their clients' opinions probably meant to conduct an opinion survey. Presently, both consumers and companies have another alternative: the enormous amount of user-generated content shared on the Web.

Over the last two decades, the propagation of broadband Internet, and most recently, the ubiquity of mobile devices, led to an ever increasing social networking. Due to its ease-of-use and reach, online social media became an important tool on the day-to-day communication process, it has been establishing trends and topics ranging from the environment and politics to technology and entertainment (LIU, 2010).

For the past 20 years or so, people have been constantly and voluntarily sharing their opinions, feelings, and desires on blogs, forums, and social media. Not only these information are readily available, they can also be more trustworthy than information provided by the sellers or vendors (BICKART; SCHINDLER, 2001). For companies, this information is a resource to be harnessed just like the traditional word-of-mouth (SCHOUTEN; FRASINCAR, 2016).

Several studies successfully correlated social media user-generated content, usually refereed as "buzz' (THOMAS, 2004), to real-life events, such as stock market or TV viewing rates (BOLLEN; MAO; ZENG, 2011; LAMPOS; CRISTIANINI, 2010; O'CONNOR et al., 2010; CHENG; WU; CHEN, 2016).

In this study we tried to forecast the opening BOR of movies using user-generated content from Twitter. We used the hypothesis that "movies that are well talked about will be well-watched", proposed by (ASUR; HUBERMAN, 2010).

We tried to verify the hypothesis aforementioned seven years after it was initially proposed using similar techniques as the ones used by (ASUR; HUBERMAN, 2010) - comparing the "attention and popularity" and the results from sentiment analysis to the box office through linear regression.

A previous and small version of this study allowed us to create a prototype capable of forecasting box office revenue (BOR). The results were published in November 2017 on the ERI-GO[2] (Escola Regional de Informática, in Portuguese).

[1] http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html
[2] http://erigo.sbc.org.br/

2 Literature Review

In this chapter we discuss (a) online social media (OSM) nowadays, (b) the definition of sentiment analysis (SA), its techniques and challenges, and (c) the methodology and results of correlated studies in the area.

2.1 Online Social Media

An online social media (OSM), also known as online social service or simply social media, is an Internet-based and user-generated content-based platform which people use to build social relations with other people who share similar personal or professional interests (OBAR; WILDMAN, 2015).

OSMs are reaching more and more people every year. In 2010, 970 million people made use of an OSM worldwide, in 2017 the number jumped to 2.46 billion people[3]. The same study suggests that by 2021 the number of people using OSMs will exceed 3 billion worldwide. In the United States, in 2010, 48% of the population had an account on at least one OSM; in 2017, this number increased to 81%[3].

A quick lookup at Alexa[4], an Amazon.com company specialized on Web marketing, reveals that on the top 10 highest Internet traffic websites worldwide, four of them can be considered OSMs or alike (see Table 1). The same website also shows the average daily time spent on site of the top 10; users spend 40% more time online on average on OSMs compared to the other websites (9m40s compared to 6m49s).

Ranking	Website	Type	Daily Time on Site
1	Google.com	Search engine	8m01s
2	Youtube.com	OSM	8m20s
3	Facebook.com	OSM	9m54s
4	Baidu.com	Search engine	7m49s
5	Wikipedia.org	Encyclopedia	4m11s
6	Yahoo.com	Internet portal	4m06s
7	Google.co.in	Search engine	7m55s
8	Reddit.com	OSM	15m52s
9	Qq.com	OSM	4m35s
10	Taobao.com	Marketplace	8m57s

Table 1: Top 10 highest Internet traffic websites.

Souce: https://www.alexa.com/topsites, as of September 2017.

[3] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
[4] https://www.alexa.com/topsites

On Play Store[5], the online store for Android devices, three apps on the top 10 free apps can be classified as social media apps (see Table 2). On App Store[6], the online store for Apple devices, there are four on the top 10 free apps (see Table 3).

Ranking	App	Category
1	Instagram	Social
2	WhatsApp Messenger	Communication
3	Facebook Messenger	Communication
4	Uber	Maps & Navigation
5	Facebook	Social
6	Facebook Lite	Social
7	Netflix	Entertainment
8	Wish	Shopping
9	Palco MP3	Music & Audio
10	GO Launcher	Personalization

Table 2: Top 10 free apps on Play Store.

Source: https://play.google.com/store/apps/collection/topselling_free, as of September 2017.

Ranking	App	Category
1	WhatsApp Messenger	Communication
2	Instagram	Social
3	Facebook	Social
4	YouTube	Social
5	Facebook Messenger	Communication
6	Uber	Maps & Navigation
7	Netflix	Entertainment
8	Spotify Music	Music & Audio
9	Afterlight	Photo & Video
10	Snapchat	Social

Table 3: Top 10 free apps on App Store.

Source: https://www.apple.com/itunes/charts/free-apps, as of September 2017.

Not only more people using OSMs, a large number of them are also contributing to it. According to (KING; LI; CHAN, 2009), 76% of all U.S. broadband users actively contribute to social media sites in one form or another and 29% contribute regularly.

2.1.1 Microblogging Networks

There are different types of OSMs, it is possible to classify them in different categories based on their main type of media. Some categories may include social networks

[5] https://play.google.com/

[6] https://itunes.apple.com/

(Facebook, QQ, Twitter, WhatsApp), media sharing networks (Instagram, Snapchat, YouTube), discussion forums (Digg, reddit, Quora), bookmarking networks (Flipboard, Pinterest), consumer review networks (TripAdvisor, Yelp, Zomato), blogging and publishing networks (Medium, QZone, Tumblr, WordPress), social shopping networks (Etsy, Fancy, Polyvore) etc.

Presently, social and media sharing networks concentrate the largest bases of active users and monthly users (see Tables 4 and 5).

Ranking	Media	Category	Active Users
1	Facebook	Social Network	2.061
2	YouTube	Media Sharing	1.500
3	WhatsApp	Social Network	1.300
4	Facebook Messenger	Social Network	1.300
5	WeChat	Social Network	963
6	QQ	Social Network	850
7	Instagram	Media Sharing	700
8	QZone	Blogging	606
9	Tumblr	Blogging	368
10	Twitter	Social Network	328

Table 4: Top 10 OSMs with highest number of active users (in millions).

Source: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, as of September 2017.

Ranking	Media	Category	Monthly Users
1	Facebook	Social Network	135,56
2	Facebook Messenger	Social Network	102,32
3	Instagram	Media Sharing	76,88
4	Snapchat	Media Sharing	51,51
5	Twitter	Social Network	48,38
6	Pinterest	Bookmarking	44,04
7	Google Hangouts	Social Network	19,23
8	WhatsApp	Social Network	17,41
9	Skype	Social Network	13,74
10	Tinder	Social Network	10,17

Table 5: Top 10 OSMs with highest number of monthly users (in millions).

Source: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, as of September 2017.

The social networks, also known as microblogging networks, are interesting to the studies involving social media. Together they concentrate the majority of active users and most of them disposes of free APIs, allowing users to easily harvest posts from the networks.

