"The Impact of Twitter Sentiment on Ryanair's Business Performance"

Ioanna Nasiara

SID: 203862331606

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in E-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

INTERNATIONAL HELLENIC UNIVERSITY

"The Impact of Twitter Sentiment on Ryanair's Business Performance"

Ioanna Nasiara

SID: 203862331606

Supervisor:	Assist. Professor Christos Tjortjis
Supervising Committee Members:	Dr. C. Berberidis
	Assoc. Prof. Papadopoulos

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in E-Business and Digital Marketing

DECEMBER 2019

THESSALONIKI – GREECE

Acknowledgements

I would like to thank my supervisor, Prof. Christos Tjortjis for his excellent mentoring and support throughout this project. I am grateful to my family, who have always supported and believed in me, as well as to Michalis, for his precious help and guidance in this project.

Abstract

Social media platforms have gained extreme popularity and strength, due to their ubiquity and the need of people for social networking and content sharing. However, social media are also used by consumers, who are willing to share online opinions and experiences from products and services. That has posed to companies an imperative need to exploit these rich online sources, in order to gain deeper understanding of their target audiences and optimise their business performance. The focus of this study is the examination of the impact that Twitter has on the performance of an airline company, like Ryanair, that attracts a lot of attention online. Tweets relevant to Ryanair have been collected and analysed, with the purpose of extracting their sentiment. The sentiment analysis was carried through with the help of Valence Aware Dictionary and Sentiment Reasoner (VADER) lexicon, as well as a machine learning classification model, based on Random Forest classifier. Sentiment analysis was followed by a statistical analysis, which included a Simple Linear Regression Analysis, a Granger Causality Analysis, as well as an Impulse Responses Analysis. Our findings from the Regression Analysis suggest that there is a statistically significant relationship between the Twitter variables and future passengers' growth rate, while Granger Causality analysis indicates a unidirectional Granger causality from Twitter. Impulse Response analysis shows that the growth rate of passengers reacts significantly to shocks in Twitter.

Keywords: Social Media, Twitter, Sentiment Analysis, Linear Regression, Vector Autoregression

Contents

List of Figures

List of Tables

1. Introduction

Social Media have become an inextricable part of most people's everyday life, around the globe. The rapid advance of technology and the advent of social networking platforms have led to the production of vast amounts of data and content, created by the users (User Generated Content, UGC) (Thelwall et al., 2011). Twitter is one of the most popular media, where users are able to share their thoughts and opinions on different topics, express themselves politically and even share their experiences from using products or services (Agarwal et al., 2011). According to Ye et al. and Schivinski et al., consumers are increasingly turning to the web for further information before making a purchase, which is an indicator of online user generated reviews' impact on consumers buying behaviour. Thus, a brand's reputation, consumers' brand perception and their decision-making process are directly affected by what people say online (Ye et al., 2011).

Research has shown that online word of mouth plays a vital role in the shaping of consumers' buying behaviours (Ye et al., 2011), which poses an imperative need for brands to monitor their customers' perceptions towards their products and services. Social listening refers to the monitoring of a brand's social media channels, in order to keep track of conversations, regarding specific topics and keywords. The process of monitoring aims to return valuable and actionable insights to the company, that will support a well-informed long-term business strategy. Moreover, Social Listening is a technique that allows companies to capture specific needs and wants of customers, discover their concerns about specific products and even redesign their marketing strategy, in case that it does not resonate with the company's audience. As an example, in 2013, the well-known travel agency Expedia launched a television advertisement encouraging Canadians to book their next travel, in order to escape winter. The advertisement featured an apparently irritating violin sound, which took people to Twitter to complain and express their completely negative sentiment towards the ad. By listening to and monitoring people's feelings online, Expedia achieved a U-turn reaction by replacing the annoying violin sound and successfully eliminated the negative conversation, by analysing large-scale electronic social data (Hudson & Hudson, 2017) (Brandwatch, 2014).

Airline companies heavily rely on customers' feedback, for the purpose of constantly improving their services and providing an excellent experience at their audience (Wan & Gao, 2015). However, in the case of airlines, traditional methods of collecting customer feedback, such as questionnaires, often fail to capture the real sentiment and are costly and extremely time consuming. Thus, the use of alternative approaches, such as sentiment analysis, seems to be crucial.

A lot of research has been done in the field of Sentiment Analysis, a data mining technique that aims to study the views, attitudes and feelings of people towards organisations, brands, individuals and

events (Liu & Zhang, 2012), by classifying a piece of text, according to the emotional tone of the author, into positive, negative or neutral. Sentiment analysis can be a challenging task, but very helpful in practice, particularly for airline companies, as people tend to increasingly share their experiences on social media, hence producing a large volume of data that could be utilised for customer feedback analysis.

The rest of the dissertation is organized as follows; In Chapter 2, we present the Theoretical Background of the study, including the Fundamental Concepts of this thesis and the Literature Review. In Chapter 3, the Purpose of the study is discussed, and the Research Questions and Research Objectives are presented. In Chapter 4, the Research Methodology framework is elaborated, while in Chapter 5, we present the Results of the study. Finally, the dissertation concludes in Chapter 6 by discussing limitations and future directions for improvements and summarizing key findings.

2. Theoretical Background

2.1 Fundamental Concepts

2.1.1 Twitter

According to Kaplan & Haenlein (2011), microblogging has already become a well-established form of communication among users, enabling them to publish or retrieve short pieces of content, mainly text, images and links. Microblogs are considered a hybrid of traditional blogs and social networking platforms, with Twitter being the most popular medium with around 330 million monthly active users as of the first quarter of 2019 (Statista, 2019). Twitter was founded in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams and was launched later this year, in July 2006. Twitter updates, named Tweets, were initially limited to 140 characters, however, as of November 2017, the company has decided to increase the character limit up to 280 characters (CNN, 2017). This message size enables users to publish timely the content they want to express in the form of short statuses, which allows users to publish several tweets during a day.

Tweets are publicly accessible, thus available to everyone to read and comment, unlike Facebook, which follows a stricter policy in its users' data accessibility. Another important function of Twitter is the liberty of its users to follow whoever they wish, without seeking any permission. This function is a kind of subscription to an account, with the purpose of receiving updates in tweets in the Home or Feed, where all tweets appear in chronological order. Twitter is famous for its ability to spread news and information around the world in seconds (Kaplan & Haenlein, 2011), an ability tremendously owed

to the Retweet function, which allows users to reshare or forward tweets into a larger audience. Moreover, Twitter users are able to highlight a keyword in their tweet using the hashtag symbol '#', in order to classify it into a broader topic or theme. Last but not least, users can post tweets intended for specific accounts, with the '@' symbol.

Twitter is used every day by millions of people around the world, who are encouraged to actively participate and generate content. They may discuss politics, business or current events. Moreover, social media users tend to share online their personal experiences from using a product or service, either positive or negative. It has been shown that product reviews generated by users can have a major impact on consumers' buying decisions, particularly when they include a negative comment about a product or a service. The survey of (Jansen et al., 2009) suggests that microblogging services, like Twitter, can extensively influence electronic word-of-mouth (e-WOM). The term Word-of-Mouth refers to the transmission of information among people about their personal experiences from a product or service, which acts as a strong determinant of consumers' buying decisions. That results from the fact that e-WOM can reach a significant number of users within a very short amount of time, it is easily accessible, and it is a reliable source of information, as it is generated by consumers that, most of the times, share personal genuine experiences. Furthermore, Twitter's ubiquity and the constant connectivity of humans has increased dramatically the quantity of tweets that refer to brands, a lot of times emotionally charged by the author. Thus, brands are directly affected, as e-WOM shapes brand perceptions and images.

Goldsmith and Horowitz (2006) researched the main reasons why consumers seek online opinions from others. Their findings showed that people's motivation in online reviews pursuit originates from their need to reduce risk related to a future purchase, to find the best price in the market, to learn about a product's characteristics and to compare products online, among others. Cheung et al. (2008) investigated the extent to which consumers who seek online opinions are ready to accept online reviews and the criteria that promote this acceptance. They found that "Information usefulness" significantly influences consumers' decisions to adopt opinions and information that is published in online reviews. In particular, "Relevance" and "Comprehensiveness" are important features of "perceived Information usefulness". On the other hand, "source credibility", "accuracy" and "timeliness" do not impact significantly the "Information usefulness". Davis and Khazanchi (2008) conducted an empirical study of online word of mouth as a predictor for e-commerce sales. They found that product sales can be predicted by the number of product views in an e-commerce website, especially when images of the product are displayed. Although product category is not considered as a predictor of sales, the combination of product category and the volume of reviews about a product has a major impact on product sales.

2.1.2 Sentiment Analysis

Sentiment Analysis or Opinion Mining refers to the sub-field of Natural Language Processing that aims to determine the sentiment that is expressed in a piece of text and classify it in one of the following categories: positive, negative or neutral. The purpose of Sentiment Analysis is to expose the author's feelings and views towards a variety of topics.

The sentiment of a piece of text can fall into more than one types of categories. Apart from the classification into positive, negative and neutral, a scaling system can be also employed, assigning a score of strongly positive, positive, neutral, negative and strongly negative. Thus, the analysis might reveal more fine-grained results, that better reflect the reality. On the other hand, another more sophisticated type of Sentiment Analysis indicates precise emotions, such as happiness, surprise and anger.

