The influence of product description characteristics on sales
Example of a video game digital distribution service Steam

Master thesis
Programme: Data Science and Marketing Analytics

Student: Sofiia Rossovskaia
Student number: 570460

Supervisor: Michel van de Velden
Second assessor: Vardan Avagyan

Erasmus University Rotterdam
Erasmus School of Economics
2022

[The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.]

Abstract

This master thesis investigates the influence of the product's description characteristics on commercial success. Since the topic is quite broad, the research focuses more on one specific segment. This segment is the video games industry: large, prosperous and developing even more during coronavirus pandemic.

In order to understand the connection between description characteristics and sales, this master thesis uses the video games descriptions from the online distribution platform Steam. The text corpus was collected, cleaned, and used to understand emotional polarity and subjectivity, emotions expressed, and topics raised in the descriptions.

After it, this research uses Random Forest and Support Vector Machines models to predict video games' commercial success using features generated from descriptions and basic product information. Text-related features, such as description's subjectivity and the share of fear and surprise emotions, are significant, as well as the game's price and release date.

Table of content

1. Introduction

In the digital and social media era, the video games industry stays active and attracts more and more customers. From 2015 the number of gamers worldwide was growing steadily, and the forecasts until the end of 2023 are optimistic as well (Statista, 2021). The industry develops and improves every year. Large companies present next-generation consoles (e.g., Xbox Series X and Series S and PlayStation5 in 2020) and devices. Some are real game changers (for instance, virtual reality headsets have introduced qualitatively new user experiences). New technologies and game engines appear on the market, allowing developers to make their products more complicated and visually attractive.

Another factor of crucially growing industry's success is the coronavirus pandemic. At the very beginning, when regulations were stringent, and a lockdown was the most common way to stop the virus from spreading, people spent plenty of time at home. This way, there was a necessity to entertain yourselves. Streaming services and online cinemas could take the most from this situation, increasing the number of customers and, of course, revenues. For example, Netflix's net income grew by 48% in 2020, compared to 2019 (Statista, 2021).

When people have watched many films and series (or found others low quality or dull), those who were not interested in video games started thinking about it. For instance, the sales of the console segment increased by 155% in 16-22 of March 2020 compared to the previous week (Statista, 2020). The multiplayer projects became extremely popular since they create the feeling of being together and spending time with friends and family like in previous, safe days. In 2020, such games as *Among Us* (online analogue of a well-known mafia game but with space entourage), *Fall Guys* or *Animal Crossing* became viral. The multiplayer game *It Takes Two* was launched in March 2021 and achieved 1 million sold copies in one month (IGN, 2021).

Because of this abnormal demand, the revenue of the video games industry increased by 20% in 2020 compared to 2019 (Witkovsky, 2020). Netflix income growth looks more impressive at this point. However, Netflix is a relatively young company. It was founded in 1997, but the international expansion started only in the 2010s and continues even now. For instance, in Russia and CIS (Commonwealth of Independent States) countries, Netflix began broadcasting only in 2020. New customers worldwide join the service since it is novel for them. This way, the massive increase in Netflix's income can be explained by both company's growth and coronavirus. At the same time, the video games industry takes its roots from the middle of the 20th century, when the first arcade games have appeared. The video games industry is a large, mature industry, and the 20% increase in revenues is exceptional.

Regarding the size of the video games industry, it is enough to say that even in 2019, before the coronavirus changed the economic landscape, the industry’s revenue was more than both film and music industries (Statista, 2020). The global revenue of this sector was equal to $145.7 billion, while film and music got only $42.5 and $20.2 billion, respectively. These statistics show us how massive and significant the video game industry is in the modern world. This industry's personal computer (PC) generated the revenue of $33.5 billion, which is more than the whole music industry revenue and equals 79% of the film industry's revenue. The noticeable fact here is that the PC segment is the smallest in the video games industry, which brings us to the structure of this industry.

The video games industry is quite heterogeneous. This variety depends on the platforms on which people can play games. The primary platform types are personal computers, consoles and smartphones. Due to the Global Games Market Report by Newzoo, the revenue of the video games industry by June 2021 was already $175.8 billion (Newzoo, 2021). Based on this data, we can understand the share of each market segment. This way, PC has the smallest percentage of 20% ($35.9 billion), console segment takes the dual role with 28% ($49.2 billion), and the mobile segment is the leader with 52% market share ($90.7 billion). The potential of the mobile segment is enormous: people can take a smartphone or a tablet wherever they want and play some game in a queue, traffic jam or during a boring conversation. This market part generates enormous revenues and develops rapidly as new smartphone models appear.

The PC segment is still the oldest and the most attractive for video games developers. Due to the State of the Game Industry Report from the 2020 Game Developers Conference, 56% of developers are currently working on projects for PC (GDC, 2020). Also, there is a difference between PC and mobile products in terms of gameplay and perception. Since smartphones have much smaller screens than personal computers have, the content is susceptible to image quality. Moreover, PCs have keyboards and computer mice that allow the introduction of more complicated gameplay techniques. The final distinction lies in the perception layer: mobile games usually have short and simple or no storyline; the primary goal of such projects is to kill time. PC games (and console ones) often contain a serious story consuming work of screenwriters and narrators, composers and fashion designers. Such projects can get be indeed called art (Forbes, 2015). Considering all mentioned factors, we choose the PC video games market as the object of this research.

There are about six platforms where people can buy games for their personal computers. This number is already quite confusing for a new user who has never purchased a video game. The situation becomes even more complicated when we mention that buying a product on the developer's website is also possible. However, a strong leader exists here, being well-known and easy to find. In 2003, Valve released a new digital distribution service — Steam, the leading PC games online store. 60% of video game developers make at least some money selling their product on this platform, 34% of developers receive the most of their income from Steam and not other services (GDC, 2020).

Of course, competition exists in the video games market for PC. In 2019, the situation between Steam and Epic Games Store (EGS) became unstable. Some games (such as *Metro Exodus*) were released exclusively in EGS, and the game's page in Steam was bombed with negative reviews (PCGamer, 2019). Later this year, *Detroit: Become Human*, previously an exclusive project for PlayStation, was also presented in EGS (VentureBeat, 2019). Both these games were available on Steam, but much later, and many unwilling to wait for release gamers created an account in Epic Games Store. Finally, EGS offered some additional bonuses for both gamers and developers. Each week the platform gave away one or two games from the store for free. The percentage EGS takes from the developer's revenue is only 12%, which is extremely attractive compared to Steam's 30%.

In 2021 Steam remains the essential digital distribution service in the PC video games industry. Even since 40% of developers are entirely sure that Epic Games Store will achieve long-term commercial success and 75% of developers believe that Steam should decrease the cut, Steam is still the most stable option.

Considering everything mentioned before and the fact that Steam is an old platform with almost twenty years of history (compared to three years old EGS), we decided to narrow PC games market research to the games from Steam.

As of September 2021, Steam presents 103,700 games in various genres such as action, indie, casual, simulation, etc. As a company, Steam does not disclose its information a lot. This way, on the game webpage, a user can find information about the game's price, genre, developers, system requirements, reviews, ratings but nothing regarding the game's sales. This information is available only to developers and Steam itself. However, there is an opportunity to access the estimation of how many people owned the game. Website SteamSpy provides this information and estimates the intervals, randomly accessing public user's accounts.

Knowing this preliminary information about the video games industry, we can return to the coronavirus case. As a consequence of the pandemic, many people came into the industry, searching for entertainment and new experience. These new users can have some knowledge about the most trending games or no information at all. As they go to the game's web page on Steam, their decision will be based on screenshots, trailers, reviews, and descriptions. Screenshots show the potential buyer how the project's graphics look like, what is the general style. Trailers allow understanding the gameplay and mechanics that players will face in the game. Reviews display the subjective opinions of people who already played this game, and descriptions explain the game's story. The story, which will become the gamer's own story for several hours. Here, developers can reach the potential players, catch them with the specific keywords, introduce them into the game's atmosphere, and enhance screenshots and trailers. Text descriptions make the whole game's story complete.

Steam provides two kinds of descriptions. The first one is short, appearing at the top of the game's webpage. It usually contains the most concise summary, the main idea. That is the first hook where the developer can catch a gamer. If the gamer goes further, he finds the second description lower on the page. This one contains much more information about the story, the game features and specifics. The second description can encourage the user to buy the game if he is still doubting after reading the first one.

1.1 Research question

The influence of the product's descriptions becomes the research's main object. This master thesis tries to answer the following *research question*: what aspects of product’s description (separate keywords, emotions or topics) give insights into the product's commercial success and what are these insights? To answer this question fully and consistently, the following sub-questions will be helpful:

- Are there any differences in descriptions characteristics (such as keywords/sentiment/topics) between genres?
- How can we measure the atmosphere of video games descriptions, and what insights on video game’s commercial success can be found using this measurement?
- Do the text characteristics of short and full descriptions (emotions shares, subjectivity, and emotional polarity) match or not?
- Is there any difference in emotional polarity and subjectivity between different groups of games? Can this difference be used to answer the main research question of this master thesis?
- Can we extract the meaningful topics from the corpus of video games descriptions? Are these topics different for short and full descriptions, and can they help to explain video games sales?

This research uses game descriptions to analyze game features and characteristics without looking at screenshots and trailers or playing the game. This way, we provide insights on how gamers perceive text information and how it influences their buying decision-making process.

1.2 Managerial relevance

There are several reasons why this research is meaningful in terms of managerial relevance. Firstly, Steam as a distribution platform and game developers could use this research's results to adjust games' descriptions and make them more attractive for customers. Knowing which topics are catchier for which genres, managers can optimize the description writing process and obtain additional revenues.

Secondly, the results of this master thesis will be a good base for further research from video games developers. For instance, which topics are more attractive for male and female gamers? Which keywords allow to describe a horror game in the most frightening but at the same time the most appealing way?

1.3 Academic relevance

This thesis relies on three streams of literature. These streams represent the main fields of knowledge on which the whole research bases and tries to improve and explore more deeply.

<u>The video game industry</u>. This literature segment gives information on how the industry functions, gamers' perception and curiosity, and the market situation and reviews. However, there is a lack of description analysis in this segment. This master thesis can expand the boundaries of the video games industry study by introducing more relevant information regarding the description effect on video games sales.

<u>Consumer behavior and product features</u>. Here we describe which product characteristics are more important for potential customers and focus more on physical goods. However, researchers do not study and describe consumer behavior in digital markets so much. This master thesis tries to take a look into consumers' incentives in the digital market of video games.