2.1.2 Twitter

Twitter[7] was founded in 2006. It has rapidly grown, reaching enormous popularity worldwide. In 2012, more than 100 million users posted an average of 340 million messages daily[8]. In the first quarter of 2017, 328 million users posted an average of 500 million messages daily[7].

Messages on Twitter are called "tweets", they are limited to 140 characters. The fact that tweets are limited to a short forces users to be more concise (THAKKAR; PATEL, 2015). This trait makes Twitter to be known as "what's happening now" network (ZHOU et al., 2013). This conciseness can lead to informal language, like abbreviations or slang, which can be a challenge when it comes to sentiment analysis (see subsection 2.1.3).

Posted tweets are shown on the user's timeline, which can be set up as public or private. Public timelines can be seen by anyone, even if they don't have an account on Twitter; private timelines can be seen only by friends. On Twitter, 90% of the users set up their timelines as public (STROHMAIER; WAGNER, 2014).

When it comes to studies regarding OSMs or user-generated content, Twitter is one of the most used networks (ARNOLD; VRUGT, 2008; BOLLEN; MAO; ZENG, 2011; ARIAS; ARRATIA; XURIGUERA, 2013; ZHOU et al., 2013; SCHOUTEN; FRASINCAR, 2016).

In September 2017, Twitter announced their plans to increase the maximum length of tweets to 280 characters[9]. For now, the new length is in the test phase for all languages except Chinese, Japanese, and Korean. The platform stated that is because these three languages can convey about double the amount of information in one character as one can in many other languages, such as English, Spanish, Portuguese, or French.

The announcement on Twitter's official blog[9] included a comparison between tweets written in English and Japanese (see Figure 1). While tweets in Japanese have a peak at 15 characters, the peak for English is at 34 characters. Also, only 0.4% of the tweets written in Japanese reach the length limit, while 9% of the tweets written in English do.

This is an interesting change for studies based on Twitter data. While the limit of characters is still low, now people can express more emotion on a single tweet.

[7] https://www.twitter.com/

[8] https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html

[9] https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html

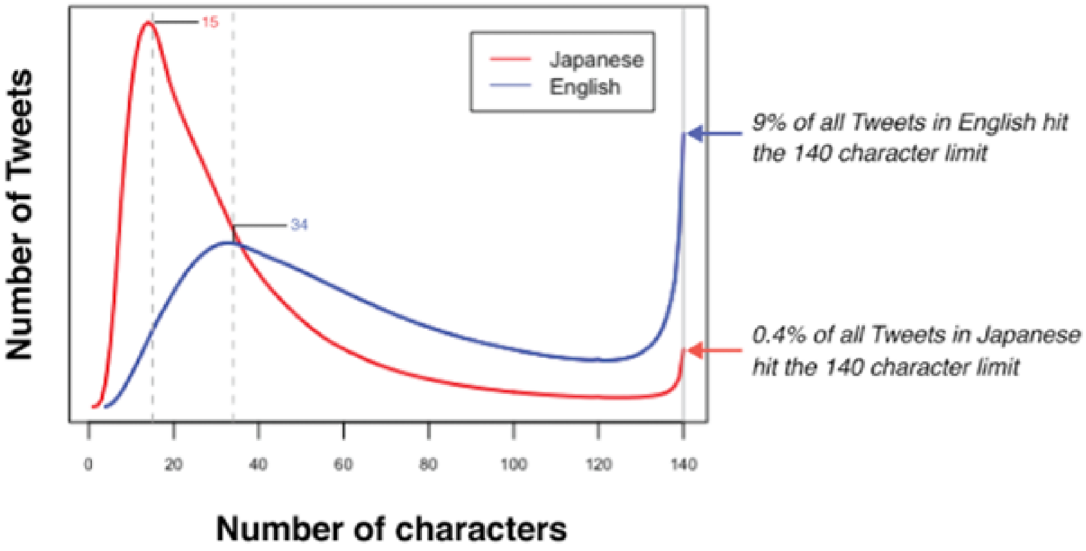


Figure 1: Number of tweets versus number of characters in English and Japanese.

Source: https://blog.twitter.com/official/en$_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html$

2.1.3 Twitter's API

Anyone with a registered account on Twitter can make use of Twitter data. Through its API[10] one can start their own application or make use of one of the several third-party libraries[11]. Twitter offers two different APIs: REST and Streaming. Both can be used under free, premium, or enterprise plans.

The REST API (see Figure 2) can be used to make query requests, such as latest tweets posted containing a keyword or phrase, latest tweets posted by an user, tweets posted on a time frame etc. The Streaming API (see Figure 3) stablishes a persistent HTTP connection between the application and Twitter's API server. The application sets a listener for tweets. All tweets matching the listener's query are broadcasted by Twitter to the application.

Twitter implements rate limiting on its API[12]. Rate limits are divided into 15 minute intervals. Each method has a limit number of available calls every window. Twitter also limits the number of results retrieved and the time scope each request can search for.

Due to the API's limitations we were forced to used a third-party application[13] capable of retrieving tweets older than two weeks. This application does not implement the Twitter's API, instead it generates a GET request on Twitter's advanced search to retrieve tweets directly from the HTML file returned.

[10] https://developer.twitter.com/en/docs

[11] https://developer.twitter.com/en/docs/developer-utilities/twitter-libraries

[12] https://developer.twitter.com/en/docs/basics/rate-limiting

[13] https://github.com/Jefferson-Henrique/GetOldTweets-python

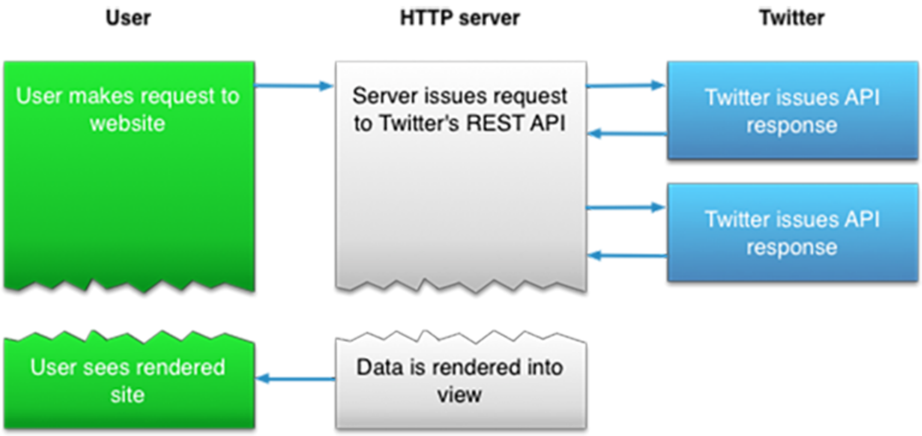

User
HTTP server
Twitter
User makes request to website
Server issues request to Twitter's REST API
Twitter issues API response
Twitter issues API response
User sees rendered site
Data is rendered into view

Figure 2: Representation of Twitter's REST API.