The importance of Sentiment Analysis has been in a constant rise, due to the explosion of Web 2.0, which brought a whole new era of social networks, blogs, forums and online reviews websites. Every second around the globe, vast amounts of data are produced by people that discuss online about current events, politics or brands. Extracting information from this quantity of unstructured data can bring us one step closer to understand human behaviour or even predict it. A remarkable amount of studies has shown that the analysis of social media data can provide information about consumers' future buying behaviour, prediction of a company's financial performance or enable market segmentation and the design of a marketing campaign.

Given the significant role of Sentiment Analysis, it has been observed a growing number of corporations that employ such data mining methods, in order to benefit from the information that people voluntarily and lavishly share online. Twitter is among the first platforms that data analysts examine and extract information from.

There are two different methods of Sentiment Analysis that are extensively used; the lexicon-based and the machine learning method.

2.1.2.1 Lexicon-based Approach

As the name indicates, the lexicon-based approach to sentiment analysis is performed with the use of dictionaries of opinion words. The term "opinion words" refers to sentiment-rich words or phrases that are used to express attitudes or beliefs. Such words or phrases that are included in the dictionaries or "opinion lexicons" are already assigned with a pre-calculated polarity score. According to Kolchyna et al. (2015) the basic idea behind the lexicon-based approach is first to create a list (bag-of-words) with all the words that are found in a text and need to be analysed. The next step is to check which

words of the list are also included in the lexicon and add their sentiment score to the total sentiment score of the text.

As mentioned above, the lexicon is the core element of the lexicon-based sentiment analysis, therefore the creation of the lexicon is the most important task in the process. Indeed, the quality of the lexicon plays a vital role in a fine classification of text in sentiment categories. There are two different approaches in the construction of a lexicon; the manual approach and the automated approach, according to Liu (2012). The Hand-Tagging method belongs to the manual approach and allows a researcher to build a lexicon, by manually selecting words from texts, that convey a sentiment and tag them as positive or negative one by one. However, the manual approach can be very time-consuming and laborious. Thus, researchers have turned to automated techniques or combinations of the two approaches. Automated techniques are mainly distinguished into two categories; the dictionary-based approach and the corpus-based approach. The dictionary-based method suggests the addition of words in existing smaller dictionaries, such as synonyms or antonyms. Specifically, this is an iterative process, which starts with the collection of words that are already sentiment-assigned. Next, the set of these words is enriched, when their synonyms and antonyms are added, by an algorithmic system that searches online libraries. As soon as the algorithm cannot find any more synonyms or antonyms, the process ends, and no further iteration is made. In order to eliminate errors in the final list of words, human inspection is suggested. Liu (2012) reports as a drawback of the dictionary-based approach, the weak context dependency of the seed words. That is, the sentiment orientation of these words heavily relies on the context of that they appear in a sentence. The basic concept behind corpus-based approach of sentiment analysis is to create a dictionary that concerns only one specific area of interest Liu (2012). The corpus-based approach uses an initial list of emotionally charged words and then searches for more words in large corpuses that have to do with a specific domain. Therefore, a larger list of opinion words with specific sentiment orientations emerges.

VADER Lexicon

VADER stands for Valence Aware Dictionary and Sentiment Reasoner and is an open source sentiment analysis lexicon, developed by C. J. Hutto and Eric Gilbert, in 2014 (Hutto & Gilbert, 2014). According to the authors, VADER has been constructed, having in mind specific goals that it would be able to reach. Namely, VADER achieves a very high performance, when analysing text, extracted from social media. This is considered an important advantage, as this kind of text exhibits peculiarities, as it usually consists of abbreviations, emoticons and slang language. Although VADER performs very well in social media text, it comprises a good solution in several other domains, as well. Moreover, high speed is one of the assets of VADER, as it is able to even work with streaming data. However, according to Hutto and Gilbert, speed does not sacrifice performance, a remarkable characteristic.

The construction of VADER lexicon was based in a number of existing, credible sentiment lexicons, such as the *Linguistic Inquiry Word Count (LIWC),* the *Affective Norms for English Words (ANEW)* and the *General Inquirer (GI)*. These sources have been supplemented with further lexical characteristics of language that is used in social media nowadays. Then, the authors employed the "*Wisdom of the Crowd*" approach, in which, humans examined in total, more than 9,000 lexical elements. Out of the total number of the candidate elements that were initially generated, 7,500 elements were finally kept, with human-validated valence scores, that not only denoted their *polarity*, i.e. positive or negative, but also their *sentiment strength*, assigning in that way sentiment scores that are consistent with the scale of the positivity or negativity of a lexical feature. *Figure 1* demonstrates the steps that Hutto and Gilbert followed, during the construction of VADER, along with the evaluation of the lexicon, as described below, in detail.

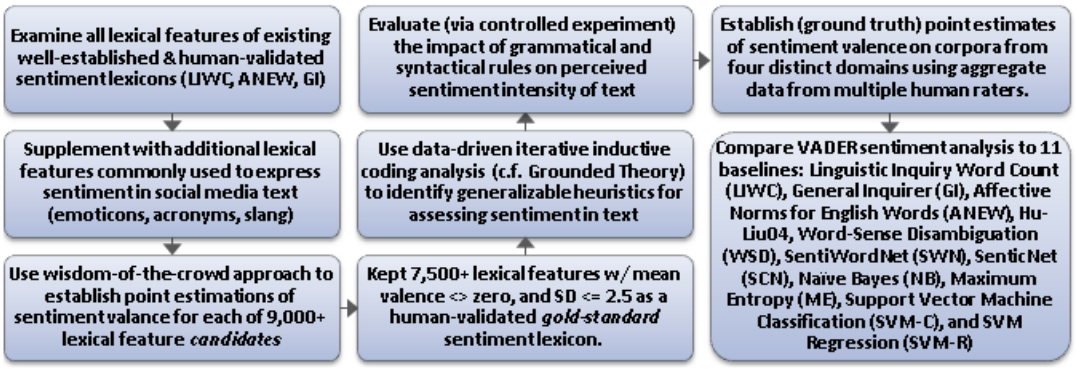


Figure 1. VADER lexicon construction and evaluation process. (Hutto & Gilbert, 2014)

In order to evaluate VADER, as a sentiment lexicon and its suitability for different domains of interest, it was tested in four domain contexts: a) in social media text, b) in movie reviews, c) in technical product reviews and d) in opinion news articles. The metrics that were used to measure the performance of the lexicon were a) the *correlation* of the sentiment valence, as calculated by VADER and the mean sentiment rating, as provided by 20 human raters, and b) the three classification metrics of *precision*, *recall* and *F1 score*. Consequently, VADER's results were contrasted against seven other important and solidified sentiment analysis lexicons' results *(Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD) using WordNet, Hu-Liu04 opinion lexicon)*. The results showed that VADER performed remarkably well in the social media domain and very adequately in the rest domains, which means that it is capable of domain generalisation. *Table 1* demonstrates the results of all sentiment lexicons and the human raters, in all four different domains.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 1. VADER classification performance contrasted with the performance of human raters and seven other sentiment lexicons, per domain. (Hutto & Gilbert, 2014)

2.1.2.2 Machine Learning Approach

The Machine Learning approach of Sentiment Analysis refers to the task of *classification* of text, into positive, negative and neutral, according to its sentiment orientation. Machine learning techniques for text classification fall into two categories; those that use supervised learning and those that use unsupervised learning. As far as supervised learning is concerned, machine learning techniques entail a training dataset and a test dataset. The training dataset is an initial set of already labelled instances that sets the rules by which an algorithm learns how to classify new unlabelled examples. Specifically, during the training process, the model is trained to classify input examples, according to the predetermined label. The training process is followed by the testing process, where the model is tested by classifying unseen data. The first step in machine learning text classification is to convert the text into a vector, i.e. a numerical representation, that the algorithm can understand. This procedure is called *Feature Extraction* or *Text Vectorisation*.

Machine Learning Algorithms

A simple but popular machine learning algorithm in text classification is Naïve Bayes. Naïve Bayes belongs to probabilistic classifiers and it computes the posterior probability of a class *c*, given a predictor (feature) *x*. Based on the Bayes Theorem, the algorithm predicts the probability that a given feature belongs to a specific label (Medhat et al. 2014). Naïve Bayes is given by the following formula:

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \quad (1)$$

Another machine learning algorithm that is widely used in text classification tasks is the Support Vector Machine. Support Vector Machines (SVM) is a linear classification method, that has been giving valuable results when it comes to textual data. The concept of SVMs is to develop a hyperplane, which separates the data points in the search space into classes. As shown in *Figure 2,* there are two different classes. Support Vector Machine algorithm determines three Hyperplanes, A, B and C. The data points that are closest to each line are called support vectors. The SVM algorithm picks as the optimal Hyperplane, the one that has the maximum distance (margin) from the support vectors. In *Figure 2*, the A Hyperplane is the one that maximizes the margin; thus, it is the optimal.