<u>Text analytics and topic detection</u>. This literature stream is the most technical one. Articles in this segment present text mining approaches, sentiment analysis and topic detection. This master thesis can introduce how all these or similar techniques work with video game descriptions data.

1.4 Structure of the thesis

This master thesis has the following structure. Section 2 presents theoretical concepts, literature review, and hypothesis, Section 3 describes the methodology, including data collection techniques, text analysis approaches, and models. The data description with exploratory data analysis and model results are presented in Section 4. Finally, Section 5 provides a discussion of this research's results and managerial and academic implications. The references are presented in Sections 6 of this thesis.

2. Theory

This section provides an overview of the literature connected to the questions of the video game industry, descriptions features, and topics extraction from text data. After it, the hypotheses are formulated, based on the theory from presented articles and researches.

2.1 Literature review

The literature review comes in line with the streams mentioned in the Academic relevance sub-section. This way, it starts with the articles regarding the video games industry.

A critical concept of human life, in the most general, and video games perception, in particular, is curiosity. Researchers from Dutch universities explored curiosity triggers in the video games industry in 2019 (Gómez-Maureira, Kniestedt, 2019). The results of the online survey conducted during the research show that games balancing between uncertainty and structure invoked the most curiosity among players. The second noticeable idea is that two genres — Exploration and Social Simulation — made gamers more curious than others. Summing up, there are two game features that the research can project to game descriptions as well: balance between structure and uncertainty and specific genres.

Regarding the genres themselves, the classification of the video games according to genres can be pretty complicated (Clarke et al., 2017). The genres can overlap in projects' characteristics, and the classification focus can change. Different researchers or game developers can assign various meanings to the same genre since they look at it from different perspectives. Genre's classification of Steam, used in this research, is not perfect as well: many games relate to the action genre just because they require fast reaction and active player's actions. However, it is still used since even not being perfect allows structuring thousands of video games.

Another point regarding the action genre and the industry itself is that belonging to this genre increases the game's sales (Sacranie, 2010). The second factor of influence is the review score. Gamers are interested in others' experiences before the purchase. Consequently, the tone of users' reviews can influence sales significantly.

The second literature stream relates to the features that consumers find attractive in the product and the product's description. The articles describe mainly physical products; however, some ideas and results can be adapted and used for such digital products as video games. The first article in this stream relates to wine products marketing (Bruwer et al., 2011). Researchers have found that even since many customers read labels on the wine bottles, it is hard for them to match described features with actual wine's characteristics. This way, customers surveys and questionnaires can be biased. Moreover, researchers extracted ten buying influencing factors and compared their importance between gender and age groups. One of the factors was label information, and it took a 6 out of 10 importance score, showing that a product's description is important for consumers.

The product's description is even more critical for online stores and digital products. It contains all needed information (and sometimes even more), allowing the creation of various features from it (Wang et al., 2018). These features can sometimes be not obvious or hidden, and the researchers should use sophisticated text analysis techniques to extract them. However, since these features are extracted, they can improve understanding of customers' behavior and choice.

Another article analyses the online product descriptions from cannabis retailers (Luc et al., 2020). The research idea was to find how cannabis products are described on the retailers' websites. The research used data from 27 retailers regarding 428 products. After it, all descriptions were manually coded, giving 38 codes in 6 categories. These categories describe different product features relating to cannabis effects, diseases treatment and side effects. As a result, the most frequently mentioned characteristics (relaxation, fruit taste or smell and pain treatment) were found, explaining retailers' presentation of their products. There is a possibility to extend this approach. Firstly, the categories can be extracted not manually but using advanced algorithms such as Latent Dirichlet Allocations or Non-Negative Matrix Factorization (mentioned further in the third literature stream and the Method section). Secondly, these topics can be used not only to understand the current situation but also for prediction tasks.

The third literature stream is the widest one since it describes different approaches to text data analysis: from sentiment analysis and topic extraction to modelling obtained features. Sentiment analysis is one of the most commonly applied to text data approaches. It allows (based on machine learning techniques or sentiment lexicon) classifying documents (text units) according to their emotional polarity (Medhat et al., 2014). Usually, there are two categories: positive and negative polarity, but sometimes neutral category appears. Mainly the lexicon-based approach is used since it is less computationally intensive and faster than the machine learning approach. However, there are also hybrid approaches, combining advantages of both techniques (Prabowo, Thelwall, 2009). The idea of sentiment analysis can be extended from understanding polarity to emotions identification. In this field, both machine learning and lexicon-based approaches are used. Extracted emotions give a better understanding of data and improve further prediction models with new features (Alm et al., 2005).

Topic extraction is a powerful method of text data analysis that allows finding hidden patterns in data. There are two main approaches — Latent Dirichlet Allocation and Non-Negative Matrix Factorization. LDA is a probabilistic model (Blei et al., 2003), and NMF works based on linear algebra and matrixes decomposition (Lee et al., 1999). Despite the differences in the methodology, both approaches give results in the same format: extracted topics with the set of words mainly associated with each topic. Comparison of the accuracy of LDA and NMF on short texts shows that NMF usually gives better predictions (Chen et al., 2019). The research used the following datasets to prove this hypothesis: snippets of data from Google, data from news websites (Reuters, USA Today, XinlangNews, etc.) and StackOverFlow. Both LDA and NMF models are described in more detail in the Method section.

2.2 Hypothesis

The hypotheses of this master thesis arise from its research sub-questions. They will be checked and either rejected or kept during the research, giving a better understanding of data and its patterns.

The first hypothesis (H1) states that there are differences in description characteristics between genres. There are topics which are specific for particular genres. Also, we expect that the share of anger and surprise emotions will be higher for descriptions of the video games in action genre.

Atmosphere and emotional tone of the text are hard to measure. The second hypothesis (H2) stated that there is a possibility to measure this factor, express it in one (or more) variables and use for prediction of the video games commercial success.

Each game on Steam has two descriptions: short and full. The third hypothesis (H3) states no difference in text characteristics between these two types of descriptions. Clarifying this idea: the length of descriptions can vary (since one of them is always more detailed), but such features as emotions, subjectivity and emotional polarity are the same.

The forth hypothesis (H4) states that emotions reflected in the description and the level of description’s subjectivity (how emotional and catchy the text is, how many modifiers (*very, a lot*, etc.) are used) positively and significantly influences the commercial success of the video game.

Finally, the fifth hypothesis (H5) relates to the topic extraction procedure. There are meaningful, logical topics that can be extracted from short or full descriptions and be used in the video game's success prediction.

3. Method

This section introduces the methodological part of this research. It explains the technical aspects of data collection and cleaning, feature engineering and model construction.

3.1 Data collection

There are several methods of collecting data from websites, and all of them in some way work with the websites' API. The difference between these methods is the amount of code the researcher needs to write and the difficulty of the webpage's structure that the method can handle.

The first group of methods requires the usage of specific Python packages such as BeautifulSoup or Selenium. The time spent on the data collection process increases for this approach, but the result data can be cleaner from unusual or wrong observations (e.g., there is no data scraped from wrong fields of the website).

The second possibility is to work directly with the website's API, writing the functions from scrape without any predetermined packages. This approach works nicely with well structured, simple webpages and allows to save some additional time.

The last data collection technique requires the WebScraper extension for the Google Chrome browser. WebScraper does not need any coding, and it is simple and easy to learn. Moreover, it allows scraping some information that is not available with two previous methods. The trade-off here lies between the simplicity and the quality of the data since WebScraper can mix up some data objects in the webpage.

In the master thesis we used two of the above-mentioned methods. The main part of information was collected using WebScraper extension on Steam website. The estimation of sales figures from SteamSpy was collected be direct access to the website’s API.

3.2 Data preprocessing

Several points should be checked during the **data preprocessing** step. First of all, abnormal and suspicious values should be dropped. Also, observations that were collected but are not helpful for research should be deleted. Secondly, the dataset should be cleaned from missing values: these should be either replaced with mean or median values or deleted. Finally, the data should be checked for outliers: abnormally large or small values should be found and analyzed. After it, these outliers should be deleted or replaced.

Additionally, the text data requires more attention than numerical one since it is more varied and contains more noise. Text data can contain emoticons, grammar, and lexical mistakes, which numerical data does not contain. The text is hard to analyze and to make measurable. Many steps of text data preprocessing exist to make this process more efficient. First of all, if the data was scraped from the website with international content, only documents in analyzed language should be used. Usually, this language is English since it is officially international and many text analytical tool are available preferably for English language. Then, stopwords should be deleted from the analyzed text. These are extremely frequently used in the language but do not have a powerful sense on their own (e.g., articles, personal pronouns or forms of the verb "to be"). There are publicly available stopwords lists, such as Porter's Stopwords Corpus, which contains 2,400 words for 11 languages, including English. The following step is removing special characters (e.g., punctuation marks), HTML tags that were occasionally scraped, accented characters. Then, the contractions are expanded, and the text is transformed to lowercase.

The final steps are lemmatization and stemming. These two processes can look similar at first, but the difference is in the approach. Lemmatization is the process of word reduction to its original form, stated in dictionaries. This form is also called the word's lemma. The stemming is faster and gives not such a meaningful result. The whole idea of stemming is to drop a word's affixes and keep the word's stem which can be not always lexicographically correct.

There are different **stemming** techniques, such as Porter’s, Lancaster, or Snowball stemmers. Porter’s stemmer is one of the most commonly used, and here is the main idea of the process. The algorithm was invented by Martin Porter and appeared for the first time in the article published in 1980 (Porter, 1980). Now Porter’s stemmer is available in the *nltk* Python package. The stemmer has five phases, where the first one is the easiest: here, the algorithm drops plural and past participles suffixes of words. The phases from the second to the fifth are more complicated and depend on the following formula:

$$[C](VC)^m[V],$$

where *C* is the notation for one or more consonant strings, *V* is the notation for one or more vowel strings (which are letters *a*, *i*, *o*, *e*, *u* and *y* if there is no consonant letter before it), and *m* is the measure of any word or word part. Square brackets mean that these parts should not necessarily be present in the word, and $(VC)^m$ means that *VC* part is repeated m times in the word. This way, the phases 2-5 depend on *m* or, in other words, on the word’s length, and based on this factor, the same suffix can be replaced with different suffixes. As a result of applying these five phases, the researcher obtains the word’s stem that can be used in further analysis.