Source: https://developer.twitter.com/en/docs

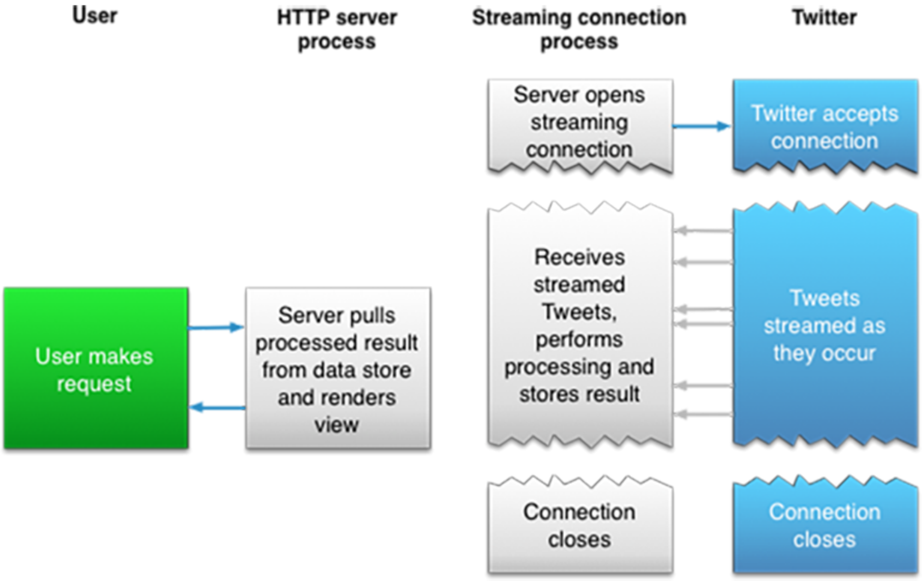

User
HTTP server process
Streaming connection process
Twitter
Server opens streaming connection
Twitter accepts connection
Receives streamed Tweets, performs processing and stores result
Tweets streamed as they occur
User makes request
Server pulls processed result from data store and renders view
Connection closes
Connection closes

Figure 3: Representation of Twitter's Streaming API.

Source: https://developer.twitter.com/en/docs

2.2 Sentiment Analysis

Sentiment analysis (also known as Opinion Mining) is a computational science branch that works on the intersection of other three branches: information retrieval, natural language processing, and artificial intelligence (SCHOUTEN; FRASINCAR, 2016). Together, they give the tools for sentiment analysis to rise to its challenges: identification, classification, and agglutination (SCHOUTEN; FRASINCAR, 2016).

Given a sentence, sentiment analysis should identify: which part of the sentence carries opinion, who wrote it, its subject, and the opinion itself (KING; LI; CHAN, 2009). Note the following example:

> *"(1) I bought a MacBook Pro a couple of days ago. (2) It's a nice notebook. (3) The screen is gorgeous. (4) The battery life is great. (5) Although the performance is not the best, that's ok to me. (6) However, my wife was mad at me as I did not tell her before I bought it. (7) She thought the notebook was too expensive, and wanted me to return it."*

Based on the example above, sentiment analysis must be able to identify that sentences (2), (3), and (4) express positive opinions, while sentences (5), (6), and (7), negative opinions. It also should identify that sentences (1) and (2) reference the notebook as a whole, while sentences (3), (5), (6), and (7), some of its attributes - screen, battery life, performance, and price. At last, it should identify that sentences (2), (3), (4), and (5) express the opinions of the author of the text, while sentences (6) and (7), the opinions of his wife.

Generally, opinion can be expressed on anything - products, services, people, organizations, events, topics etc. The term "object" is commonly used to refer to the entity target of the opinion (LIU, 2010). The object may have a set of components and a set of attributes, the union of these two sets is the set of features (LIU, 2010). On the example above, "MacBook" is the object; "screen" and "battery life" are components, and "performance" and "price", attributes.

From a given document d, a set of objects $\mathrm{O} = \{o_1, o_2, ..., o_\mathrm{p}\}$, which has a set of features $\mathrm{F} = \{f_{i1}, f_{i2}, ..., f_{i\mathrm{q}}\}$, and a set of authors $\mathrm{H} = \{h_1, h_2, ..., h_\mathrm{r}\}$, the opinions in d can be represented by the quintuple $(o_i, f_{ij}, p_{ijkl}, h_i, t_l)$, on which o_i represents the object; f_{ij}, the set of features; p_{ijkl}, the polarity of the opinion; h_i, the opinion holder; and t_l, the moment the opinion was shared (LIU, 2010).

Each element of the quintuple is an open challenge for the natural language processing (NLP). There are methods (see Subsection 2.2.1) to tackle each individual problem, but a combined solution usually requires a combination of different approaches.

Since most websites display information on the author of the text or sentence, as well as its timestamp, by analyzing texts referencing a single object and feature at a time, the quintuple mentioned above has only the polarity of the opinion p_{ijkl} left undiscovered.

Due to the large number of product and movie reviews shared on the Web and the interest of companies in the opinion of their public, this type of message has been the main subject for sentiment analysis studies (DAVE; LAWRENCE; PENNOCK, 2003;

POPESCU; ETZIONI, 2007; ZHANG; YU; MENG, 2007; WIJAYA; BRESSAN, 2008; BRODY; ELHADAD, 2010; ZHOU et al., 2013).

2.2.1 Entity and Aspect Detection

In a broad analysis, the entity, the object of the sentiment, and its aspects, characteristics or properties of the entity, may not be known. The task of identifying those elements are known as entity and aspect detection. There are basically five approaches for that: frequency-based, syntax-Based, Supervised Machine Learning, Unsupervised Machine Learning, and Hybrid (SCHOUTEN; FRASINCAR, 2016).

The first approach, frequency-based, relies on the fact that there is a set of words that are used more frequently that the rest. Those words, usually simple and compound nouns, very often represent aspects of sentiment. For an noun to be considered an aspect it has to used in sentences more often than the expected average. There are some words that are frequently used, but can not be considered as entity or aspects; for example, “dollar” or “bucks” in consumer reviews (DODDS; DANFORTH, 2010).

The syntax-based method find aspects by means of the syntactical relations they are in. A very common relation is the adjectival modifier between a sentiment word and an aspect, as in “awesome movie”, where the word awesome, as sentiment, modifies the word movie, as aspect (SCHOUTEN; FRASINCAR, 2016). A big advantage of this method is its ability to detect low-frequency aspects, but it usually requires a high number of well described syntactical relations (ZHAO et al., 2010).

There are not a lot of methods that employ only a supervised approach. Other approaches, like frequency-base, are usually applied in conjunction (SCHOUTEN; FRASINCAR, 2016). In general, those methods make use of supervised classification, where sentiments are manually classified in positive or negative, which can make this approach very costly (GONÇALVES et al., 2013).