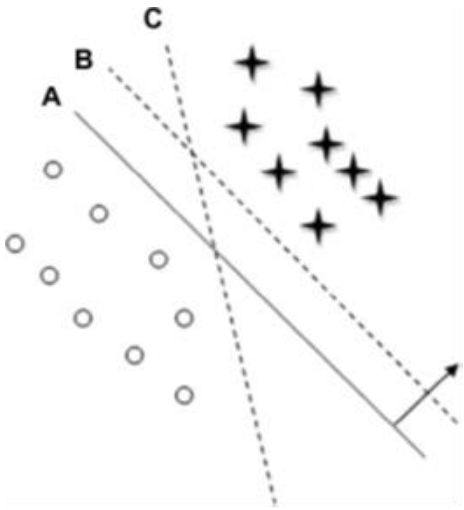


Figure 2. Support Vector Machine on a Classification Problem (Medhat et al., 2014)

Another important algorithm in classification is Random Forests. Random Forests is an ensemble method that works by building up numerous decision trees, using statistical measures, such as Information Gain and Gini Index. In order to classify a new object, each individual tree gives a classification, while the frequency of appearance of the classes is counted. The classification with the highest frequency of class appearance is selected over all the other trees in the forest. Random Forests handle overfitting well and work well in large datasets with high dimensionality.

Classification Evaluation Metrics

Confusion Matrix is of significant importance in the evaluation of the performance of an algorithm, as it allows for visualisation of the actual and the predicted classes. As shown in *Table 2,* each row of the matrix represents the predicted class and each column represents the actual class. (Tzirakis & Tjortjis, 2017)

	ACTUAL CLASS	
PREDICTED CLASS	*Positive*	*Negative*
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

Table 2. Confusion Matrix

- True Positive (TP) refers to the number of instances that were correctly classified as positive.
- False Positive (FP) refers to the number of instances that were incorrectly classified as positive.
- False Negative (FN) refers to the number of instances that were incorrectly classified as negative.
- True Negative (TN) refers to the number of instances that were correctly classified as negative.

Classifier Accuracy or *Recognition Rate* refers to the percentage of all correctly classified instances of the test set. *Accuracy* is given by the following formula:

$$Accuracy = \frac{(TP + TN)}{All\ instances} \quad (2)$$

Precision refers to the percentage of instances that were classified as positive and are actually positive. *Precision* is given by the following formula:

$$Precision = \frac{TP}{(TP + FP)} \quad (3)$$

Recall refers to the percentage of actually positive instances that were classified as positive. *Recall* is given by the following formula:

$$Recall = \frac{TP}{(TP + FN)} \quad (4)$$

F-score or *F measure* refers to the harmonic mean of Precision and Recall. *F-score* is given by the following formula:

$$F - score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (5)$$

2.2 Literature Review

Lassen et al. have thoroughly researched the usefulness of Twitter social data on future smartphone sales forecasting. Specifically, they aimed at predicting iPhone sales by implementing the Asur and Huberman methodology, which investigates whether a tweet can reflect the interest of a person in a product and their intention to buy or endorse the product. Lassen et al. consider the AIDA model

(Awareness, Interest, Desire, Action), which concerns the four stages within a sales process, from the time that a consumer becomes aware of a product to the time that they decide to act by either purchasing it or not. The authors deploy the Hierarchy of Effects, a theory based on the three stages of Cognition, Affect and Behaviour, that sheds light on the way that advertisement works and its impact on a consumer's buying decision. The authors claim that a tweet about a specific product (e.g. iPhone) reveals a level of engagement of a consumer in one of the AIDA and Hierarchy of Effects stages. In their empirical work, they develop a forecasting regression model. They use time lagged and season weighted Twitter data (number of Tweets mentioning the iPhone) and their estimated sentiment, as predictors of the future iPhone sales. Their results indicate that there is a strong correlation between iPhone sales and iPhone tweets. Hence, social media data can act as a predictor for sales and that social media interactions influence the consumers' purchasing behaviour.

Asur and Huberman (2010) investigated how social media data are strong enough to predict the results of real-life issues. In particular, the purpose of their research was to forecast box-office revenues for movies, taking into consideration the conversation around these movies on Twitter. Asur and Huberman collected 2.89 million tweets for 24 movies and predetermined a ''critical period'' of a week before and two weeks after the movies' release, in order to run their experiments. The authors measured the attention each movie received by the audience, by the rate that tweets for a specific movie were created per hour. The tweet-rate was proven to be strongly correlated with the box office profits, as shown by the linear regression model that was used. Furthermore, Asur and Huberman examined the relationship between sentiment and box office revenue, with the use of sentiment analysis of all tweets. According to their findings, sentiments improved the prediction, however they did not manage to outperform the rate of tweets per hour. It is also important to mention that the results of their research predicted the actual box office sales with higher accuracy, than the Hollywood Stock Exchange.

In their work, Fan, Che and Chen (2017) combined the Bass/Norton model with sentiment analysis, in the scope of forecasting product sales, with the use of product reviews. They collected both online reviews and historical sales data about a Hyundai automobile from 2007 to 2014. After assigning a sentiment to each piece of review data using the Naïve Bayes classification algorithm, they proceeded with the sales prediction using the Bass-emotion and Norton-emotion models, in which they integrated the sentiment index. Their experiments resulted in an accurate method for predicting product sales.

Bollen, Mao and Zheng (2011) investigated whether public mood, as captured by analysing a large scale of tweets, is related with and can predict the closing values of the Dow Jones Industrial Average (DJIA). The sentiment analysis of the tweets was conducted with the help of two different tools; the *Opinion Finder*, which distinguishes the daily public sentiment into positive or negative and the

GPOMS, that provides an in-depth view into six specific human feelings (Calm, Alert, Sure, Vital, Kind and Happy). Subsequently, the authors employed the Granger Causality econometric method, in order to discover whether the daily sentiment on Twitter can act as a predictor of the changes of the DJIA daily values. Their results indicate that public mood has an effect on DJIA prices. In particular, Bollen et al. observed that the mood dimension of "Calm" was strongly correlated with stock market fluctuations. Furthermore, in an attempt to precisely evaluate the extent to which online public sentiment affects the stock market, the authors implemented a fuzzy neural network system that forecasts DJIA prices, taking into consideration first the prices of the last three days and second the prices of the last three days-along with the mood variations. Their findings highlight the importance of the public sentiment in the stock market, as the fuzzy neural network model achieved a prediction accuracy of 87.6%.

The prediction of a company's stock price movements was the subject of Bing, Chan and Ou (2014) research, which gave some insightful results. First, Bing, Chan and Ou (2014) thoroughly investigated the structure of each piece of data and the relationship among its attributes. They formed hierarchical connections among attributes of products, creating layers of information. These layers of information about a product or brand act as supporting information, which, however, are taken into account by the algorithm, in an attempt of classifying moods about a top layer attribute. As an example, "Apple Inc." can be considered a top layer attribute, "iPhone" and "MacBook" belong to a second layer and "Battery" and "Design" to a third one. However, all three aforementioned layers of information contribute to the assignment of a sentiment to "Apple Inc.". Bing et al. propose a twofold methodology. First, each tweet is assigned with a sentiment, that ranges from negative- to positive+. Second, attributes that have been defined in the previous step, are studied in order to draw association rules, that will support the forecast of stock prices movements. Concerning the sentiment analysis, Bing et al. used the TF-IDF model, in order to create a word list and SentiWordNet lexicon, in order to assign a sentiment score in each word. Consequently, the authors investigated the possible association rules among all attributes, by applying a chi-square test. Furthermore, in order to deal with the question if specific sentiment classes are associated solely with specific attributes, they employed the adjusted residual method. Bing, Chan and Ou (2014) proceeded with the selection of the real stock prices from Yahoo! Finance for 30 companies of different industries and plotted the correlation between the stock price values and the public mood towards each company. The implementation of their proposed algorithm returned an average accuracy of 66.48% across different industries, with an optimal time lag of 3 days. The highest accuracy was this of the IT industry, which reached 76.12%.

Tabari et al. (2018) investigated the role of social media in the stock market. In particular, in their empirical analysis, the authors created a dataset that consisted of tweets, selected in a three-month

period. The tweets were selected according to their relevance with stock market, that is, only if they contained at least one stock symbol. The dataset was then submitted to Amazon Mechanical Turk in order to be labelled. Tabari et al. used the labelled dataset, in order to create a classification model using Support Vector Machine, with an accuracy of 79.9%. Subsequently, the authors implemented a Granger Causality analysis, which showed that there is a statically significant causality between stock returns and Twitter sentiment. In particular, they employed two models; the first one researched whether stock returns cause Twitter sentiment scores and the second one, if Twitter sentiment scores cause stock returns. According to their findings, specifically for Apple, the Granger Causality analysis showed that Twitter sentiment affects the stock returns, with a lag of two days. The same result was observed in the case of Facebook, as well.

3. Research Problem

This chapter aims to present the purpose of this dissertation, as well as to define the research problem and the research questions that assist the investigation of the research area.

3.1 Research Questions and Purpose

The purpose of this dissertation is to examine the importance of social media in the growth of a low-cost airline company and the extent to which opinion sharing in social media can affect the future sales of the organisation. This examination will be carried through by assigning a sentiment class to Twitter data, using sentiment analysis. Subsequently, the relationship between metrics that represent the Twitter data and the growth rate of passengers will be investigated, using regression analysis. In addition, we will attempt to clarify whether it is Twitter that causes changes in the passengers' growth rate or the other way around, using VAR methodology. Last, we will explore how the growth rate of passengers is affected by exogenous shocks, with the help of VAR Impulse Responses.