3.3 Feature engineering

Various techniques were used for feature engineering in this research. These techniques extracted the following features from the text data: emotions, topics, emotional polarity, and subjectivity.

3.3.1 Emotions extraction

Extraction of emotions from text data is quite a complicated task. There are not so many publicly available packages that could provide such functions and open code, allowing the understanding of the algorithm. The *text2emotions* package is available for Python and has open and detailed documentation. The *get_emotion* function works with the following logic: looking at the emotional sentiment of each word in the document (which is predefined in the dictionary), it shows the shares of five primary emotions presented in this document. These emotions are happiness, anger, sadness, surprise and fear. As a result, the researcher has a vector with the proportion of five emotions for each document in the dataset. Emotions with the larger proportion are more present in the document (there are more words relating to these emotions), and all proportions sum up to 1. Possibly, there can be a sum equal to 0 if there is a neutral emotional sentiment on the text.

3.3.2 Topics extraction

Topic extraction is a powerful, unsupervised technique that allows finding unexpected connections that were previously hidden in the data and using them in further modeling and research. Also, documents in the sample can be clusters based on the main topic highlighted in the document. There are several approaches to topic extraction, and the most well-known are Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Here, the we shortly explain the main idea of both approaches.

Latent Dirichlet Allocation algorithm was introduced in 2000, and then firstly applied in the machine learning field in 2003 (Blei et al., 2003). LDA is a generative probabilistic model of a corpus, and its main idea is to find the probability of belonging into specific topic for each word in the corpus. The main assumptions of this algorithm are as follows:

- The order and grammatical form of words in a document do not matter: the LDA algorithm analyses bag-of-words, without connections between these words.
- Documents should be cleaned from stopwords in their classical meaning (articled, personal pronounce, etc.) and from words which occur in the documents with a high frequency and do not have additional meaning. For instance, in data science related documents such words as “variable”, “model” or “analysis” should be dropped.
- The number of topics defined by LDA is known and fixed by the researcher.

The algorithm of Latent Dirichlet Allocation can be described in two steps. Firstly, for each word *w* in each document d one of the k topics is assigned randomly. Then, for each word in each document, two probabilities are checked:

- $p(topic\ t|document\ d)$: if the proportion of words (other than *w*) belonging to the topic *t* in the document d is high, then, the word *w* probably belongs to the topic *t* as well.
- $p(word\ w|topic\ t)$: if the proportion of documents assigned to topic *t* because of the presence of word *w* in these documents is high, then, the word *w* probably belongs to the topic t. LDA represents all documents as separated, independent words with some

probability of belonging to specific topic. This way, words with strong probabilities will influence the topic definition for the whole document.

This way, the final definition for the probability of the word w being associated with the topic k is: $p(w\ belongs\ to\ t) = p(topic\ t|document\ d) * p(word\ w|topic\ t)$.

Moreover, there is online learning for LDA available, introduced in 2010 (Hoffman et al., 2010). This algorithm is based on online stochastic optimization with a natural step, and its main advantage is that it allows analyzing huge number of documents in a relatively small time.

As a result, the LDA algorithm gives k topics which are sets of words with the highest probabilities of belonging to this exact topic. Through multiple iterations, the algorithm achieves stable and meaningful results. These words can be chosen either by number (for instance, ten words with the highest probabilities) or by some threshold (all words that have probability higher than 60%). Also, for small corpus all words can be displayed.

Non-Negative Matrix Factorization is a specific case of matrix factorization technique, where the matrices elements are restricted to be non-negative. And this is the main assumption of the NMF algorithm. The method was previously mainly used in chemistry data analysis, and then it became highly used in different fields after 1999 (Lee & Seung, 1999). The key idea of NMF is to obtain a low-dimensional representation of high-dimensional vectors. This approach is frequently used in recommendation systems construction, where sparse data is a usual thing. Regarding the topic extraction in text data case, for which NMF is also commonly used, the problem of sparse data is still actual. NMF works with a word-document matrix *V*, where rows represent words and columns represent documents. If a word appears in a document, there is 1, otherwise there is 0. There should be such matrixes W and H that represent the original matrix V in a way $V = WH$, where W contains topics extracted from the corpus and H contains the weights of these topics in original documents. The optimization process of the NMF method is described with the following objective function:

$$\frac{1}{2}||V - WH||_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(V_{ij} - (WH)_{ij})^2,$$

where *F* represents the Frobenius or Euclidian norm, *m* is the number of documents and *n* is the number of words.

As a result of NMF application, the researcher gets the topics and words, belonging to these topics with weight. The higher the weight of the word is, the better is its representation of the topic.

3.3.3 Sentiment analysis (emotional polarity and subjectivity)

The sentiment of the text data is often the key feature for the research with text data. Usually, the concept of sentiment is associated with the emotional polarity of the document, which means is the general mood of the document positive or negative (sometimes neutral category appears as the third option). Sentiment analysis of customers' reviews allows understanding which products gain best reactions or which product's features appear more in positive feedback.

Sentiment analysis in Python is presented with the *TextBlob* package. The *sentiment* property of this package returns two scores for the document which are polarity and subjectivity. Polarity lies in the range [-1; 1], where -1 means that the document's sentiment is completely negative, and 1 means it is completely positive. Subjectivity belongs to the range [0; 1], where 0 means that the word or document is objective, and 1 means it is completely subjective.

The calculation of these scores is quite intuitive. The package contains the *en-sentiment.xml* file, the dictionary, which provides information regarding polarity, subjectivity, and intensity of English language words. However, some words have different meanings that affects these three parameters. This way, to find the final polarity and subjectivity of a word the scores for each meaning are averaged. Then, if there is a negation (such words as "no" or "not") before the word, its polarity is multiplied by -0.5 and subjectivity remains the same. If there is a modifier (for instance, "really"), polarity and subjectivity are multiplied by the intensity of modifier, stated in the dictionary. Finally, if there are both negation and modifier, the word's polarity is multiplied by $-0.5 * \frac{1}{intensity\ of\ modifier}$, and subjectivity is multiplied just by $\frac{1}{intensity\ of\ modifier}$. Final polarity and subjectivity of the document are calculated as the average of scores for words and phrases in this document.

3.3 Model

When the data is collected and additional features are created, the research moves to the prediction part. This thesis tries to classify the video games into different sales categories and uses two machine learning models to achieve this goal. These models are random forest (RF) and support vector machine (SVM), powerful approaches, making classification tasks easier to solve.

Random forest is the model based on several less sophisticated decision trees (Breiman, 1999). This approach is the bagging of decision trees when each decision tree acts on a random sub-sample of features. The significant advantage of RF is that it reduces the risk of overfitting, which is a massive problem of the DT approach. However, it is relatively more challenging to interpret.

The algorithm of RF can be described in the following way. For each $k = 1, \dots, N,$ where N is the number of decision trees, a bootstrap sample X_k is created. Then, the decision tree b_k is built on the X_k sample. Splits of the tree are created based on the criterion (which can be entropy, misclassifications error or Gini). The splitting continues until the tree reaches the maximal allowed depth or the minimal number of observations in the leaf, n_{min}. The set of m features for each decision tree split is randomly chosen from d original features.

For classification tasks the final prediction is chosen with the majority voting rule, for regression tasks — the mean of decision trees predictions is used. The optimal *m* for classification models is usually considered as $\sqrt{d}$ and for regression models it is $^{d}/_{3}$.

The idea of **Support Vector Machine** (SVM) model is to divide observations into classes with a separating hyperplane (Boser et al., 1992). In p dimensional there is a flat affine subspace (hyperplane) of dimension $p-1$. This way, in two-dimensional space the hyperplane will be a line and in tree-dimensional space it will be a plane. This can be shown with the following expression:

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = 0,$$

where $\beta_1, \beta_2, \dots, \beta_p$ are parameters and $X = (X_1, X_2, \dots, X_p)^T$ is a point in p-dimensional space. Since *X* satisfies the equation, it lies on the hyperplane.

The algorithm of SVM is a bit more complicated, than the hyperplane itself. The hyperplane has margins around it that can and should be maximized. The whole optimization task of the classifier with soft margins, meaning that some misclassifications are allowed, is described below:

$$\underset{\beta_0,\beta_1,\dots,\beta_p,\epsilon_1,\dots,\epsilon_n}{\text{maximize}} \quad M,$$

$$subject\ to \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) \geq M(1-\epsilon_i)\ \forall i = 1, \dots, n$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Since the optimization task is described mathematically, the explanation of the variables should be given. *M* is the margin's width, $\epsilon_{1,\dots,}\epsilon_n$ are the slack variables representing misclassification. If the *i*th observation lies on the correct side of the hyperplane $\epsilon_i = 0$, if it lies on the correct side of the hyperplane but on the wrong side of the margin $\epsilon_i \in (0; 1)$ and if it lies on the wrong side of the hyperplane $\epsilon_i > 1$. Finally, *C* is the non-negative regularization parameter of the model. Smaller value of *C* gives wider margin that results in low bias and high variance. For higher values of *C*, the margin becomes narrower, resulting in high bias and low variance. This way, *C* parameter should be tuned to achieve the optimal point.

The most important part of SVM model is the kernel. Some data can be not linearly separably with the hyperplane. Possible solution here is to transform the data into higher dimension, where it can be separated. However, this procedure can be extremely computationally expensive. At this point appears the kernel trick. It allows providing dot product of transformed vectors from the initial dimension without really transforming it.

The kernel can be simply linear (in this case the model is called Support Vector Classifier, SVC), which can be described as follows:

$$K(x_i, x_i') = \sum_{j=1}^{p} x_{ij} x_{ij}',$$

where x_{ij} and x_{ij}' are classified observations. SVC works with linear data and with two classes tasks. However, researchers do not always work with perfectly linear data and there can be more than two categories. To solve these tasks, such kernels as polynomial and radial basis function (RBF) appear.

Polynomial kernel is more general representation of the linear one:

$$K(x_i, x_i') = (1 + \sum_{j=1}^{p} x_{ij} x_{ij}')^d,$$

where *d* is the degree of the polynomial. This kernel already can be used with multiclass tasks and non-linear data (opposed to the linear one) since the data is transformed to nother coordinates; however, it is usually the least effective from non-linear kernels.