In the unsupervised approach, the detection requires only labeled data to test and validate the model. The methods usually employ a lexicon of words (sometimes refereed as “bag of words”), which operates like a dictionary, but instead of the definition of the words, it contains their syntactical classification (HOFMANN, 2000).

Finally, when two types of methods are used, they are called hybrid methods. Methods can combine other approaches to generate more salient features. For example, a method can employ a supervised approach combined with frequency-based and syntax-based approaches, a list of adjectives, and a list of emoticons (GONÇALVES et al., 2013).

2.2.2 Sentiment Classification and Subjectivity

Textual information can be broadly classified in two main categories: facts and opinions. While facts are objective expressions about something, opinions are usually subjective expressions about people's sentiments or feelings (LIU, 2010). Sentiment analysis is only interested in textual information containing opinions. The task of identifying those sentences is called subjectivity classification.

Two analysis are possible regarding the length of the text: document-level or sentence-level. The first takes the whole text or document into account, and tries to estimate the polarity of the opinion carried by it. The second, sentence-level, tries to estimate the polarity of the opinion carried by each sentence individually (LIU, 2010).

The problem of classification, regardless of the length of text, can be tackled similarly as entity and aspect detection, with dictionary-based, supervised machine learning, and unsupervised machine learning approaches (SCHOUTEN; FRASINCAR, 2016). The goal is not detect sentiment, but to assign a score to the sentiment previously identified.

The dictionary-base approach compare each word in a sentence with the words in the dictionary. Words in the dictionary - adjectives and nouns - are labeled in positive or negative. When a word in the sentence matches a word in the dictionary, the score of the sentence is incremented if the word is positive, decremented if the word if negative. The sentence is classified as positive if its score is positive, and vice versa.

This approach is very simple and easy to implement, but can present some poor results, which is why other methods are usually employed in combination (SCHOUTEN; FRASINCAR, 2016). For example, when the final score is zero, the method checks for the word distance between the sentiment adjective or noun to the aspect. The sentiment word closest to the aspect breaks the tie. The method also check for a negation word near the sentiment word, which causes its score to flip.

Sentiment classification can be easily formulated as a supervised learning problem since there will be only labels, positive and negative. Not only that, review-like sentences usually have an assigned rating, like 1 to 5 stars or 0 to 100 score, which makes training and test data readily available (SCAFFIDI et al., 2007).

In unsupervised learning, explicit aspects are used to find a potential sentiment phrase by looking for a sentiment phrase in its vicinity. Each potential sentiment phrase is examined, only the ones that show a positive or negative sentiment are retained. The final output is a set of sentiment phrases with their most likely polarity label (SCHOUTEN; FRASINCAR, 2016).

2.3 Correlated Studies

Over the last decade, several studies successfully correlated social media user-generated content, usually refereed as “buzz’ (THOMAS, 2004), to real-life events, ranging from stock market volatility and epidemics to public elections and TV viewing rates.

(BOLLEN; MAO; ZENG, 2011) tried to verify if the public mood is correlated to or even predictive of economic indicators. They investigated whether measurements of collective mood states derived from large-scale Twitter feeds were correlated to the value of the Dow Jones Industrial Average over time. They obtained an accuracy of 86.7% in predicting the daily up and down changes in the closing values of the DJIA.

(LAMPOS; CRISTIANINI, 2010) reported on a monitoring tool to measure the prevalence of disease in a population by analyzing the contents of social networking tools, such as Twitter. Their method was based on the analysis of tweets daily, searching for symptom-related statements, and turning statistical information into a flu-score. They compared the flu-score with data from the Health Protection Agency, obtaining a correlation greater than 95%.

(O’CONNOR et al., 2010) connected public opinion measured from polls with sentiment measured from text. They analyzed several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and found they correlated to sentiment word frequencies in contemporaneous Twitter messages. In several cases they obtained correlation as high as 80%.

(CHENG; WU; CHEN, 2016) examined the relationship between social media site Facebook and TV ratings. The study applied multiple regression models and determined that the key social media measures correlated with TV ratings. In essence, TV shows with higher number of posts and engagement are likely to relate to higher ratings, obtaining correlation as high as 90%.

(ASUR; HUBERMAN, 2010) tried to verify the hypothesis that “movies that are well talked about will be well-watched”. In 2013 (ARIAS; ARRATIA; XURIGUERA, 2013) did a similar study. They collected 2.89 and 100 million tweets, referring 24 and 100 movies, respectively. They used linear regression and Support Vector Machine (SVM) to correlated the buzz and the positive-negative ratio to the BOR, obtaining correlation of 97% and 68%, respectively.

(LIU et al., 2016) also tried to forecast the BOR by correlating three different aspects, purchase intention, attention and popularity, and sentiment analysis. They used linear and support vector regression, and obtained a correlation of 94%.

3 Methodology and Results

In this chapter we discuss (a) the dataset acquisition, (b) a statistical analysis on the dataset, (c) the used sentiment analysis technique, and (d) the results obtained.

3.1 Dataset Composition

In this study we tried to verify the hypothesis that "movies that are well talked about will be well-watched". This hypothesis was proposed by (ASUR; HUBERMAN, 2010) in 2010.

We chose 24 movies for the study (see Table 6). We selected the movies with highest opening BOR released up until September 2017.

Movie Name	Opening Box Office	Closing Box Office
Beauty and the Beast	174,750,616	504,014,165
Spider-Man: Homecoming	117,027,503	332,707,249
Despicable Me 3	72,434,025	262,271,065
The LEGO Batman Movie	53,003,468	175,750,384
Kong: Skull Island	61,025,472	168,052,812
Cars 3	53,688,680	152,476,561
War for the Planet of the Apes	56,262,929	146,458,374
Transformers: The Last Knight	44,680,073	130,168,683
Girls Trip	31,201,920	115,004,375
Baby Driver	20,553,320	107,668,101
Annabelle: Creation	35,006,404	101,578,510
John Wick: Chapter Two	30,436,123	92,029,184
The Emoji Movie	24,531,923	84,912,232
The Mummy	31,688,375	80,101,125
Alien: Covenant	36,160,621	74,262,031
Atomic Blonde	18,286,420	51,573,925
The Dark Tower	19,153,698	50,548,577
Smurfs: The Lost Village	13,210,449	45,020,282
Going in Style	11,932,330	45,018,541
All Eyez on Me	26,435,354	44,922,302
47 Meters Down	11,205,561	44,300,147
Valerian	17,007,624	40,479,370
King Arthur: Legend of the Sword	15,371,270	39,175,066
Fist Fight	12,201,873	32,187,017

Table 6: Movies selected for the study and their respective opening and BOR (in US Dollars).

Source: http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm, as of July 2017

We searched for tweets using the hasthags used by promotional material (e.g.: #BeautyAndTheBeast, #Cars3, #AlienCovenant), and the movie title striped of special characters like hyphens and colons (see Table 7). We excluded movies whose name could easily blend into day-to-day conversation, it would be too difficult to correctly identify tweets that were relevant to the study. Example of those movies are “Dunkirk”, “Get Out”, “It”, “Logan”, and “Split”.