This dissertation is comprised of two main research questions:

1. *Can the sentiment of Twitter messages (tweets) have an impact on Ryanair's monthly traffic numbers?*
2. *Can tweets be used as predictor variables?*

In light of these questions, the research is governed by the following objectives, that constitute the research skeleton, in order to address these questions:

1. Collect data about Ryanair's monthly traffic information (number of passengers), for the period July 2017 till September 2019.
2. Collect Twitter messages (tweets) that concern Ryanair as an airline company and express an opinion about the company.
3. Preprocess the tweets, in order to bring them in a structured format, that can support the further analysis.
4. Classify each tweet, according to the sentiment that it expresses, into three sentiment categories: positive, negative and neutral.
5. Identify and assess the features that will be used as predictors.
6. Perform regression analysis, in order to examine the impact that Twitter has on the growth rate of Ryanair's traffic, while taking into account the sentiment that is expressed in Twitter about the company.

4. Methodology

This chapter details the process that we followed, in order to investigate and answer the research questions. First, we present the way that data from two online sources was collected. Second, we describe the two different methodologies that were followed, in order to perform sentiment analysis. Third, we introduce the methodology of the statistical analysis that we conducted, in order to explain the relationship between the dependent and the independent variables.

4.1 Data Collection

For the purposes of the empirical analysis, we used data over the period 07/2017 – 09/2019, for in total 27 months. The data were acquired from two online sources. First, we collected tweets, from Twitter, using Brandswatch API, which allowed us to gain access to tweets, without having to deal directly with Twitter API. The tweets that were collected were all in English language and all of them were targeted to Ryanair, thus included the mention *"@Ryanair"*. This way, we ensured that all noisy tweets, i.e. tweets irrelevant to Ryanair would be excluded from our final dataset. The final dataset consists of 382.425 instances and was saved in our computer as a csv file. Second, we collected traffic data for Ryanair, which is publicly available on Ryanair's official website. Traffic data consists of the monthly number of passengers, in millions, that flew with Ryanair, during the period of examination, i.e. July 2017 until September 2019.

4.2 Sentiment Analysis

4.2.1 Data Preprocessing

Data cleansing is considered an important step in natural language processing tasks, as sentiment analysis, especially when it concerns Twitter data. That is, tweets, due to their short size, usually consist of a lot of abbreviations, emojis and slang. Moreover, sentiment analysis tools perform more accurately when they are fed with words and sentences that do enrich the sentiment score with a sentiment value. The data cleansing process that we followed is briefly described in *Figure 3*.

1. We filter out content such as URLs, Hashtags, Mentions and Reserved Words (e.g. RT and FAV), using a Python preprocessing library for Twitter data, named "*Tweet Preprocessor*"[1].
2. All punctuation marks and digits are also removed from the tweets, however we keep the exclamation mark (*!*) and the question mark (*?*), as we have found that they are not negligible in the sentiment assignment process.
3. Contractions, such as "*haven't*" and "*won't*" are converted to their original forms; "*have not*" and "*will not*". Noise, such as whitespaces, new lines and tabs are also removed from the text.
4. Stopwords are eliminated, using NLTK stopwords list[2]. Stopwords are the most common words in a language, that are met in every sentence, however they do not carry any sentiment. Such words are "*the*", "*am*", "*and*".
5. Tweets are further preprocessed by tokenizing them, viz spliting them into words and symbols that are called tokens, using *NLTK Tweet Tokenizer*[3].
6. The last step of the preprocessing includes the Lemmatization of the tweets, using *NLTK Lemmatizer*.

[1] https://pypi.org/project/tweet-preprocessor/

[2] https://www.nltk.org/book/ch02.html

[3] https://www.nltk.org/api/nltk.tokenize.html

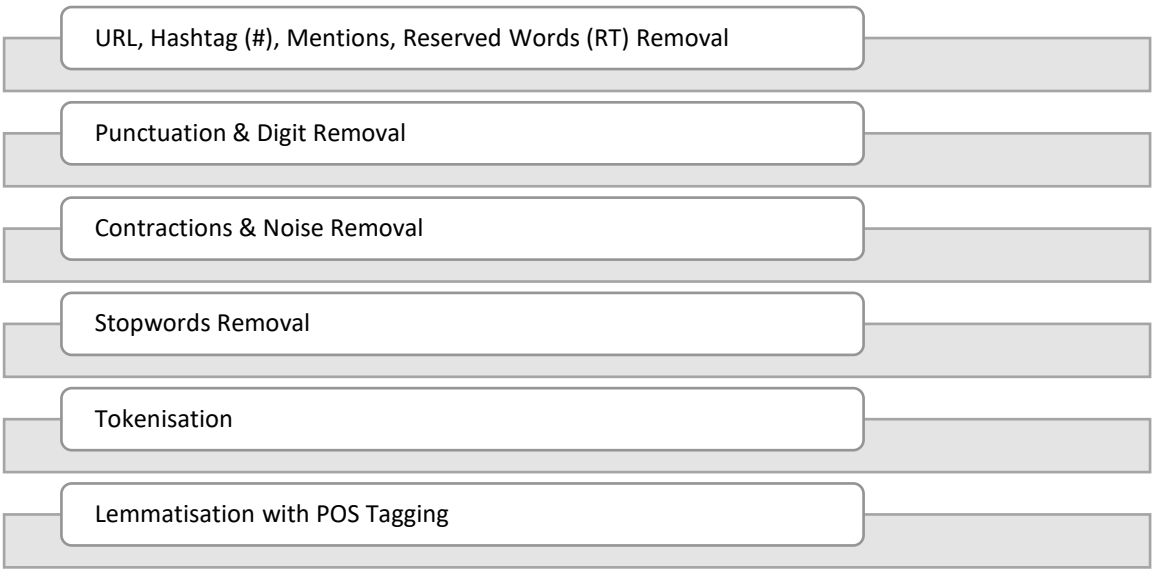


Figure 3. Data Preprocessing Process

4.2.2 Lexicon-based Sentiment Analysis

First, we decided to perform a lexicon-based sentiment analysis, using VADER lexicon. As mentioned in section 2.1.2.1, VADER is specialised in social media data, a domain that it performs exceptionally well.

Having completed the data pre-processing phase, each tweet was split into tokens, forming a bag of words. Each bag of word was then fed to VADER, which assigned each token a sentiment score. Basically, VADER sentiment analysis employs a dictionary that assigns lexical features to sentiment scores.

Next, each token went through a rule-based checker, in order to identify the intensity of their sentiment. The intensity of the polarity is determined by a number of heuristics, apart from lexical features. In particular, punctuation is very important to VADER, as it considers a sentence more emotionally intense, when it includes punctuation, such as exclamation marks. Capitalisation of words is also taken into account, as capitalised words are considered to convey a more intense sentiment. Another heuristic is the use of degree modifiers in a sentence. Degree modifiers alter the emotional intensity of a text, according to their distance from the specific word that they modify. Furthermore, VADER considers the change in polarity of two clauses, when they are connected with a “but”. Particularly, VADER boosts the strength of the sentiment of the second clause, as it is considered to

affect more the polarity of the whole text. The last heuristic is negation, where a tri-gram is examined, as it can change completely the sentiment score.

The total sentiment score of each tweet is obtained by summing up the sentiment score of each token in the tweet.

Sentiment score of individual words is calculated on a range from -4, which is the most negative, to +4, which is the most positive. A sentiment score of 0 represents a neutral sentiment. However, the sentiment score of a sentence ranges between -1 and +1. That is, the total sentiment score of a sentence, which is calculated by the sum of scores of each word in the sentence, is normalised by the following model:

$$\frac{x}{\sqrt{x^2 + a}} \quad (6)$$

where x is the sum of sentiment scores of the words in a sentence and a is a normalisation parameter whose value is set to 15. As x increases, the sentiment score tends to move closer to -1 or +1. Therefore, we understand that a document with a large number of words is likely to have a sentiment score close to the extreme values of the range.

4.2.3 Machine Learning Sentiment Analysis

Training Phase

Supervised machine learning entails a labelled dataset that is used for training purposes of the classifier. The training dataset consists of instances that include an input object, as well as a label or a class. At first, an algorithm analyses the labelled data, then it extracts features, that include the required information from the input data, so that the sentiment analysis can be accomplished, by using only a representation of the initial data. Last, the algorithm creates a function, that is later used in the classification of unseen data or the testing dataset.

In our study, the labelled dataset that we choose to train our classifier on, consists of tweets about six major airline companies in US and can be found here. The "*Twitter US Airline Sentiment*" dataset that is used for training purposes, consists of 14.641 tweets, which show how travellers expressed their feelings on Twitter in February 2015.

As far as the training dataset from Kaggle is concerned, first, we exclude all the tweets that were labelled as neutral, in order to enable the algorithm to focus on content with exclusively positive or negative sentiment. Thus, the final training dataset that we use, consists of 11.541 tweets, classified

as either positive or negative. Consequently, we proceed with the data pre-processing, by removing HTML tags and URLs, digits and punctuation, stopwords and noise. We further tokenised the text and applied part-of-speech tagging, in order to mark each token with the part of speech that it belongs to, i.e. noun, adjective, verb and adverb. Moreover, we create unigrams and bigrams of the tokens, in order to improve the quality of text classification. The last step in data pre-processing was the lemmatisation of the tokens.