RBF kernel is also known as exponential one. It divides data with a circle, that can be seen from the second degree in its formula:

$$K(x_i, x_i') = exp(-\gamma \sum_{j=1}^{p} x_{ij} x_{ij}')^2,$$

where γ is a parameter that belongs to the interval from 0 to 1. This parameter is usually set to 0.1 or to $\frac{1}{number\ of\ features}$, but it can also be tuned to achieve higher accuracy.

There are two main approaches to handle multiclass data with SVM. The first one is the one-versus-one classification approach. There are $\frac{N(N-1)}{2}$ classifiers (where *N* is the number of classes) that compare all classes, one by one. The final class is assigned to the observation with the majority of classifiers. The disadvantage of the one-vs-one approach is its computational intensity. The second method to handle multiclass data is one-versus-all classification. It requires only *N* classifiers and chooses the final class based on the maximal distance from the hyperplane. The one-vs-all approach is less computationally intensive, but it struggles with imbalanced data.

4. Results

The most intensive part starts after introducing this thesis's research topic and giving an extensive overview of literature and methodology. This section describes the application of all mentioned methods and approaches: from the data collection process to the model results.

4.1 Data description

The Data description sub-section is divided into several parts. First of all, the data collection process is explained. Then, data cleaning segment presents how the dataset size decreased step by step. Finally, the collected and cleaned data is explored and explained.

4.1.1 Data collection

For data collection, this research used two websites and two approaches mentioned in the Data collection sub-section. The primary dataset was scraped from Steam using WebScraper browser extension. Less massive but much more critical in the research context information — games' owner's number estimation — was collected from SteamSpy database using API.

Steam has an enormous number of games published, which makes the scraping process long and multistage. No laptop could collect more than 100,000 observations at a time. This way, to avoid days of computationally expensive and inefficient work, this research collected the data from Steam in tree steps. At the first step, we scraped nine medium size sets of randomly chosen games. Then, at the second step, the sets were collected for five genres, which are indie, strategy, simulation, RPG and adventure. Finally, the last two sets of video games were obtained by scraping random games which do not belong to the most supplied genres (action, adventure, casual and RPG). The second and the third steps were needed to collect representative sample of the real population: without them the distribution was skewed to the action genre.

Some datasets contained the same observations since games were chosen randomly, so all duplicates were dropped after merging the datasets into one. As a result, the Steam data consists of 62,095 video games observations. The SteamSpy database is also quite large and was scraped into four attempts. After merging and duplicates cleaning, the final SteamSpy dataset consists of 47,326 rows. Finally, these two datasets were merged by the video game's name, and the research data of 33,197 observations was obtained.

4.1.2 Data cleaning

The initial features of this research are the following: game title, price, discount (if applicable), reviews number and review tone (e.g.: Positive, Mixed, etc.), developer, release date, genre, short and full descriptions and the number of game's owners (target variable). All these columns contained some missing values. Games that were scraped without a title also did not have information for other features. These products were mostly different bundles such as several games from one publisher, game + soundtrack, or game + DLC. This master thesis aims to research games and not the bundles of games and additional products, so these observations were dropped.

The data cleaning process removed text information from the columns with numeric data (e.g.: *(99 reviews)* changed to *99*), transformed release date to the *datetime* format, and cleaned for missing values. The missing values for discount and reviews number were replaced with zeros since there are games without reviews or discounts for the data scraping moment. This research used the median release date (June 2018) for games without a release date and deleted games missing full description or information regarding developer and genre.

Finally, the *genre* column was checked. Steam provides video games but also various software for video production, game development, animation, etc. These products did not match the research goals (since they are not video games) and were dropped. Also, video games that did not have information about genre and full or short descriptions were deleted. After all the cleaning procedures, the dataset consists of 32,147 observations, meaning that 1050 observations were deleted during the cleaning process.

Since this phase is done, we applied some outliers cleaning and exploratory data analysis (EDA). These two procedures are often used at the same time, because EDA makes finding outliers easier with the help of visual presentation. The first variable was *price,* and the maximal value was €800 that is for sure outlier. This observation appeared to be the huge bundle of 58 games. The maximum price was set to €30 and only 580 observations were lost. The number of reviews also had sufficient outliers, which can be explained with the extreme popularity of these video games (the maximum value was 6,000,000 reviews written to one game). Observations with more than 1,000 reviews were dropped, since their success is defined by side factor (for instance, developer's popularity or artificially created high demand) and not by the game's description, which is the key feature of this research. 3,598 observations were dropped after the outliers cleaning.

The last step before looking at EDA results and searching for the first meaningful results was to delete games with descriptions written in other languages than English. The main language of this research is English and mostly packages for NLP work with English words as well. There were 3,658 video games with descriptions in languages other than English.

The final number of observations in the dataset after all these steps is 24,891. This way, 8,306 observations were lost due to language, outliers and missing values cleaning procedures.

4.1.3 Exploratory data analysis

Since there are a lot of variables used by this research, the analysis part is structured in a more intuitive way. Firstly, there are initially scraped variables, and secondly — variables obtained during the feature engineering process.

Initially scraped variables

Moving to the exploratory data analysis (EDA), the first feature is the game's price. Figure 1, the box plots of price per genre (prices are given in euros and genres are defined by Steam). The

Massively Multiplayer genre has the lowest average price, which corresponds to the industry's logic. The specific of this genre is that massively multiplayer games require in-game donations that are not shown in the price. This way, developers lower the price, knowing that they will receive more from donations. The most expensive (on average) genres are Sports and Simulations. We can explain the high price for sports games with the specific target group interested in sports and ready to pay more.

Figure 1. Box plots of price per genre

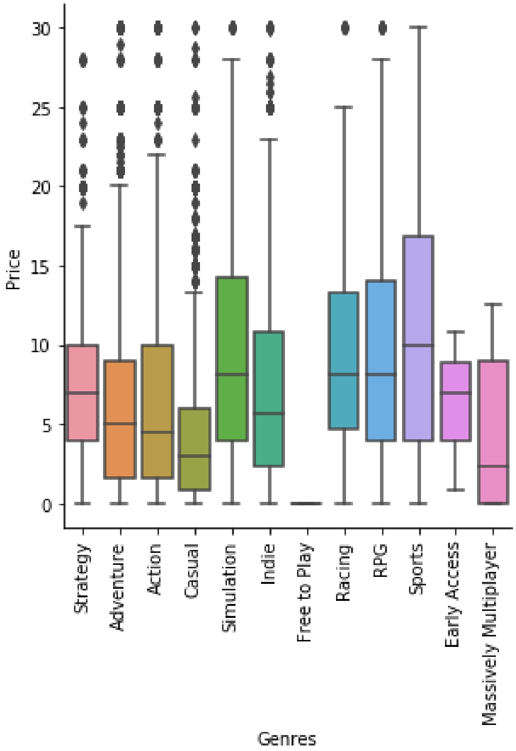
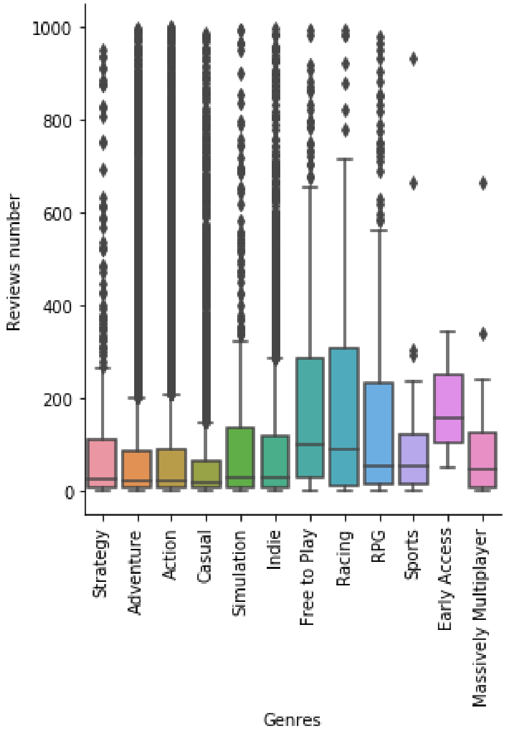


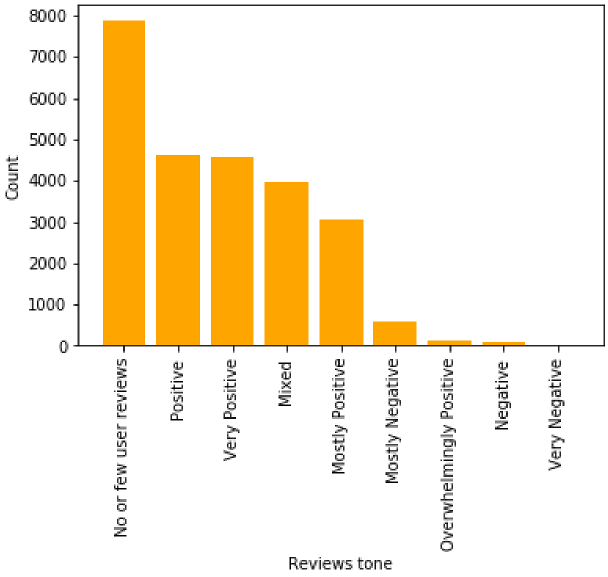
Figure 2 represents the box plots of the number of reviews written per genre. There are still some values beyond the box plots' whiskers. However, these values are normal compared to the dropped ones and deleting them would lead to a dramatic decrease in the sample size. This situation could be not a problem, considering that the dataset contains almost 25 thousand observations. Still, these outliers belong to the classes of the target variable, which are not so big, as is shown further. As a result, these outliers remain in the sample. The highest average number of reviews get games in the early access category, which is logical as well. Developers that sell their games in early access usually ask for opinions, for feedback, whether it is positive or not, and players give it to them, increasing the number of reviews. For other genres, except for free to play, racing, RPG, sports and massively multiplayer games, the average number of reviews is relatively low and lies around 50.

Figure 2. Box plots of reviews number per genre



The number is not the only feature of games' reviews, and there is also the emotional tone. Figure 3 shows that there are mostly positively reviewed projects or projects with only a few reviews in our sample. The gradation of tone from "Overwhelmingly Negative" to "Overwhelmingly Positive" is shown in Appendix 1.