Searched Named	Searched Hashtag
Beauty and the Beast	#BeautyAndTheBeast
Spider Man Homecoming	#SpiderManHomeComing
Despicable Me 3	#DespicableMe3
The LEGO Batman Movie	#TheLegoBatmanMovie
Kong Skull Island	#KongSkullIsland
Cars 3	#Cars3
War for the Planet of the Apes	#WarForThePlanet
Transformers The Last Knight	#Transformers
Girls Trip	#GirlsTrip
Baby Driver	#BabyDriver
Annabelle Creation	#AnnabelleCreation
John Wick Chapter Two	#JohnWick2
The Emoji Movie	#EmojiMovie
The Mummy	#TheMummy
Alien Covenant	#AlienCovenant
Atomic Blonde	#AtomicBlonde
The Dark Tower	#TheDarkTower
Smurfs The Lost Village	#SmurfsMovie
Going in Style	#GoingInStyle
All Eyez on Me	#AllEyezOnMe
47 Meters Down	#47MetersDown
Valerian	#Valerian
King Arthur Legend of the Sword	#KingArthur
Fist Fight	#FistFight

Table 7: Movie names and hashtags used for the study.

For our time scope, we the used a “critical period” (ASUR; HUBERMAN, 2010), defined as the time when promotional campaigns are in full swing, which is usually the week before and the week after the movie’s premiere.

We only collected tweets posted in the English language. That was due to the fact that we only used the opening and closing BORs in the United States. This information can be easily found on the Web[14].

Tweets are returned as objects, they have many attributes - tweets’ id, author’s id, date and time of creation, message text, language, geolocation coordinates, number of

[14] http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm

retweets, number of favorites etc. The tweets we collected were stored in a database; we choose to store only the tweets' id, its author's id, the date and time of creation posted, and the message text.

Since we ran the application multiple times to ensure getting all tweets in the defined time scope, we ended up collecting some tweets multiple times. We then ran a query on the database to delete duplicate data. In the end, we collected 4.31 million unique tweets.

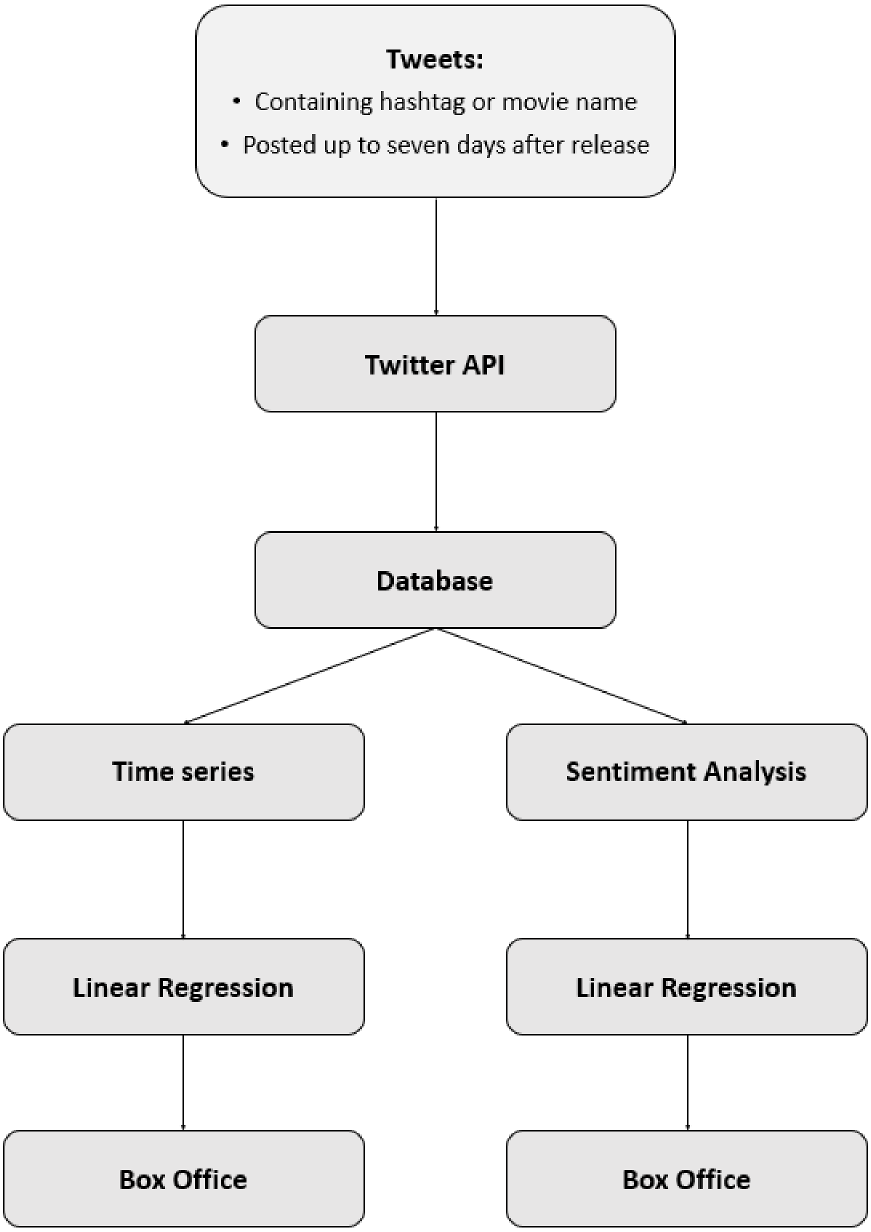


Figure 4: Model of collecting and analysis.

3.2 Statistical Analysis

The plot of the data suggested a linear behaviour (see Figure 5). The number of tweets referring each movie (popularity) followed an increase the week before the premiere day, then a decrease a week after it, making the premiere weekend the busiest period for movies on Twitter.

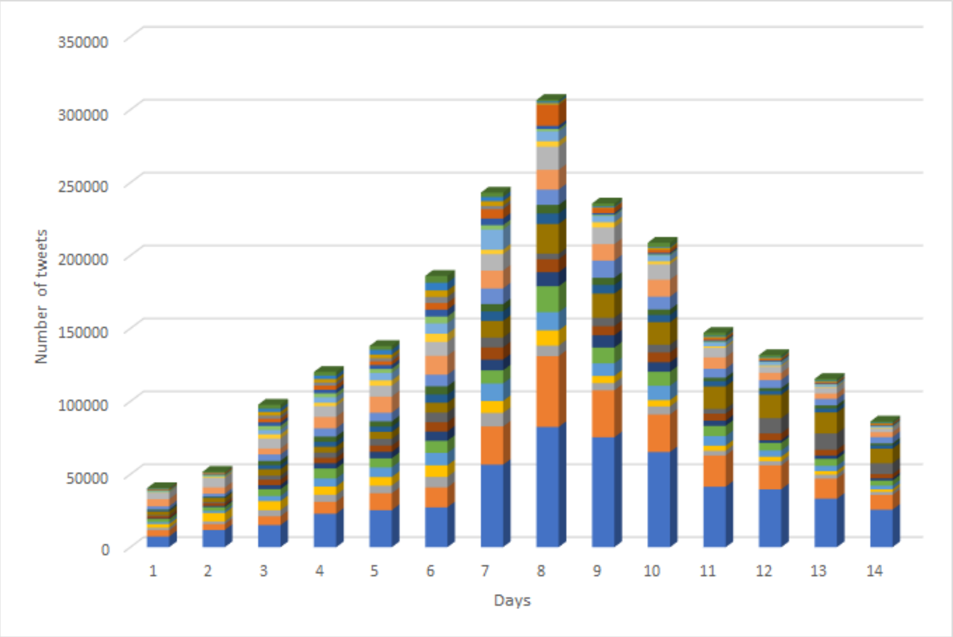


Figure 5: Number of tweets daily.