Overcome Class Imbalance

As shown in *Figure 4* our training dataset is highly imbalanced. Class imbalance happens when the instances of a certain class (in our case, negative) by far outnumber the instances of another class (positive). Class imbalance is considered a serious problem, as it can cause significant bias towards the majority class. Moreover, it can decrease the machine learning algorithm performance and increase the number of false negatives.

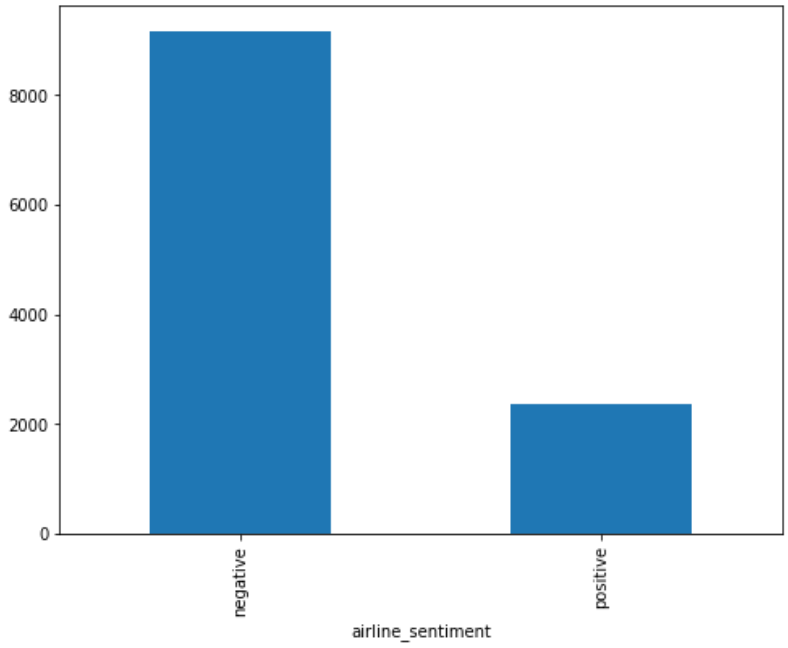


Figure 4. Statistics of the training dataset from Kaggle. The bar on the left represents the number of negative tweets in the dataset and the bar on the right represents the number of positive tweets.

In order to alleviate this problem, we employ an undersampling technique, in advance of the algorithm training. Undersampling is the removal of a number of instances that belong to the majority class, with the purpose of equating the number of instances between the two classes. Specifically, we perform undersampling based on the Neighbourhood Cleaning Rule (NCL). NCL removes majority instances, in the following way: For every instance in the training set, NCL finds three nearest neighbours. If the instance belongs to the majority class and the classification given by its three neighbours is the opposite of the class that the instance belongs, then the instance is removed. On the other hand, if the

instance belongs to the minority class and the classification given by its three neighbours is again the opposite, the three neighbours that belong to the majority class are removed (Laurikkala, 2001).

Feature Extraction/Vectorisation with Bag of Words

Proper feature extraction is critical for the performance of a machine learning classifier. A simple method that is extensively used in the bibliography, is the Bag of Words model. A Bag of Words is a representation of text, that illustrates the frequency of occurrence of each token in the text. A Bag of Words model does not take into account any grammar or even any sequence of the words.

In our study, we use a Bag of Words model, that includes both unigrams and bigrams that are represented by the frequency of occurrence of each unigram or bigram in the text. In terms of feature selection criteria, we have already narrowed down our features set by excluding common words, in the preprocessing step of removing stopwords. Moreover, we have set a threshold in minimum number of documents (tweets), that a feature appears. That threshold is set to 20 tweets. In that way, we remove terms that appear too infrequently. No upper limit in frequency of features appearance is set. Our Bag of Words model consists of in total 748 features. Therefore, each of 11.541 tweets is represented by 748 features, representing the Bag of Words score for different unigrams and bigrams.

Model Training and Validation

As soon as the features are generated, each tweet is represented as the bag of words of all tweets. Next, the dataset is split into two subsets, the train set, which is the 70% of the total dataset and the test set, which is the remaining 30% of the total dataset. The purpose of dividing the dataset into two subsets is to avoid overfitting. Overfitting occurs when a classifier is both trained and tested in the same data. Thus, it scores perfectly, when classifying already seen data, however it totally fails in the classification of unseen data.

Moving forward, we are ready to train our model, using different classifiers. In particular, we employed the Gaussian Naïve Bayes, the Support Vector Classifier and the Random Forest Classifier. Finally, the best scoring classification model is saved, in order to be applied on the dataset that includes all Ryanair's unlabelled tweets.

Prediction Phase

Having completed the model training process, we are ready to proceed with the actual sentiment analysis of our Ryanair tweets. During the prediction phase, the Ryanair dataset is preprocessed exactly as the training dataset, obtained by Kaggle. Next, we incorporate the Bag of Words model, as well as the trained classification model, both created during the processing of our training Kaggle dataset.

Subsequently, we proceed with the final predictions on Ryanair's tweet sentiment, by employing our previously trained classification model. Thus, each tweet is now classified either as positive or as negative.

In order for our prediction model to be able to assign a numerical value of polarity to each tweet, we utilise the probability of a tweet to be negative and the probability to be positive. In particular, the polarity score of each tweet is given by the following formula:

$$Sentiment = \frac{(Probability\ of\ a\ tweet\ to\ be\ negative - 0.5)}{(-0.5)} \quad (7)$$

4.3 Statistical Analysis

4.3.1 Independent Variables

This part of our study introduces the independent variables that are used in the statistical analysis. As already described in section 3.1, the purpose of our statistical analysis is to explain the relationship between Twitter and Ryanair's business performance. Twitter is represented by a number of metrics that we use as our independent variables, which are also extensively used in the literature. Specifically, the *extracted sentiment or polarity of the tweets*, the *volume of tweets*, the *volume of positive tweets*, the *volume of negative tweets* and the *ratio (volume of positive tweets) / (volume of negative tweets)* belong all to the set of our independent variables X_t.

All of the above independent variables derive from the aggregation of daily data by month, since the dependent variable of Ryanair passengers Y_t is also calculated per month. Moreover, all variables, except for the volume of tweets, have been computed using both the lexicon-based and the machine learning sentiment analysis. Therefore, we use in total, nine independent variables, as presented in *Table 3.*

Independent Variables	*Type*	*Range*
vader_polarity	Real	(-1) - (+1)
ml_polarity	Real	(-1) - (+1)
tweets_volume	Integer	>= 0
vader_positive_tweets	Integer	>= 0
vader_negative_tweets	Integer	>= 0
vader_pos_neg_ratio	Real	>= 0
ml_positive_tweets	Integer	>= 0

ml_negative_tweets	Integer	>= 0
ml_pos_neg_ratio	Real	>= 0

Table 3. Independent Variables

4.3.2 Regression Analysis

Having completed the sentiment analysis, we proceed with the statistical analysis, in order to investigate the relationship between buzz in Twitter and Ryanair's performance. In particular, we aim to research the role of the volume of tweets, as well as the sentiment that they convey, in the volume of sales of the company, as expressed with the number of passengers. The investigation is carried out with the use of Regression Analysis. We implement a Linear Regression Analysis, using the method of Ordinary Least Squares (OLS). The method of OLS computes the regression line, by minimizing the sum of squared distances of the data points and the regression line. We implement a Simple Linear Regression, as we have observed highly correlated independent variables.

The Regression model that is used, is given by the following equation:

$$Y_t = \beta_0 + \beta_1 X_t + e_t \quad (8)$$

where Y_t illustrates the dependent variable (number of passengers), X_t represents the independent variables (Twitter metrics), β_0 is the intercept term and β_t is the slope coefficient. Last, e_t is the error term.

4.3.3 Vector Autoregression (VAR)

In the last part of the analysis, we investigate the interdependencies between Ryanair's business performance and Twitter metrics, using a Vector Autoregressive Analysis (VAR). Vector Autoregressive models were introduced by Sims (1980) and contain both contemporaneous and lagged values of all variables, which are all considered end. The simplest form of a VAR model is the one that we use in our empirical analysis, the bivariate VAR. The bivariate VAR involves two variables, y_t and x_t, and each one of them depends on the present and past values of the other, as well as an error term.

A bivariate Vector Autoregressive model is given by the following equations:

$$y_t = a_{10} + a_{11}y_{t-1} + \cdots + a_{1k}y_{t-k} + \beta_{11}x_{t-1} + \cdots + \beta_{1k}x_{t-k} + e_{1t}$$

$$x_t = a_{20} + a_{21}x_{t-1} + \cdots + a_{2k}x_{t-k} + \beta_{21}y_{t-1} + \cdots + \beta_{2k}y_{t-k} + e_{2t} \quad (9)$$

In our analysis, we choose to examine the relationship between variables y_t and x_t and their first lagged values, i.e. y_{t-1} and x_{t-1}.

One substantial weakness of VAR methodology is the high complexity in interpretability of the estimated models, due to the large number of parameters that can be entailed and the VAR methodology a-theoretical nature. By a-theoretical, we mean that there is no underlying theoretical background that the model's construction and use is based upon. However, in order to overcome VAR methodology's weakness, we also implement in our analysis the following statistics tests: block significance tests and impulse responses.