Figure 3. Bar chart of reviews tone distribution



Figures 4 and 5 present the box plots for games' discounts. The difference between these two is that the first one is for all observations in the sample, and the second one shows box plots only for games with discounts that are strictly larger than 0. Before interpreting these plots, it is essential to mention that we collected the data in "usual" conditions. These conditions mean no seasonal sales during the data scraping process, and there were no outside factors influencing discounts. Figure 3 gives the following insight: in general, games on Steam have low, almost 0 discounts, and only games in early access have visible positive discounts in the box plot. This fact can refer to the developer's idea of giving significant discounts to games in early access, attracting more players and getting more feedback.

Figure 4. Box plots of discount per genre

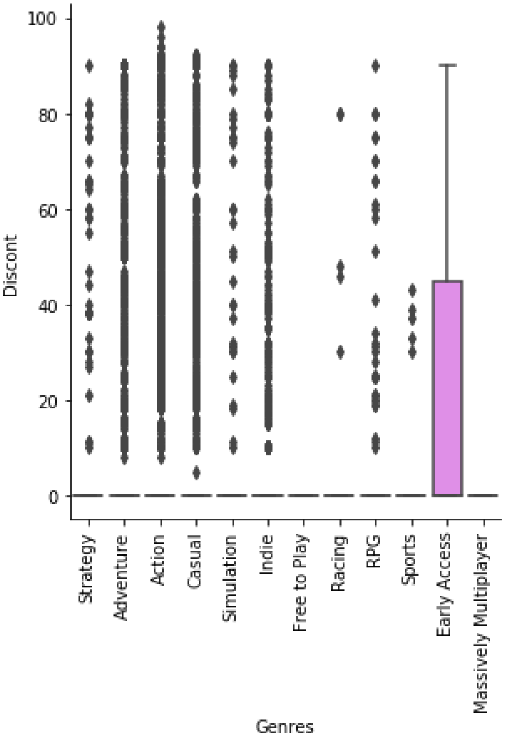


Figure 5. Box plots of non-zero discount per genre

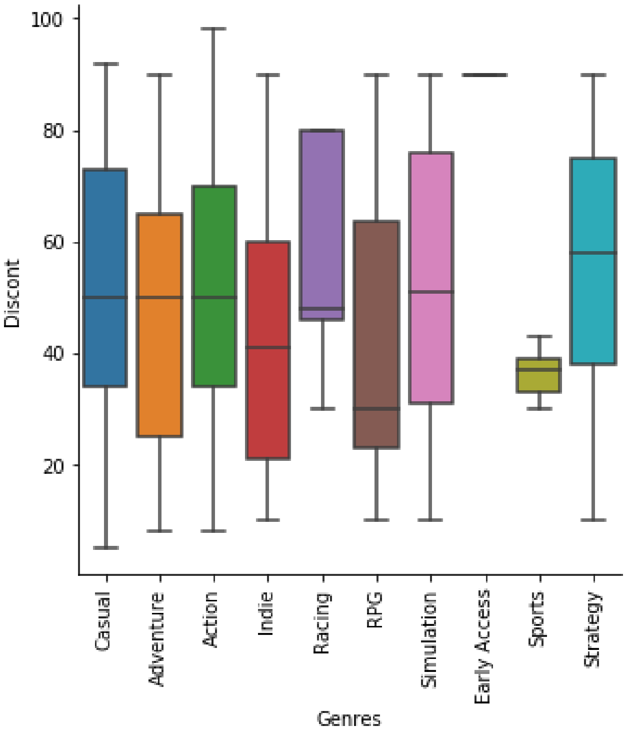
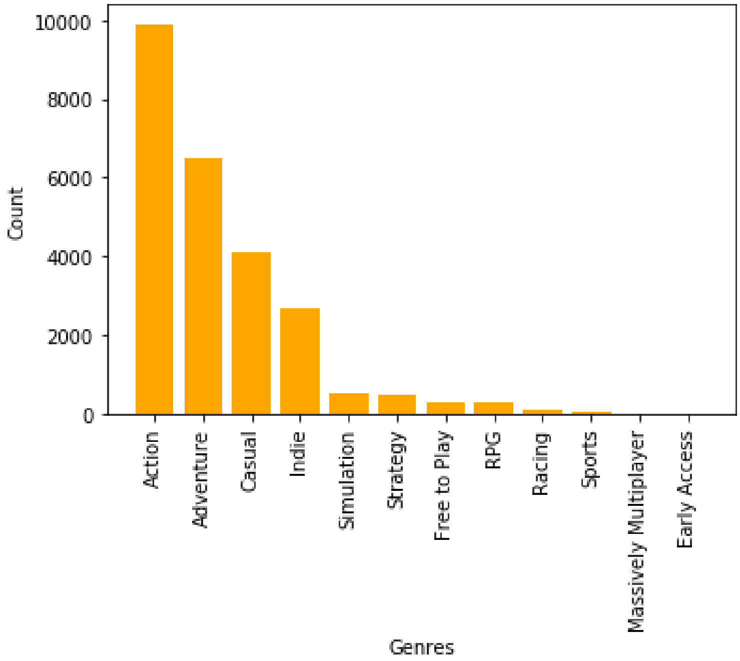


Figure 5 considers games with positive discounts, and the early access games still have the largest ones. The following genres have the lowest average discounts. RPG (since this genre is relatively not popular), sports (that matches interestingly with relatively high average price) and indie (since indie developers are small independent teams, they are not willing to give big discounts to their products). Except for games in early access, the highest discounts belong to projects in strategy and simulation genres.

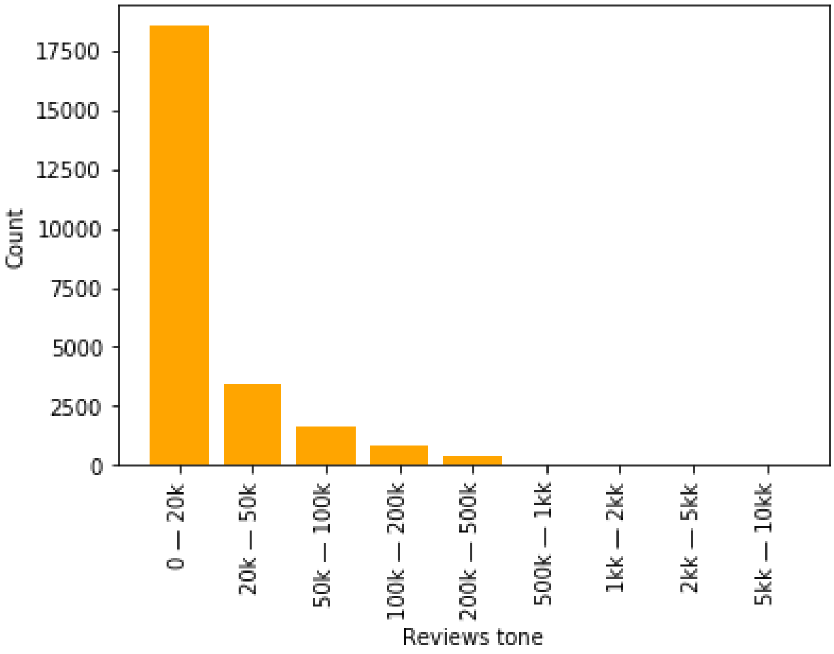
The distribution of games between genres can be seen in Figure 6. The majority of video games in the sample belongs to action, adventure and casual genres. The difference between action and other genres could be even more enormous if we did not scrape additional data from different genres as described before.

Figure 6. Bar chart of genres distribution



The last but not the least variable from the initial scraping is the amount of people owning the game. This variable is categorical, and Figure 7 shows that most video games in the sample belong to the first category, that is, below 20,000 downloads. Moreover, this bar chart gives an understanding of the enormous class imbalance in the data that should be fixed with the undersampling of the majority class and merging categories with the largest downloads.

Figure 7. Bar chart of owners groups distribution



Also, this research uses some minor features: dummy variables for specific n-grams (one or more words appearing in the description), the video game's age, and length of descriptions in words. Dummies present one or several words in the short descriptions (the short one is used following the same logic as with the topics: the short description is the first that the gamer reads).

There are the following words and combinations: *fantasy*, *team*, *first person*, novel (includes *novel* and *visual novel*), *online*, *zombi*, *retro*, *gameplay*, fight (includes *attack*, *kill*, *fight* and *enemy*), *single player*, *tower defense*, horror (includes *horror* and *horror game*) and *old school*. The presence of these n-grams in the sample is not higher than 4%, except for the *fight_n* dummy variable: it equals 1 for 14.9% of observations.

From the release date, we derived the "age" of the game in years, as for September 30 of 2021. The average is 3.5 years, and the maximum is 38 years. For several games that were not released yet, the age was set to be 0.1 years to avoid negative values. The last variables are simply the number of words in short and full descriptions. The average values are 33 and 212 words, respectively.

Shares of emotions in descriptions

Now, when the variables from the initial dataset are clear, the data description moves to variables created during the feature engineering process. Following the order from the Method section, this research shows emotions firstly, then topic extraction and, finally, polarity and subjectivity.

Using the *text2emotion* package, we extracted the shares of five primary emotions for all observations in the sample for short and full descriptions. Figures 8 and 9 show that, on average, fear has the largest share in video games. Such a prevalence of this emotion is explained with the dictionary of text2emotions package and the specificity of video games descriptions. Such words as "horror" are marked with fear emotion, and horror games are popular and multiple in the video games industry. Another reason is that this emotion in the dictionary also relates to words like "risky" or different phobias, such as coulrophobia. Many action games suggest handling risky situations, and many horror or thriller games are based on phobias.

Figure 8. Box plots of emotions shares in short descriptions

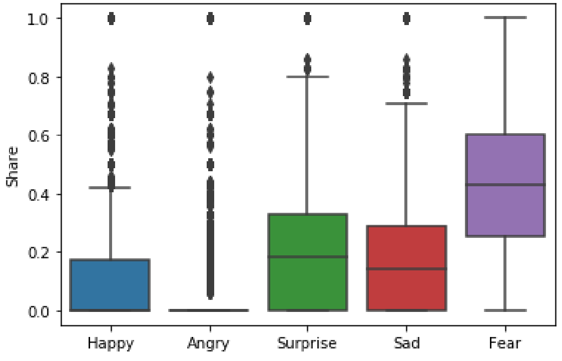
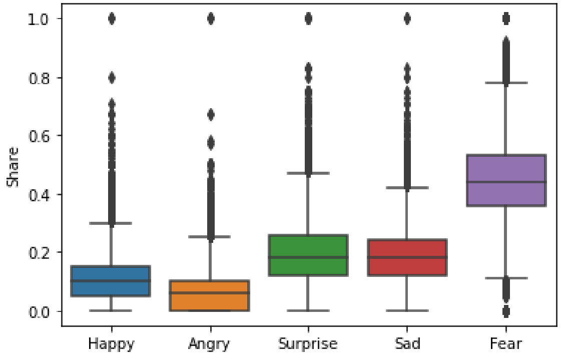


Figure 9. Box plots of emotions shares in full descriptions



The following important insight regarding emotions definition in the video games descriptions is that the emotions in full descriptions distribute more clearly. The interquartile ranges are narrower and lie further from 0 (except for angry), meaning that the shares of emotions are less volatile for full descriptions. This fact can relate to the idea that, in short descriptions, due to the small length, developers try to use as many emotional, catchy words as possible, mixing different emotions.