Linear regression is used when there is a linear relation between the input and the output - an increase or decrease on the input will cause a proportional increase or decrease on the output. Any linear regression model can be expressed as:

$$y = A_1 * x_1 + A_2 * x_2 + ... + A_\mathrm{x} * x_\mathrm{n} + B$$

where y is the output; x, the input; A, the coefficient of the input; and B, the linear coefficient (also refereed as error).

The accuracy of linear regression model are usually measured by two parameters, R^2 and *p-value*. The first tells how close the data are to fitted regression line, or the percentage of the response variable variation that is explained by a linear model; a high R^2 indicates a strong correlation between the input and output, and vice versa.

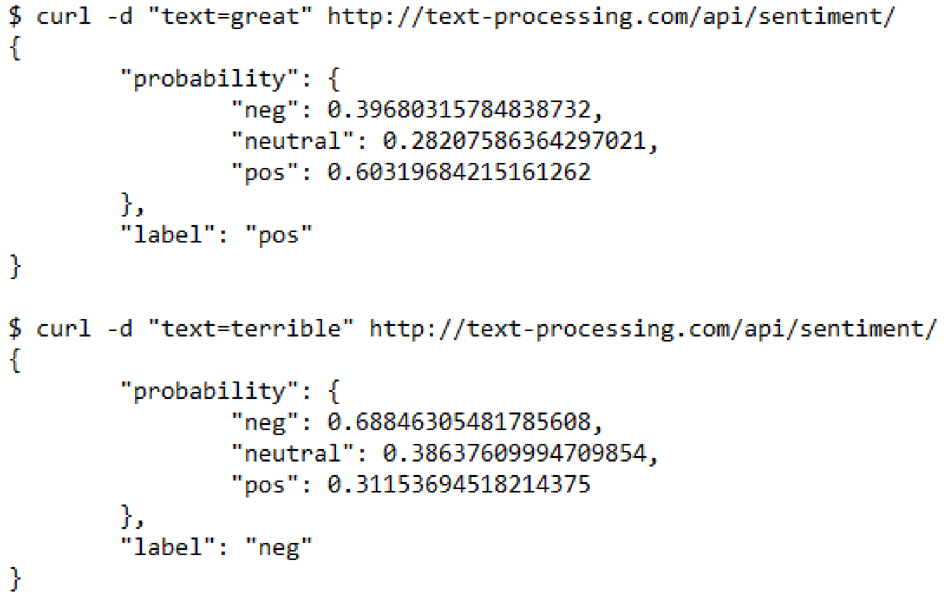
The second, p-value, indicates the probability of accepting the null hypothesis, which is the lack of significant correlation between data. Therefore, a p value below the adopted significance level of 5% ($p < 0.05$) rejects H0, that is, there is correlation between the data. The low p-value ($p < 0.001$) observed indicates that there is a very strong correlation, even at a significance level of 0.1%.

Linear regression is a simple, yet powerful model. There are studies in the OSM and SA fields that obtained solid results using it (ASUR; HUBERMAN, 2010; ARIAS; ARRATIA; XURIGUERA, 2013; CHENG; WU; CHEN, 2016).

3.3 Sentiment Analysis

There are libraries, like NLTK[15], and websites, like Text-Processing[16], capable of performing sentiment analysis right out of the box for free. We decided to use them to label our data since the goal of study was not to compare techniques, or to find the best technique, but to use SA as a tool to find or improve our model of forecasting.

Text-Processing is a service that offers free and paid text processing services like extraction of features, sentiment analysis, part-of-speech tagging etc. It also disposes of a large set of trained data and classifiers. Its API receives a string of text, labels it, and returns the results as a JSON object with the label and probability for each label (see Figure 6).



```
$ curl -d "text=great" http://text-processing.com/api/sentiment/
{
        "probability": {
                "neg": 0.39680315784838732,
                "neutral": 0.28207586364297021,
                "pos": 0.60319684215161262
        },
        "label": "pos"
}

$ curl -d "text=terrible" http://text-processing.com/api/sentiment/
{
        "probability": {
                "neg": 0.68846305481785608,
                "neutral": 0.38637609994709854,
                "pos": 0.31153694518214375
        },
        "label": "neg"
}
```

Figure 6: Example of Text-Processing's API use.

Source: https://market.mashape.com/japerk/text-processing

Initially, we considered using Text-Processing to label all our dataset once it already has a large set of labeled data, but since its free plan allows about 4.000 requests per hour per IP address we only used it to label 10% of our selected dataset.

[15] http://www.nltk.org/

[16] https://market.mashape.com/japerk/text-processing

The rest of the dataset was labeled using Python's NLTK (Natural Language Toolkit). This library implements useful methods for text processing, just like Text-Processing, and machine learning techniques for SA. We used a Naive Bayes classifier, which we trained with the data we labeled earlier.

For this part of the study we excluded from the dataset tweets containing URLs as they can be considered "promotional material" (ASUR; HUBERMAN, 2010). Those tweets may not be interesting to SA because they usually have neutral or deliberately positive polarity, which can tamper the results.

Also, it is common to find studies pre-processing their dataset before employing SA, among them: elimination of stop-words (that, the, this, to etc.), elimination of special characters (exclamation points, question marks etc.), elimination of emojis, possible HTML tags etc. The goal of pre-processing is to ensure the elimination of noise and uninformative information, which can yield slightly better results (HADDI; LIU; SHI, 2013).

We did not do any pre-processing in this study because both Text-Processing and NLTK are capable of dealing with them, but mostly because the results after it are slightly better (an increase of 3.5%) (HADDI; LIU; SHI, 2013).

3.4 Results from Attention and Popularity

At first, we investigated the impact of promotional material on the BOR. Prior to the release of a movie, media companies, and producers generate promotional information in form of trailer videos, news, blogs, and photos, in order to promote word-of-mouth cascades. This kind o material usually has an URL linking to external media.