Granger Causality

Our purpose is to identify the significance in effects that each variable has in another. In order to do so, we implement Granger causality tests, which were introduced by Granger (1969). Granger Causality tests are a statistical hypothesis testing method, which addresses the question whether changes in a variable y_t cause changes in a variable x_t. By definition, Granger Causality does not really imply causality, i.e. movements of one variable cause movements of another. Its meaning is rather of an increased predictability of one variable's contemporaneous value, from another's lagged value. That is, if past values of a variable x_t explain (in part) and improve the prediction of current values of a variable y_t, better than past values of variable y_t explain and improve the prediction of current values of variable y_t, it is said that variable x_t Granger causes variable y_t. Statistically, we jointly test the significance of the coefficients of variable x_t in the equation of variable y_t and vice versa. Granger causality can be uni- or bi-directional.

Impulse Responses

Further to our analysis, we attempt to identify the type of statistically significant effect that changes in the value of a variable x_t have in a variable y_t, in a VAR system. By type of effect, we mean a positive or a negative effect. Furthermore, we aim to investigate the responsiveness of a dependent variable y_t, when a shock is applied to each of the variables. In that hope, we examine the VAR's impulse responses.

5. Results

This chapter details the results from the sentiment analysis and the statistical analysis that we previously conducted (see section 4). The first part of this chapter graphically presents the volume of the collected tweets per month, as well as the number of passengers per month. In the next part of this chapter, the results of both the lexicon-based and the machine learning sentiment analysis are thoroughly described. Finally, in the last section of this chapter, we detail our findings from the statistical analysis.

5.1 Data

As discussed in the methodology (see section 4.1) 382.425 tweets that refer to Ryanair were collected, all published from July 2017 to September 2019. For the same period, we collected information about the number of passengers of Ryanair per month.

Figure 5 demonstrates a visual representation of the volume of tweets per month, over the span of the research. The number of tweets has been transformed to logarithmic form, for normalisation purposes and easier interpretability of patterns in the data.

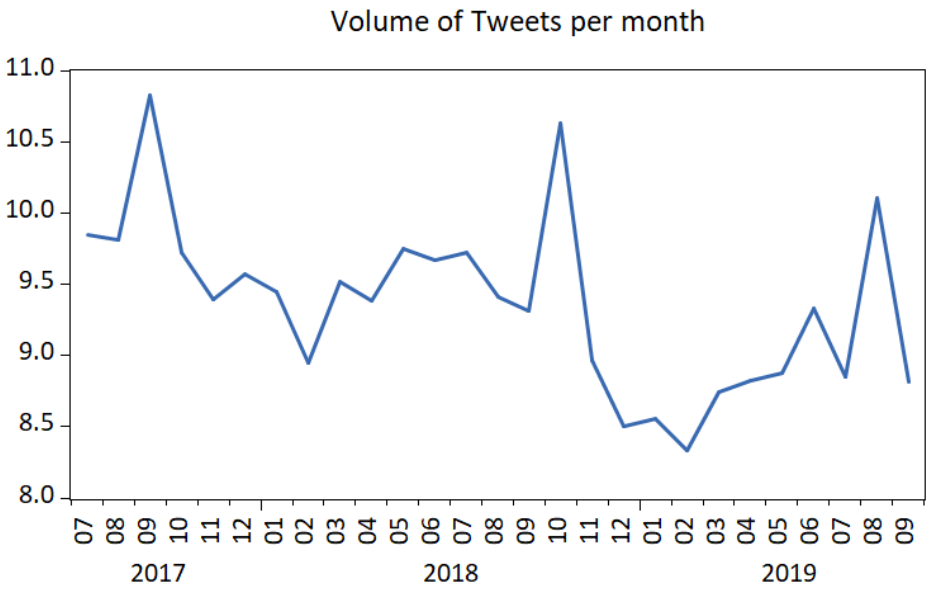


Figure 5. Volume of Tweets per month

As it can be observed, the volume of tweets displays an unusual behaviour on three particular periods, as the spikes in the graph indicate. We presume that each spike corresponds to a major event that concerned Ryanair during that period and highly raised the activity of users on Twitter. In September 2017, where the first spike is presented, BBC News reports major Ryanair flight cancellations, which affected thousands of travellers. The second spike appears in October 2018, when according to BBC News, a racial abuse incident took place on a flight from Barcelona to London. The incident was video-recorded and went viral online, while Ryanair has been criticised for failing to remove the abusive passenger from the flight. The third spike appears in August 2019, when Ryanair failed to prevent pilot strikes, that threatened a lot of flights, as the Guardian reports. The above incidents are presented in *Figure 6*.

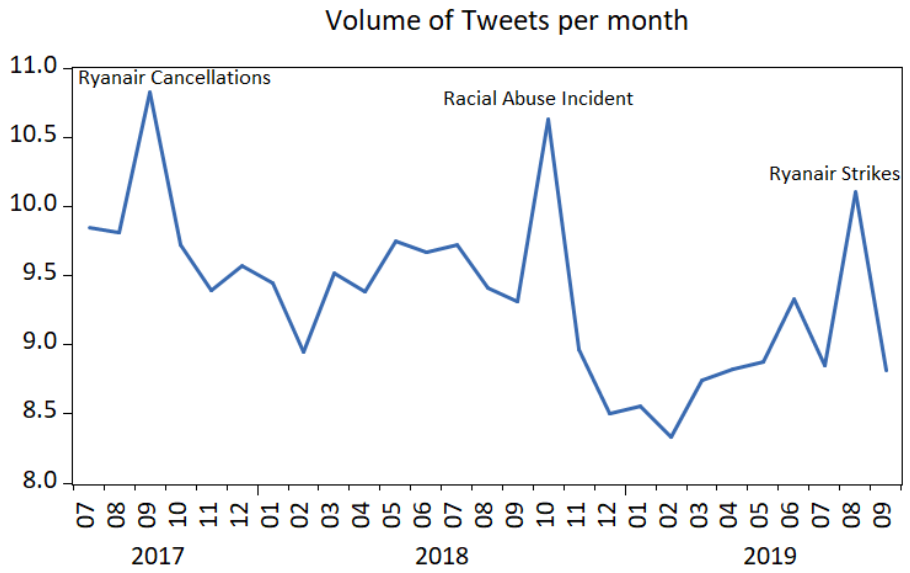


Figure 6. Spikes in Tweets Volume

In terms of the traffic data, this indicates the number of Ryanair's passengers per month, in millions. *Figure 7* demonstrates the volume of traffic, throughout the period of examination.

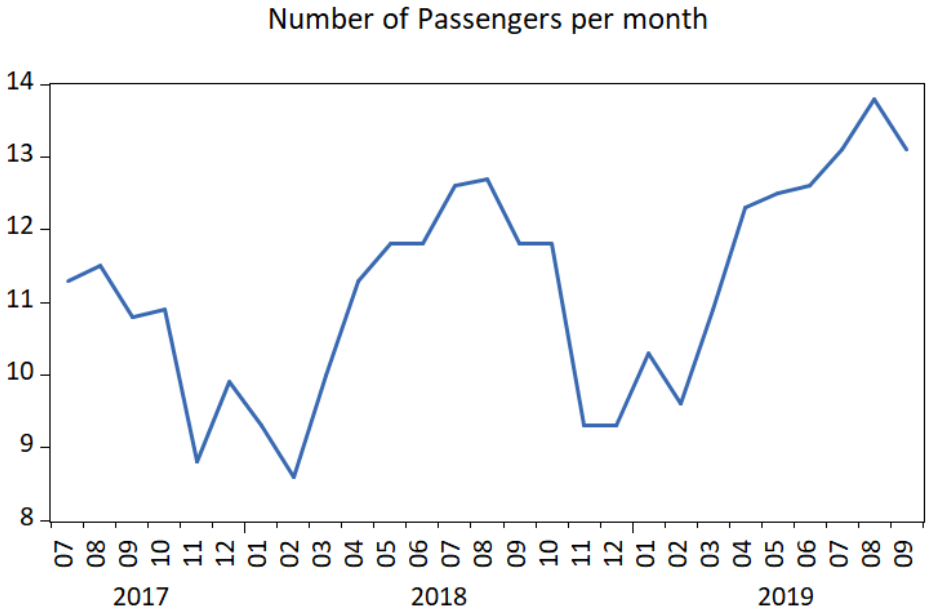


Figure 7. Number of Passengers per month

Figure 8 presents both the volume of tweets and the number of passengers, per month. Both series have been transformed to logarithmic form. Interestingly, the volume of tweets and the number of passengers seem to have a similar tendency, almost throughout the period of examination. That could be explained by the assumption that as the number of passengers grows, there is an increasing Twitter activity, as more and more people tweet about Ryanair.