Regarding the difference in emotions between short and full descriptions, we found that there are no genres with a notably large share of sad or happy words. However, the other three emotions give some patterns. The first one, anger, appeared mainly in short descriptions of genres RPG, Adventure and Massively Multiplayer and in full descriptions for Sports, Action and Massively Multiplayer. For these genres, the prevalence of anger looks logical since they contain elements of competition and fight.

The second emotion, having outstanding genres in short and full descriptions, is fear. Genres with the highest fear shares in short descriptions are Massively Multiplayer, Racing and Strategy, in full ones — Strategy, Racing and Early Access. Since fear emotion in *text2emotion* package also describes risk, the prevalence of these genres looks natural. Also, two genres appear in both descriptions, meaning that fear emotion is more consistent in the video game's presentation.

The last emotion, surprise, had a visible pattern only in full descriptions, and for short ones, the average share was almost equal for all genres. The genres with the highest surprise emotion share are Adventure, Sports and Action. Considering that the Adventure genre assumes world exploration and Sports and Action games are highly dynamic, the pattern looks logical and meaningful.

These facts show that we cannot reject the H1 of this research: there is difference in descriptions' emotions between genres and Action genre games have more anger and surprise emotions in full descriptions.

Topics extracted with LDA and NMF

For topic extraction, we tried both LDA and NMF to understand which algorithm is more efficient (better separates topics from massive text data). Sparse document-word matrices were created to apply these two approaches, having 24,891 rows (documents) and 7,133 columns (words) for short descriptions and 18,928 columns (words) for long descriptions.

Latent Dirichlet Allocation realization from *scikit-learn* library for Python was used. The online learning method mentioned in the Method section was chosen as the best approach to handle huge document-word matrices.

The number of topics was checked from five to twenty-five. These boundaries were chosen because it would be difficult to represent the large text corpus with less than five topics and hard to interpret more than twenty-five. However, the results of the LDA algorithm were disappointing for both short and full descriptions. The same words representing the topic with the highest probabilities appeared in different topics (meaning that topics were not clearly separated). Moreover, topics generated by LDA did not present descriptions and their games in a good way.

For instance, one topic based on *Japanese*, *ninja* and *samurai*, was mainly presented in the casual snowboarding game, having nothing in common with Japan. This way, it was decided to move to the NMF approach.

Non-Negative Matrix Factorization is also available in the *scikit-learn* library. The same two document-word matrices were used for NMF. Specifying the method, the Non-Negative Double Singular Value Decomposition (NNDSVD) was chosen for the initialization of the procedure since it works better with sparse data. The number of topics was used from five to twenty-five, the same as for LDA.

The topics for short and full descriptions were logical and meaningful: ten topics for short and seven for full descriptions. Then, when the topics were extracted, observations were clustered according to these topics. This clustering was done with k-means approach, where *k* is equal to the number of topics. At this stage, topics for full descriptions showed much worse results than topics for short ones. There was always one topic that took around 60% of observations, and this situation could be realistic, but at each iteration, that was a different topic. Also, the results of such clustering were counterintuitive: the primary class could be presented by the topic describing puzzle game, but there are not so many of them in the sample.

Considering the results of both LDA and NMF techniques for short and full descriptions, we decided to focus on topics extracted with NMF from short descriptions. This decision also makes sense since potential players will first look at the short description, and only if this catches them, they continue with the full one. This way, successful topics of short description are more important in terms of further managerial implication.

Table 1 presents the topics extracted from the short descriptions with NMF. There are ten words (the number which is enough to understand the topic and to mot be confused with too many words) with the highest weights for each topic and the definition of the topic given by this research. The words are presented after the stemming procedure, meaning that they can be not grammatically correct.

Table 1. NMF topic for short descriptions

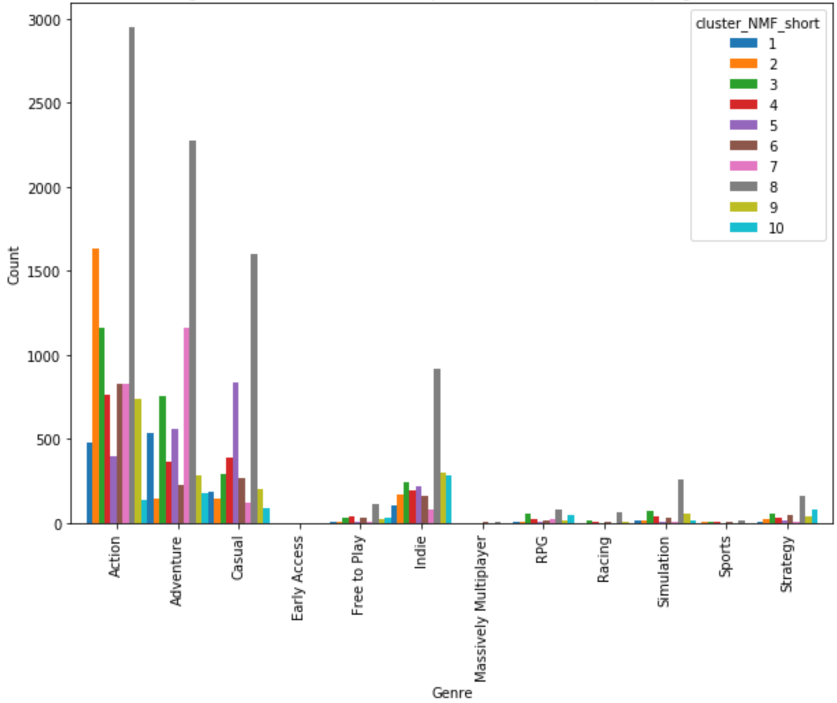
	Topic definition	Top-10 words for topic
1	Mysterious games with escape mechanics	find, way, explor, mysteri, escap, find way, help, home, secret, go
2	Fantasy games requiring exploring and saving the world	world, explor, open, open world, set, save, magic, fantasi, war, save world
3	Puzzle games	puzzl, puzzl game, solv, solv puzzl, challeng, level, relax, platform, jigsaw, use
4	Adventure games	adventur, adventur game, stori, explor, mysteri, action, action adventur, rpg, person, click
5	Turn-based strategic games	base, turn, turn base, strategi, battl, strategi game, tactic, rpg, combat, war

6	Gameplay related descriptions	play, game play, friend, role, play game, role play, mode, vr, onlin, fun
7	Game format	player, singl, game player, singl player, person, multiplay, first, mode, first person, vr
8	Action games	action, enemi, fight, level, platform, use, fast, battl, shooter, challeng
9	Recently released games	new, time, take, make, one, experi, life, stori, get, vr
10	Survival tower defense games	surviv, build, explor, citi, tower, resourc, horror, defens, manag, craft

With this table, we answer to one of the research sub-questions and show that we cannot reject H5. There are meaningful and clearly separable topics, extracted from the corpus of video games descriptions.

Figure 10 shows the distribution of NMF topics in short descriptions per genre. Since it is one of the research sub-questions, it makes sense to look at the difference. Firstly, for the Action genre, there are many descriptions with topic 2 (which describes world exploration and saving). There are relatively many games in the Adventure genre with topics 1 and 7 in their short descriptions. These topics are about mystical escape games and the game format, respectively. Finally, Strategy and RPG genres have frequently mentioned topic 10, describing survival and tower defense games, which align entirely with these genres.

Figure 10. Bar charts of NMF topics for short descriptions per genre



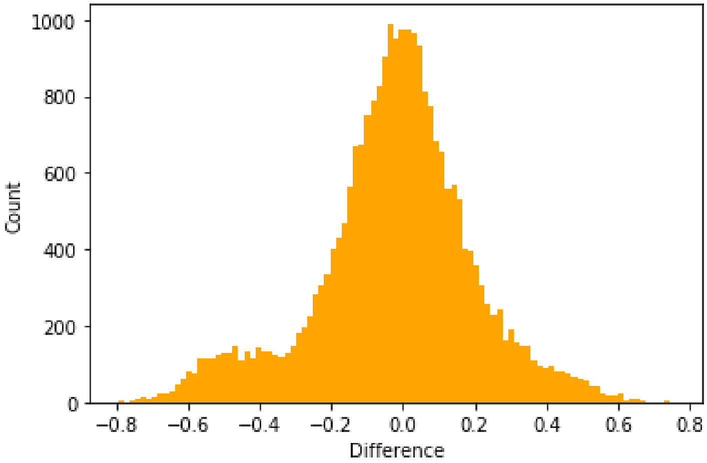
Emotional polarity and subjectivity

The following feature engineering approach was using the *TextBlob* package to extract the polarity and subjectivity of the video game descriptions. Polarity was set as a dummy variable: if the polarity score was more or equal to 0, the description was considered positive or 1. Otherwise, the description tone was deemed to be negative or 0. As a result, there are 62% and 71%

descriptions with positive polarity for short and full type, respectively. Another fact is that for 67% of observations, the polarity of short and full descriptions is the same. 21% of observations have negative short descriptions and positive full descriptions. The remaining 12% of the sample have positive short descriptions and negative full descriptions. This way, the general emotional tone is mainly the same for both descriptions of the same video game.

Subjectivity is a continuous variable, showing how emotional (in terms of language, this is expressed with modifiers and their intensity) the description is. The average subjectivity is 0.49 and 0.51 for short and full descriptions, respectively. The difference in subjectivity between the two types of descriptions is shown in Figure 11 and looks almost perfectly normally distributed. This fact means that most observations have the same subjectivity, in the same way as it was with the polarity. A small peak in the left part of the histogram means that there are more observations for which the subjectivity of full description is higher than the subjectivity of short description.