We used only the tweets containing URLs. We used linear regression, using as input the percentage of tweets containing URLs (PURL), the total number of tweets containing URLS (TURL), the number of theaters (THTR), and the aggregate score from critics (ASCR), ranging from 0 to 100, and the BOR as output. Except for the ASCR, these are the same input used by (ASUR; HUBERMAN, 2010) in their study. Data are shown in Table 8, results are shown in Table 9.

The number of theaters premiering a movie can be easily found on the same website we used to get the opening BOR[17]. The aggregate score of each movie can be easily found on IMDb[18].

Except for the results of the percentage of tweets containing URLs, all the other results showed a correlation around 85%, and a very high significance ($p < 0.001$). A

[17] http://www.boxofficemojo.com/yearly/chart/?yr=2017&p=.htm

[18] http://www.imdb.com

Movie Name	PURL	TURL	THTR	ASCR	TTWT	PNRT
Beauty and the Beast	33.77%	180.040	4.210	65	549.551	3,82
Spider-Man: Homecoming	24.74%	67.214	4.348	73	271.803	1,95
Despicable Me 3	45.42%	31.411	4.529	49	69.185	2,65
The LEGO Batman Movie	46.02%	42.479	4.088	75	92.341	1,78
Kong: Skull Island	43.68%	46.498	3.846	62	106.484	2,51
Cars 3	32.11%	32.250	4.256	59	100.816	1,31
Planet of the Apes	30.16%	16.559	4.022	82	54.932	0,88
Transformers	46.44%	26.962	4.069	28	58.077	1,23
Girls Trip	33.53%	13.389	2.591	71	39.954	2,20
Baby Driver	18.40%	19.098	3.226	86	103.827	1,18
Annabelle: Creation	30.82%	13.466	3.502	62	43.706	2,18
John Wick: Chapter Two	48.01%	21.754	3.113	75	45.328	1,75
The Emoji Movie	26.50%	17.161	4.075	12	64.798	0,85
The Mummy	31.09%	40.606	4.035	34	130.669	1,29
Alien: Covenant	36.12%	47.505	3.761	65	131.592	1,23
Atomic Blonde	34.42%	11.261	3.304	63	32.734	0,99
The Dark Tower	35.92%	22.957	3.451	34	63.941	0,74
Smurfs: The Lost Village	62.41%	10.886	3.610	40	17.452	0,89
Going in Style	51.60%	8.931	3.061	50	17.315	1,25
All Eyez on Me	11.32%	5.868	2.471	38	51.880	0,72
47 Meters Down	37.60%	1.114	2.270	52	2.964	1,14
Valerian	44.46%	6.820	3.553	51	15.347	0,68
King Arthur	37.54%	4.936	3.702	41	11.716	0,37
Fist Fight	35.90%	8.465	3.185	37	23.591	0,57

Table 8: Numbers of PURL and TURL obtained.

Input	R^2	p-value
PURL	0.110	0.612
TURL	0.848	1.14e-10
TURL + THTR	0.851	8.70e-10
TURL + ASCR	0.862	2.62e-09
TURL + THTR + ASCR	0.862	2.62e-09

Table 9: Results for promotional material.

further investigation revealed that the percentages of promotional material for each movie were similar to each other, an average of 36.53%, and an average deviation of 8.15%, while the BORs are widely spaced out, therefore resulting in a bad correlation.

Then we tried to forecast the opening BOR, now using the total number of tweets (TTWT), containing promotional material or not. We applied the same methodology above. Results are shown in Table 10.

Just as the previous model, all the results showed a correlation around 85%, and a very high significance ($p < 0.001$).

Input	R^2	p-value
TTWT	0.835	2.02e-09
TTWT + THTR	0.862	3.52e-10
TTWT + ASCR	0.837	8.52e-10
TTWT + THTR + ASCR	0.874	2.02e-09

Table 10: Results for popularity.

3.5 Results from Sentiment Analysis

Later we used the results from sentiment analysis to forecast the opening BOR. First we used the ratio of positive to negative tweets (PNRT), the number of theaters (THTR), and the aggregate score from critics (ASCR) as the input, and the opening BOR as the output. Results are shown in Table 11.

(ASUR; HUBERMAN, 2010) used the same input in 2010, excluding the score from critics. They also used the ratio of positive to negative tweets, which indicates how well, or badly, a movie is talked about. Neutral tweets were excluded as they do not express direct opinion.

Input	R^2	p-value
PNRT	0.839	2.11e-10
PNRT + THTR	0.837	2.09e-10
PNRT + ASCR	0.831	2.95e-09
PNRT + THTR + ASCR	0.829	1.77e-09

Table 11: Results for sentiment analysis.

The results from the model using SA are very similar to the ones we obtained from the model using promotional and popularity. We obtained a correlation around 84%, and a very high significance ($p < 0.0010$).

We repeated the earlier analysis, but instead of considering the whole dataset for the positive to negative ratio, we started using 1% of our data, chosen randomly, and increased it by 1% each interaction until we found a correlation close to the previous ones. In this analysis, we only used the positive to negative ratio (PNRT). Results are shown in Table 12.

We accomplished the goal at a percentage of 6% - with a correlation of 83.9%, and a very significance ($p < 0.001$) - since it is greater than the lowest percentage we got in the results of promotional material and popularity, 83.5%.

Percentages	R^2	p-value
1%	0.612	6.75e-05
2%	0.699	5.34e-07
3%	0.743	1.92e-08
4%	0.789	2.39e-08
5%	0.817	2.75e-09
6%	0.836	2.90e-10

Table 12: Results for sentiment analysis.

3.6 Results from the combination of Attention and Popularity with Sentiment Analysis

Finally, we combined the attention and popularity with SA. We used combinations of the same input as before - the total number of tweets containing URLS (TURL), the total number of tweets (TTWT), the number of theaters (THTR), the aggregate score from critics (ASCR), and the ratio of positive to negative tweets (PNRT). Results are shown in Table 13.

Input	R^2	p-value
TURL + PNRT	0.869	2.10e-10
TTWT + PNRT	0.875	1.21e-10
TURL + PNRT + THTR + ASCR	0.877	3.55e-09
TTWT + PNRT + THTR + ASCR	0.906	2.90e-10

Table 13: Results for sentiment analysis.

The results above show that SA improved the results we obtained before, increasing the average correlation from 85% to 88%. It also helped this model obtain a correlation over 90%.

4 Conclusion

Movies are just like any investment. Investors - motion picture studios, producers, distributors - all want to know if they will make a profit or lose money, if a movie will be a blockbuster or a flop.

A simple linear regression using the closing BOR as output, and the opening BOR as input indicates a correlation of 92%. Add the number of theaters and the number of days the movie will stay on screen, and the correlation will go up to 97%. But forecasting the opening BOR is more difficult. Every year we see big productions with really bad opening revenues, and budget movies doing really well.

For the investors - producers and studios - knowing before hand how a movie is supposed to perform can be an important tool in order to reallocate resources in terms of promotional material, evaluate its impact on the company's stock value etc. Not only that, OSMs can give a glimpse on broader aspects of the civilization. It can be used to identify and analyze aspects of human interaction, the wisdom of crowds etc.