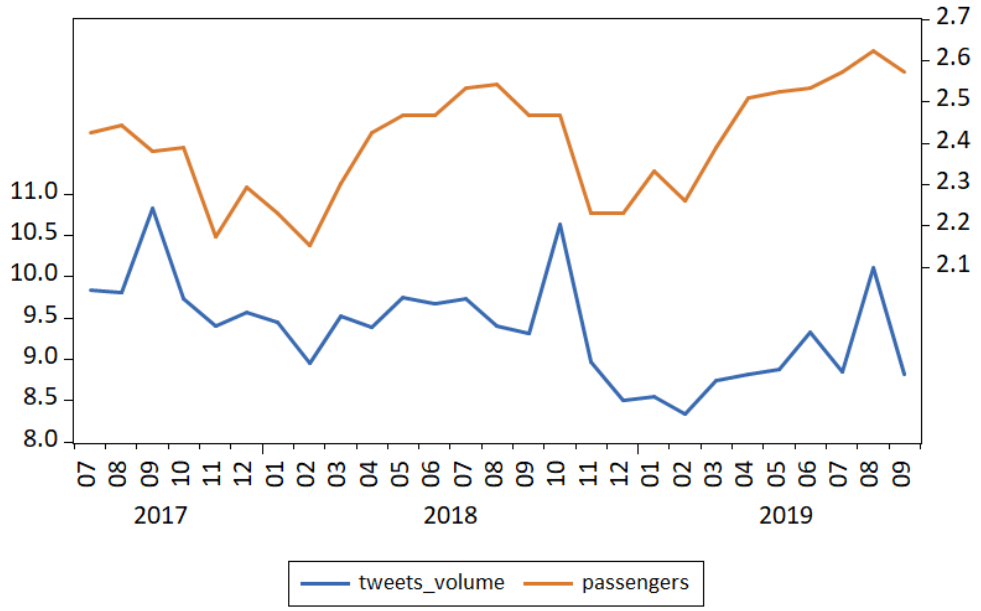


Figure 8. Volume of Tweets and Passengers

5.2 Sentiment Analysis

5.2.1 Lexicon-based Sentiment Analysis

After collecting the Twitter data and preprocessing it, we proceed with the sentiment analysis, using the VADER lexicon. The VADER sentiment analyser returned in total, 142.923 positive, 136.854 negative and 102.648 neutral tweets. As already discussed, tweets sentiment ranges from -1 to +1. We consider a tweet positive, when its sentiment polarity is greater than +0.2. A tweet with sentiment score smaller than -0.2 is considered negative and tweets with sentiment score between -0.2 and +0.2 are classified as neutral. The arithmetic mean of the sentiment per tweet, as computed by VADER is ~0.015. All positive, negative and neutral tweets were then aggregated by month, in order to further examine their fluctuations over time, as shown in *Figure 9.*

Figure 9 shows that the volume of positive tweets fluctuates throughout the examination period, presenting a few spikes in September 2017, with 15.894 tweets and in August 2019, with 9.956 tweets. The same behaviour displays the volume of negative tweets, although their highest spike is presented

in October 2018, with 26.046 tweets. Neutral tweets increase sharply in September 2019, with 14.912 tweets. We can also see that the total volume of tweets has a declining tendency from November 2018 until May 2019. Interestingly, that does not happen during the same period in 2017, therefore we cannot blame the seasonality for this decrease.

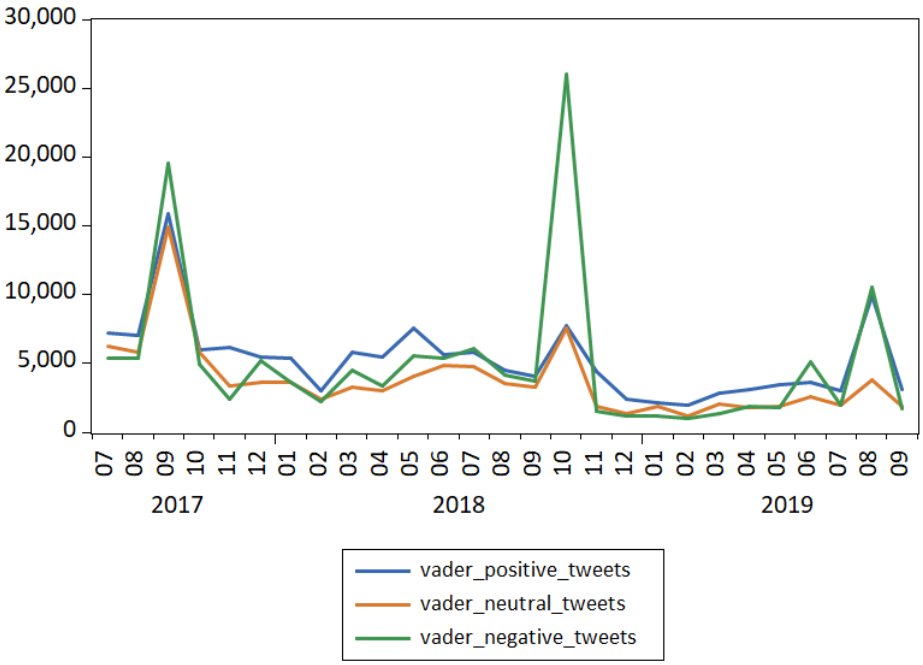


Figure 9. VADER Positive, Negative, Neutral Tweets

Figure 10 illustrates the distribution of the polarity of tweets. As we can observe, the sentiment of most tweets ranges from 0 to approximately -0.20.

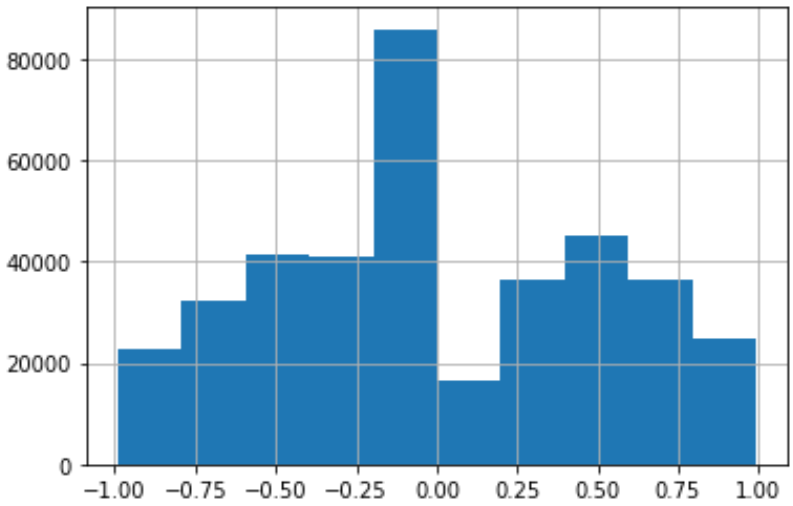


Figure 10. Distribution of Sentiment (VADER)

5.2.2 Machine Learning Sentiment Analysis

Having completed the sentiment analysis using a lexicon-based approach, we decided to analyse the tweets, using a machine learning approach, as well. As described in the Methodology (see section 4.2.3), our classification model is trained on a pre-labelled dataset, that was obtained by Kaggle. However, before proceeding with the actual model training, we tested three different classification algorithms, in order to choose the one that had the best performance. In *Table 4* we present the metrics of performance of the three algorithms.

Classification Report

Classifier	*Precision*	*Recall*	*F1-Score*	*Accuracy*
Gaussian Naive Bayes	0.68	0.74	0.64	65.9%
Support Vector	0.86	0.70	0.73	84.2%
Random Forest	0.90	0.89	**0.90**	**92.5%**

Table 4. Performance of Machine Learning Algorithms

Random Forest appears to have the best performance, with F1-score = 0.90 and Accuracy = 92.5%. *Table 5* illustrates the prediction metrics of Random Forest in the training Kaggle dataset. We can see that our classification model scores high in both precision and recall for both classes (positive and negative). Our model performs well in the classification of instances (tweets) that are labelled as positive, although they belong to the minority class, after we implement an undersampling technique, as discussed in section 4.2.3, in order to address the class imbalance problem.

Random Forest Classification Report

	Precision	*Recall*	*F1-Score*
0 : {negative}	0.94	0.96	0.95
1 : {positive}	0.87	0.82	0.84
macro average	0.90	0.89	0.90

Table 5. Random Forest Prediction Metrics

The machine learning sentiment analysis returned in total, 88.497 positive tweets, 255.377 negative and 38.551 neutral tweets. As described in section 5.2.1, positive tweets are tweets with sentiment score greater than +0.2, negative are tweets with sentiment score smaller than -0.2 and all tweets with scores within -0.2 and +0.2 are considered as neutral. The arithmetic mean of the sentiment per tweet, as calculated by our classification model, is -0.32. Again, tweets polarity was aggregated to monthly averages.

Figure 11 depicts the variation of positive, negative and neutral sentiment of Ryanair tweets, as extracted using the Random Forest classifier. We can observe that the negative sentiment outnumbers positive and neutral. That is interesting to note, since the results of VADER sentiment analyser are quite different from our classification model. However, we do see that the trends in data, in *Figure 9* and *Figure 11* are quite similar. In particular, we notice the spikes in September 2017, when negative tweets reached their peak at 36.177, October 2018, with 28.361 negative tweets and August 2019, with 19.089 negative tweets. During the same periods, positive tweets exhibit a similar behaviour, however to a smaller degree.