Figure 11. Histogram of the subjectivity difference between short and full descriptions



This fact together with the trends of emotional shares for short and full description support the H3 of this master thesis. There is no difference in text characteristics between short and full descriptions. Continuing with subjectivity, we conducted the statistical t-test to measure the significance of difference in subjectivity between games in category 0-20k owner (majority category) and all other games. The test has shown that the difference is highly significant (with *p-value = 0)*, meaning, that H5 of this research also cannot be rejected.

Now, when the whole dataset with all generated features is described carefully, this thesis moves to the model's results based on this data.

4.2 Model results

Steam presents many games from small independent game developing studios. These developers do not have a budget for marketing companies, and it is hard to attract gamers since there are thousands of such projects. As a result, this research uses with highly imbalanced data. 75% of the dataset are games with less than 20,000 owners. Of the sample, only 13.7% and 6.6% belong to the category with 20-50k and 50-100k owners, respectively. The most popular projects share (more than 500,000 owners) equals 0.35%. There is no sense in predicting each category separately since there is too significant class imbalance.

To solve this issue, we merged owners categories. The first option was to keep three classes: 0-20k, 20-50k and more than 50k owners, allowing better differentiation of the video games. The second option considered only two classes: 0-20k and more than 20k owners. This variant has the following logic: since the lowest category is the major one, that is already a success for the developer to overcome the threshold of 20 thousand owners. Also, this is important to understand which factors influence this success.

The imbalance is still strong, even after classes merge. This way, the majority category of 0-20k owners was undersampled. 6,500 and 7,000 were randomly sampled from the majority category for three and two classes cases, respectively. After it, categorical variables (genre and review tone) were transformed into dummy variables.

The Model results sub-section presents Random Forest and Support Vector Machine for three classes and then the same models for two classes. After it, we compare the results and interpret the best model. All the models were trained on 70% of the dataset and tested on the remaining 30%.

4.2.1 Models for three classes

The main challenge in building a machine learning model is the hyperparameters tuning. This way, before moving to the model interpretation, we describe how the best model was found. First of all, the base RF model was constructed with default parameters for the RF model in *scikit-learn* library for Python. This model gives the accuracy of 61% and this the benchmark result.

Random forest allows making the hyperparameters tuning easier and faster compared to other machine learning methods. Since each tree is trained on the bootstrap sample of the initial dataset, the out-of-bag (OOB) error can be calculated. This quality measure checks for good the model performs on the observations that were not included into bootstrap samples. 240 models with different hyperparameters sets were trained, and the best one was chosen based on the lowest OOB error. The full list of checked hyperparameters can be found in Appendix 2.

As a result, the final RF model for three classes has the following hyperparameters: 500 trees in the forest, minimal sample to continue split equal to 6, at least 2 observations should be in the final leaves, 10 features considered during the split and maximal depth of the tree is 50. The accuracy of the model is 62.74%.

The accuracy of 62.74% is not bad for the classification task with three classes, where the chance to choose the correct class randomly equals 33%. However, accuracy is not the only metric to check the quality of the model. Table 2 is the confusion matrix for this model and Table 3 shows precision, recall and F1-score for each class, giving a more detailed overview.

Table 2. Confusion matrix for RF for three classes

Class	Predicted 0-20k	Predicted 20-50k	Predicted 50k+
Actual 0-20k	1720	123	104
Actual 20-50k	506	244	271
Actual 50k+	233	199	455

Table 3. Quality metrics for RF for three classes

Class	Precision	Recall	F1-score
0-20k	0.70	0.88	0.78
20-50k	0.43	0.24	0.31
50k+	0.55	0.51	0.53

The model predicts the games from 0-20k category with the highest metrics. It works worse but still adequately with 50k+ class but struggles with the medium class of 20-50k owners. This means that either the second category is not different enough or the features used by this research cannot explain this difference.

Regarding the feature importance, the most valuable features of this model are time from the game's release, reviews tone "Very Positive", price, length of the full description, sentiment and subjectivity of both full and short descriptions, length of the short description and the fear emotion in full description.

The next model built for three classes is the **Support Vector Machine** with one-vs-one scheme. This model requires normalized variables since it works with distances, and measures influence the results. The variables were transformed with max-min normalization: the minimum and maximum values equal 0 and 1, respectively, and all other values lie between these limits.

As it was mentioned in the Method section, the SVM has several kernels to work with. This research uses RBF and polynomial kernels for three classes case since the linear kernel is not suitable for multi-class tasks. Also, the SVM algorithm is computationally more intensive than RF, and the best model search can be highly time-consuming. To solve this issue, we used grid search cross-validation. The parameters for SVM grid search cross-validation are presented in Appendix 3.

The grid search cross-validation used three folds and 280 models to find the best combination of hyperparameters. This combination is RBF kernel, *C* equal to 1000 and *gamma* equal to 0.01. The accuracy of this model equals 59.16% that gives, compared to the base SVM model (RBF kernel, *C* equal to 1 and *gamma* equal to 0.21), 1.71% accuracy improvement. However, compared to the results of the RF model, such an accuracy score looks disappointing. Moreover, Tables 4 and 5 present the confusion matrix and the remaining quality metrics, showing that SVM model performs worse than RF model for all three classes.

Table 4. Confusion matrix for SVM for three classes

	Predicted 0-20k	**Predicted 20-50k**	**Predicted 50k+**
Actual 0-20k	1594	217	139
Actual 20-50k	422	338	264
Actual 50k+	225	307	349

Table 5. Quality metrics for SVM for three classes

Class	**Precision**	**Recall**	**F1-score**
0-20k	0.71	0.82	0.76
20-50k	0.39	0.33	0.36
50k+	0.46	0.40	0.43

Another disadvantage of the SVM algorithm with non-linear kernels (compared to the linear one) is the lack of interpretability. Using the RBF kernel, the model transforms variables, making the calculation of feature importance impossible. This way, the conclusion for three classes is that Random Forest is more accurate and provides the possibility of meaningful interpretation.

4.2.2 Models for two classes

Moving to the two classes tasks, this research starts with the **Random Forest** approach. The same out-of-bag error approach was used to find the best model. This is the model with the following hyperparameters: 700 trees in the forest, minimal sample to continue split equal to 6, at least 5 observations should be in the final leaves, 20 features considered during the split and maximal depth of the tree was not limited.

The accuracy score of this final model is 75.76% that is a good result for the two categories classification task. The confusion matrix for this model is presented in Table 6. Also, precision, recall and F1-score equal 0.76, 0.73 and 0.74, respectively. This result shows that the prediction of the possibility to get more than 20,000 owners of the game is more accurate than the prediction of belonging to a specific category.

Table 6. Confusion matrix for RF for two classes

	Predicted 0-20k	**Predicted 20k+**
Actual 0-20k	1636	458
Actual 20k+	513	1398

The feature importance for the final RF model is almost the same as for the RF model for three classes. This can be explained with the idea that these variables are enough to understand how the gamer's perception works and which factors are the catchiest. The most important features are the following: time from the game's release, reviews tone "Very Positive", price, length of the full description, sentiment and subjectivity of both full and short descriptions, length of the short description and the sad emotion in full description.

The last model of this research is **Support Vector Machine** for two classes. Since there are only two categories now, it is possible to try the linear kernel. We used the same (see Appendix 3) grid search cross-validation, but a linear kernel option was added. As a result of the three-fold cross-validation of 82 models, the following hyperparameters combination is optimal: RBF kernel, *C* equal to 1000 and *gamma* equal to 0.018. This result means that even when the linear kernel is available, the RBF kernel performs better on the data of this research.

The accuracy of the most optimal SVM model equals 74.05%, which improves the base model's accuracy by 1.02% but is still worse than the accuracy of the RF model. Table 7 presents the confusion matrix for this model.

Table 7. Confusion matrix for SVM for two classes

	Predicted 0-20k	**Predicted 20k+**
Actual 0-20k	1521	579
Actual 20k+	460	1445

The precision, recall and F1-score metrics are equal to 0.71, 0.76 and 0.74, respectively. The recall of the SVM model is higher than the same metric for RF, meaning that SVM catches a bigger share of actual 20k+ owners games. However, the difference is slight, and there is no sense to choose the model with lower accuracy in this case. Moreover, SVM requires more time and does not allow the interpretation of feature importance. This way, Random Forest is the best model in this research.

4.2.3 Final model interpretation

The final model is Random Forest for two classes with the accuracy score of 75.76%. Feature importance allows understanding what factors influence the game's commercial success: whether it will be good enough to achieve more than 20,000 downloads. The first variable is the video game's "age", which is logical. The older the project is, the greater is the number of people who can know about it. The second factor is the reviews' tone or, being more precise, the "Very Positive" reviews tone. This fact relates to the idea that potential buyers look at the experience of others and consider it during the decision-making process. Next, the price is also important for gamers, and this goes in line with the demand theory: the quantity (number of downloads or owners of the game) is correlated with the product's price.

The following factors relate to the descriptions' characteristics. The length of the full and short description influences the probability of buying the video game. Possibly, too small descriptions do not give enough information about the project, and too extensive ones look overwhelming, complicated and tedious. The next block of features relates to the sentiment and subjectivity of both short and full descriptions. As presented in the Data description sub-section, the higher the subjectivity of the descriptions is, the more successful the project is. The model caught this pattern and also another one related to emotional polarity. The final variable from ten the most important is the share of sad emotion in full descriptions. This emotion can appear in the descriptions of the video games with the dramatic storyline or since it also refers to such words as "have nerves of steel" in horror games, which are pretty popular. Finally, since the final model is built and interpreted, this research paper moves to the General discussion section.

5. General discussion

This section introduces the discussion of this master thesis, considering theoretical and methodological base, data collected, and models built. There are answers to the research question and sub-questions, formulated in the Introduction section, and managerial and academic implications, based on the Results section.

5.1 Answer to research question(s)

Before answering the main research question, we found the answers for five research sub-questions.

The first sub-question: *are there any differences in descriptions characteristics (such as keywords/sentiment/topics) between genres?* The research has shown that there are actually some patterns. Regarding the topics and keywords in short descriptions, world exploration and saving are often mentioned for the Action genre, mysterious and related to the game format keywords are used for the Adventure genre, and survival tower defense topic appears in the descriptions of Strategy and RPG genres. Emotional tone also varies across genres. The anger appears more often in the descriptions of Massively Multiplayer, Action and Adventure games. The developer of Racing and Strategy projects describe them using fear emotion (which is strongly related to the risk and speed). Finally, the surprise emotion presents mainly in Adventure, Sports and Action genres descriptions.