Back in 2010, (ASUR; HUBERMAN, 2010) proposed and verified the hypothesis that "movies that are well talked about will be well-watched". Back then, OSMs were not as popular and ubiquitous as they are now. Just over the last seven years, 1.5 billion people joined at least one OSM worldwide[19].

In terms of attention and popularity, (ASUR; HUBERMAN, 2010) obtained correlations ranging from 80% to 97%, while we obtained correlations ranging from 83% to 87%. A possible reason to explain that is the fact that they used the movies that were being released at the time, while we selected the top movies of 2017. The movies they selected are more widely spread, both in terms of attention, and in terms of BOR, while the movies we selected have closer BOR, specially at the top.

They also used SA to try to forecast the BOR. They employed similar techniques to the ones we did in this study, but instead of NLTK, they used LingPipe. The combination of SA improved some of their results, just like it did for some of ours, but the best result they obtained combining SA was 94%, still lower than the 97% they obtained combining popularity and the number of theaters.

The addition of SA improved the models from attention and popularity and helped us achieve our best result, a correlation of almost 91%. But SA's true strength lies on how little data we needed to achieve similar results compared with attention and popularity.

(ASUR; HUBERMAN, 2010; ARIAS; ARRATIA; XURIGUERA, 2013) collected

[19] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

2.89 million and 100 million tweets, respectively, to analyze the attention and popularity features though linear regression. They obtained correlations as high as 93% and 97%, respectively. With SA we obtained a correlation of 83.6% with just 120.000 tweets. That is 10% and 14% lower, respectively, but we only need 4% and 0.12% of the data they need, respectively.

This feature open doors for real-time applications. The small percentage of data required helps creating a less expensive model, yet capable of yielding solid results.

Bibliography

ARIAS, M.; ARRATIA, A.; XURIGUERA, R. Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 5, n. 1, p. 8, 2013. Citado 4 vezes nas páginas 24, 30, 35, and 41.

ARNOLD, I. J.; VRUGT, E. B. Fundamental uncertainty and stock market volatility. *Applied Financial Economics*, Taylor & Francis, v. 18, n. 17, p. 1425–1440, 2008. Citado na página 24.

ASUR, S.; HUBERMAN, B. A. Predicting the future with social media. In: IEEE. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. [S.l.], 2010. v. 1, p. 492–499. Citado 8 vezes nas páginas 19, 30, 31, 32, 35, 36, 38, and 41.

BICKART, B.; SCHINDLER, R. M. Internet forums as influential sources of consumer information. *Journal of interactive marketing*, Elsevier, v. 15, n. 3, p. 31–40, 2001. Citado na página 19.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. *Journal of computational science*, Elsevier, v. 2, n. 1, p. 1–8, 2011. Citado 3 vezes nas páginas 19, 24, and 30.

BRODY, S.; ELHADAD, N. An unsupervised aspect-sentiment model for online reviews. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.], 2010. p. 804–812. Citado 2 vezes nas páginas 27 and 28.

CHENG, M.-H.; WU, Y.-C.; CHEN, M.-C. Television meets facebook: The correlation between tv ratings and social media. *American Journal of Industrial and Business Management*, Scientific Research Publishing, v. 6, n. 03, p. 282–290, 2016. Citado 3 vezes nas páginas 19, 30, and 35.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: ACM. *Proceedings of the 12th international conference on World Wide Web*. [S.l.], 2003. p. 519–528. Citado 2 vezes nas páginas 27 and 28.

DODDS, P. S.; DANFORTH, C. M. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of happiness studies*, Springer, v. 11, n. 4, p. 441–456, 2010. Citado na página 28.

GONÇALVES, P. et al. Comparing and combining sentiment analysis methods. In: ACM. *Proceedings of the first ACM conference on Online social networks*. [S.l.], 2013. p. 27–38. Citado na página 28.

HADDI, E.; LIU, X.; SHI, Y. The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, Elsevier, v. 17, p. 26–32, 2013. Citado na página 36.

HOFMANN, T. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2000. p. 914–920. Citado na página 28.

KING, I.; LI, J.; CHAN, K. T. A brief survey of computational approaches in social computing. In: IEEE. *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. [S.l.], 2009. p. 1625–1632. Citado 2 vezes nas páginas 22 and 27.

LAMPOS, V.; CRISTIANINI, N. Tracking the flu pandemic by monitoring the social web. In: IEEE. *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. [S.l.], 2010. p. 411–416. Citado 2 vezes nas páginas 19 and 30.

LIU, B. Sentiment analysis and subjectivity. *Handbook of natural language processing*, v. 2, p. 627–666, 2010. Citado 3 vezes nas páginas 19, 27, and 29.

LIU, T. et al. Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, Springer, v. 75, n. 3, p. 1509–1528, 2016. Citado na página 30.

OBAR, J. A.; WILDMAN, S. S. Social media definition and the governance challenge: An introduction to the special issue. 2015. Citado na página 21.

O'CONNOR, B. et al. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, v. 11, n. 122-129, p. 1–2, 2010. Citado 2 vezes nas páginas 19 and 30.

POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. In: *Natural language processing and text mining*. [S.l.]: Springer, 2007. p. 9–28. Citado 2 vezes nas páginas 27 and 28.

SCAFFIDI, C. et al. Red opal: product-feature scoring from reviews. In: ACM. *Proceedings of the 8th ACM conference on Electronic commerce*. [S.l.], 2007. p. 182–191. Citado na página 29.

SCHOUTEN, K.; FRASINCAR, F. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 3, p. 813–830, 2016. Citado 5 vezes nas páginas 19, 24, 26, 28, and 29.

STROHMAIER, M.; WAGNER, C. Computational social science for the world wide web. *IEEE Intelligent Systems*, IEEE, v. 29, n. 5, p. 84–88, 2014. Citado na página 24.

THAKKAR, H.; PATEL, D. Approaches for sentiment analysis on twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*, 2015. Citado na página 24.

THOMAS, G. M. Building the buzz in the hive mind. *Journal of Consumer Behaviour*, Wiley Online Library, v. 4, n. 1, p. 64–72, 2004. Citado 2 vezes nas páginas 19 and 30.

WIJAYA, D. T.; BRESSAN, S. A random walk on the red carpet: rating movies with user reviews and pagerank. In: ACM. *Proceedings of the 17th ACM conference on Information and knowledge management*. [S.l.], 2008. p. 951–960. Citado 2 vezes nas páginas 27 and 28.

ZHANG, W.; YU, C.; MENG, W. Opinion retrieval from blogs. In: ACM. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. [S.l.], 2007. p. 831–840. Citado 2 vezes nas páginas 27 and 28.

ZHAO, Y. et al. Generalizing syntactic structures for product attribute candidate extraction. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.], 2010. p. 377–380. Citado na página 28.

ZHOU, X. et al. Sentiment analysis on tweets for social events. In: IEEE. *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. [S.l.], 2013. p. 557–562. Citado 3 vezes nas páginas 24, 27, and 28.