The second sub-question: *how can we measure the atmosphere of video games descriptions, and what insights on video game's commercial success can be found using this measurement?* The atmosphere can be measured with the emotional tone of the words used in descriptions. In this research, five emotions were used: anger, fear, surprise, happiness and sadness. The Random Forest model with the highest accuracy score had the following emotions in the list of the most important features: sadness, fear and surprise (all of them in full descriptions). This way, the answer to this research sub-question is: yes, it is possible to measure the description's atmosphere using emotional tone, and it influences the commercial success of the video game.

The third sub-question: *do the text characteristics of short and full descriptions (emotions shares, subjectivity, and emotional polarity) match or not?* These characteristics are emotions, subjectivity and polarity. As shown in the Data description sub-section, for most observations, the characteristics of full and short descriptions match. This way, if the short description is highly subjective, the long one will also be very emotional. However, there is an interesting point in the difference distribution (Figure 11) that full descriptions are sometimes more subjective (due to the small peak in the left side of the distribution).

The fourth sub-question: *is there any difference in emotional polarity and subjectivity between different groups of games? Can this difference be used to answer the main research question of this master thesis?* The H4 states that these factors influence the commercial success of video games. Due to conclusions from sub-sections Data description and Model results, this hypothesis cannot be rejected. Both subjectivity and emotional polarity influence the number of owners of the game. Moreover, the more subjective and emotional the description is, the higher the probability of the commercial success of the video game.

The fifth sub-question: *can we extract the meaningful topics from the corpus of video games descriptions? Are these topics different for short and full descriptions, and can they help to explain video games sales?* This research has shown that the topics can be extracted with the help of either LDA or NMF algorithm. The topics found with NMF in short descriptions of video games are meaningful and interpretable; however, they did not appear at the top of the most important feature in Random Forest. Also, the topics for full descriptions were messy and did not fully match the topics from the short ones.

Finally, the answer to the main **research question** comes. The results of this master thesis show that the following aspects of the video game's description give insides into its commercial success: length, emotional tone, emotional polarity and subjectivity. The more emotional and subjective the description is, the more it attracts potential buyers and the higher the video game sales. These results can be used by independent developers and game development studios' managers, which is described in more detail in the following sub-section.

5.2 Managerial implications

The results of this research have several managerial implications. First of all, increasing the level of the subjectivity of the descriptions. They should be still informative but adding expressive and catchy words will attract gamers.

The second piece of advice is to pay more attention to the length of the short and full descriptions. Both should not be too long, even the full one, because potential buyers can be bored. However, they should stay informative.

The last suggestion is to use more words related to fear, surprise, and sadness in the descriptions. The model results show that these three emotions influence the commercial success of the video game. These emotions are strong and can catch the attention even if the person is just quickly looking through the page.

5.3 Academic implications

The influence of descriptions and their features on sales is hardly researched. Much more attention authors pay to the reviews, the reaction of customers regarding their experience. However, the triggers in the text description, triggers that allow attracting potential buyers and force them to make a decision, are crucial. This master thesis can be helpful for those researchers who want to explore this field.

On the other side, this research can be useful for those interested in the video games industry and how it works. This thesis provides general industry information and describes patterns on Steam, the giant video games online distribution platform. This overview can be improved and extended to the full, up-to-date video games industry analysis.

5.4 Limitations and further research

Even since this research has shown promising results, there is some space for improvement. First of all, the actual number of downloads per game could make the model even more accurate and give more possibilities for interpretation. However, this is the information that Steam does not disclose. This way, the second point relates to other features available only from the platform itself. These features could be, for instance, the developer's rating or marketing-related information.

The last possible improvement lies in the technical field. More powerful computer could be used to scrape even more observations from the platform, and these data could be used to improve the model results.

6. References

1. Alm, C. O., Roth, D., & Sproat, R. (2005, October). Emotions from text: machine learning for text-based emotion prediction. In Proceedings of human language technology conference and conference on empirical methods in natural language processing(pp. 579-586).
2. Bankhurst A. (2021). It Takes Two Selling 1 Million Copies Shows That Players Want Co-op Only Games, Josef Fares Says. Retrieved September 17, 2021, from https://www.ign.com/articles/it-takes-two-selling-1-million-copies-shows-that-players-want-co-op-only-games-josef-fares-says
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
4. Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. JMIR Public Health and Surveillance, 6(4), e21978.
5. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).
6. Breiman, L. (1999). Random forests. UC Berkeley TR567.
7. Bruwer, J., Saliba, A., & Miller, B. (2011). Consumer behaviour and sensory preference differences: implications for wine product marketing. Journal of Consumer Marketing.
8. Chalk A. (2019). Players protest Epic's Metro Exodus exclusive by review-bombing the series on Steam. PC Gamer. Retrieved September 29, 2021, from https://www.pcgamer.com/metro-review-bomb-steam/
9. Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems, 163, 1-13.
10. Clarke, R. I., Lee, J. H., & Clark, N. (2017). Why video game genres fail: A classificatory analysis. Games and Culture, 12(5), 445-465.
11. Clement J. (2020). Increase in sales in the video game industry during the coronavirus (COVID-19) pandemic worldwide as of March 2020, by type. Retrieved September 13, 2021, from https://www.statista.com/statistics/1109979/video-game-console-sales-covid/
12. Clement J. (2020). Number of active video gamers worldwide from 2015 to 2023. Retrieved September 13, 2021, from https://www.statista.com/statistics/748044/number-video-gamers-world/

13. Deardorff N. (2015). An Argument That Video Games Are, Indeed, High Art. Forbes. Retrieved September 29, 2021, from https://www.forbes.com/sites/berlinschoolofcreativeleadership/2015/10/13/an-argument-that-video-games-are-indeed-high-art/?sh=1ecd70a97b3c

14. Game Developers Conference. (2020). 2020 State of the Game Industry Report. Retrieved September 16, 2021, from https://reg.gdconf.com/gdc-state-of-game-industry-2020

15. Gómez-Maureira, M. A., & Kniestedt, I. (2019). Exploring video games that invoke curiosity. Entertainment Computing, 32, 100320.

16. Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent Dirichlet allocation. advances in neural information processing systems, 23, 856-864.

17. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

18. Loria S. TextBlob. (2013). GitHub repository, https://github.com/sloria/textblob

19. Luc, M. H., Tsang, S. W., Thrul, J., Kennedy, R. D., & Moran, M. B. (2020). Content analysis of online product descriptions from cannabis retailers in six US states. International Journal of Drug Policy, 75, 102593.

20. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113.

21. Minotti M. (2019). Detroit: Become Human comes to PC via Epic Games Store on December 12. VentureBeat. Retrieved September 29, 2021, from https://venturebeat.com/2019/11/19/detroit-become-human-comes-to-pc-via-epic-games-store-on-december-12/

22. Porter, M. F. (1980). An algorithm for suffix stripping. Program.

23. Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143-157.

24. Python Package Index - PyPI. (n.d.). Python Software Foundation. Retrieved from https://pypi.org/

25. Richardson, L. (2007). Beautiful soup documentation. April.

26. Richter F. (2020). Gaming: The Most Lucrative Entertainment Industry By Far. Retrieved September 16, 2021, from https://www.statista.com/chart/22392/global-revenue-of-selected-entertainment-industry-sectors/

27. Sacranie, J. (2010). Consumer Perceptions & Video Game Sales: A Meeting of the Minds.

28. Selenium documentation. (2011). Selenium. Retrieved September 20, 2021, from https://www.selenium.dev/selenium/docs/api/py/api.html

29. Stoll J. (2021). Netflix's net income from 2000 to 2020. Retrieved September 13, 2021, from https://www.statista.com/statistics/272561/netflix-net-income/

30. TextBlob: Simplified Text Processing. (2013). TextBlob. Retrieved October 18, 2021, from https://textblob.readthedocs.io/en/dev/index.html

31. Wang, W. M., Li, Z., Tian, Z. G., Wang, J. W., & Cheng, M. N. (2018). Extracting and summarizing affective features and responses from online product descriptions and reviews: A Kansei text mining approach. Engineering Applications of Artificial Intelligence, 73, 149-162.

32. Web Scraper Documentation. (2019). Web Scraper. Retrieved September 20, 2021, from https://webscraper.io/documentation

33. Witkowski, W. (2020). Videogames are a bigger industry than movies and North American sports combined, thanks to the pandemic. MarketWatch. MarketWatch, December, 22.

7. Appendixes

7.1 Appendix 1. Steam reviews tone definition

The reviews tone in Steam is defined by the percentage of positive reviews and the number of reviews in general. The table below shows the percentages for each review tone category.

Review tone	**Percentage of positive reviews**
Overhwelmingly Positive	95-99%
Very Positive	94-80%
Positive	80-99% (and few review)
Mostly Positive	70-79%
Mixed	40-69%
Mostly Negative	20-39%
Negative	0-39% (and few review)
Very Negative	0-19%
Overwhelmingly Negative	0-19% (and many reviews)

7.2 Appendix 2. Hyperparameters for Random Forest

Hyperparameters

Hyperparameter	**Possible values**
n_estimators	500, 700
max_features	2, 5, auto, log2, 10, 20, 30, 40, 50, 55
max_depth	30, 50, None
min_samples_split	2, 6
min_samples_leaf	2, 5

Here and further for Random Forest models hyperparameters mean the following: *n_estimators* is the number of trees in the forest, *max_features* is the number of features considered during the node split, *max_depth* is the maximal depth of the tree in the forest, *min_samples_split* is the minimum number of observations in the node to continue split, *min_samples_leaf* is the minimum allowed number of observations in the final leaves.

7.3 Appendix 3. Grid search cross-validation parameters for Support Vector Machine

Grid search cross-validation parameters

Hyperparameter	**Possible values**
C	0.1, 1, 10, 100, 1000
gamma	1, 0.1, 0.01, 0.001, 0.0001, auto, scale
kernel	rbf, poly
degree	2, 3, 4, 5

The hyperparameters of Support Vector Machine mean the following: *C* is the regularization parameter, showing how wide the margin will be, *gamma* shows how sensitive the model is to a single training observation (*auto* means that $gamma = \frac{1}{number\ of\ features}$, *scaled* relates to $gamma = \frac{1}{number\ of\ features * Var(X)}$), *kernel* shows which function is used by SVM and *degree* relates to polynomial kernel and its degree.