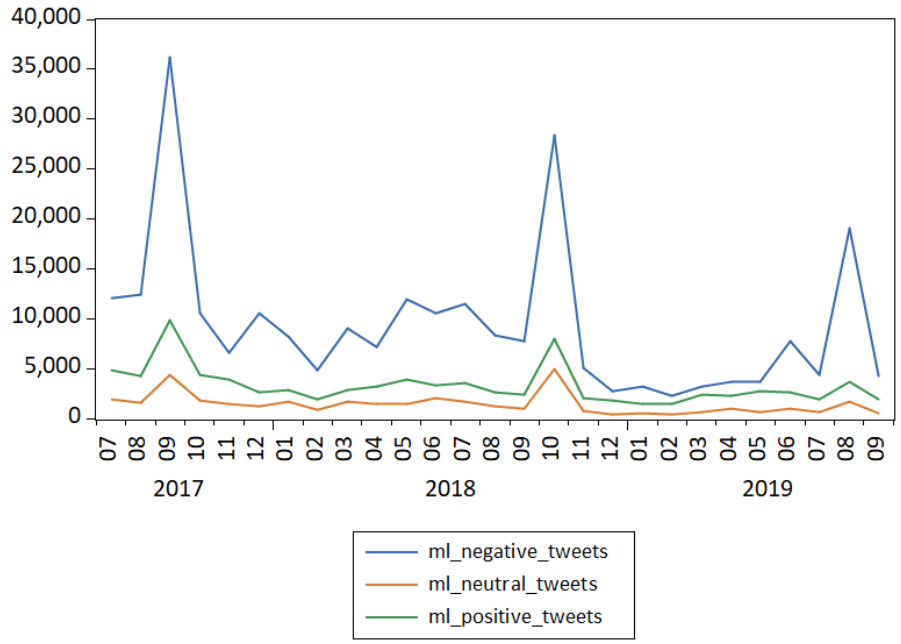


Figure 11. ML Positive, Negative, Neutral Tweets

Figure 12 depicts the distribution of sentiment in tweets, as measured by our classification model. Most tweets appear to convey a negative sentiment, with a polarity score ranging from -0.6 to -1.

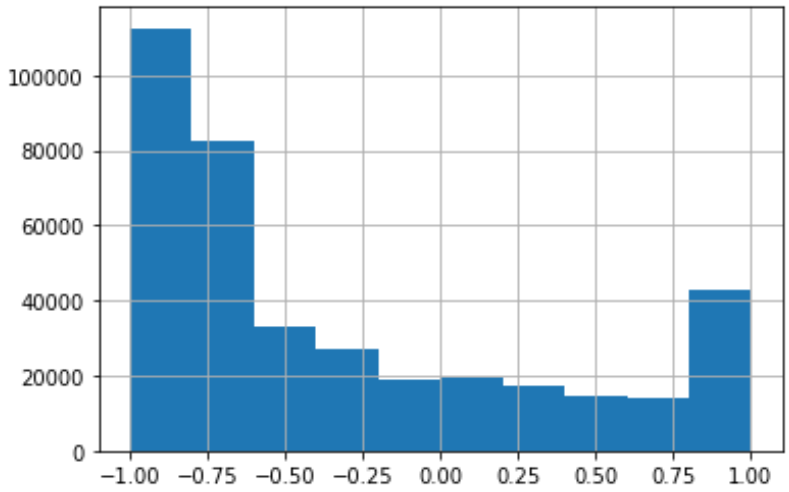


Figure 12. Distribution of Sentiment (ML)

Through our machine learning analysis, we can draw conclusions about the importance of features that were used. *Figure 13* depicts the top 25 contributor features, which acted as indicators of positive and negative sentiment. Features such as “cancel”, “bad” and “delay” indicate negative sentiment. On the other hand, features like “amazing”, “awesome” and “thank you” indicate positive sentiment.

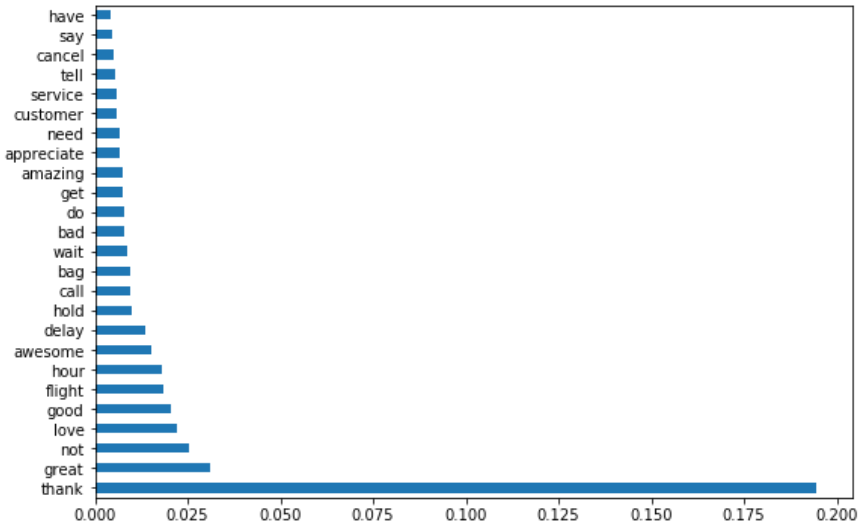


Figure 13. Top 25 Contributor Features

Last, we attempted to shed light in instances that our prediction model misclassified. For that purpose, we compared the actual labels with the predicted labels of tweets that belong to the pre-labelled dataset obtained by Kaggle. From a total of 11.541 tweets, 1.089 tweets were misclassified by our classification model. The majority of misclassifications happened on tweets that were actually negative. In particular, 965 originally negative tweets were classified as positive. On the other hand, only 124 tweets that were actually positive, were misclassified as negative. Hence, it is obvious that our classification model incorrectly classifies tweets that actually convey negative sentiment.

Specifically, we manually cross-checked a number of tweets that were misclassified as positive. According to our findings, a remarkable number of misclassified tweets included sarcasm, which our classification model was not able to detect.

5.3 Statistical Analysis

5.3.1 Regression Analysis

Table 5.3.1:
Regression Analysis Results

	Dependent Variable: Δlog (*Passengers*$_t$)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ln(tweets_volume$_{t-1}$)	-0.08 (-2.87)								
vader_polarity $_{t-1}$		0.46 (3.38)							
ml_polarity$_{t-1}$			0.36 (2.19)						
ln(vader_positive$_{t-1}$)				-0.07 (-2.01)					
ln(vader_negative$_{t-1}$)					-0.06 (-3.04)				
vader_ pn_ratio$_{t-1}$						0.07 (2.53)			
ln(ml_positive$_{t-1}$)							-0.09 (-2.36)		
ln(ml_negative$_{t-1}$)								-0.07 (-2.96)	
ml_ pn_ratio$_{t-1}$									0.27 (2.45)
R-squared	0.255	0.323	0.167	0.143	0.27	0.211	0.189	0.267	0.201

Note: t-statistic in parenthesis

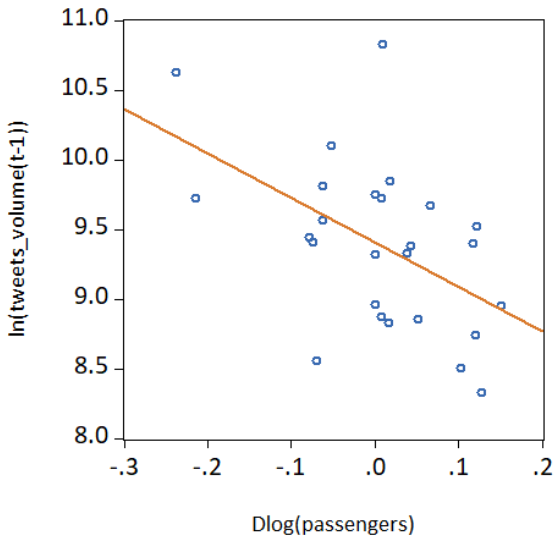
Table 5.3.1 summarizes the regression analysis results. All twitter metrics affect significantly the growth rate of passengers of the next time period (one month forward). Among the predictors, sentiment, as well as positive/negative tweets ratio, both as measured by VADER and by Random Forest classifier, have a positive effect on the dependent variable. All four positive signed variables capture the sentiment that tweets about Ryanair convey. This result confirms our assumption that the growth rate of passengers grows, when the sentiment on Twitter tends to be positive. That is also confirmed by the relationship of positive/negative ratio and passengers growth rate, since positive/negative ratio and sentiment increase proportionally.

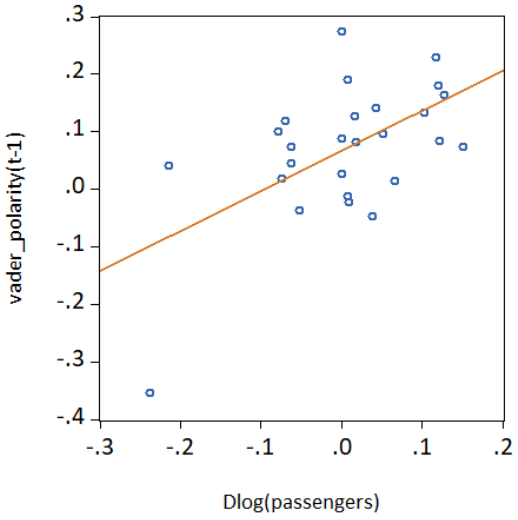
On the other hand, the effect of variables that are related to the volume of tweets is negative. Specifically, volume of tweets, volume of positive tweets and volume of negative tweets, as measured by VADER and by Random Forest classifier seem to have a negative effect on the growth rate of number of passengers. It is interesting to stress that although we would expect the growth rate of passengers to increase, as long as the volume of tweets increases, our analysis indicates otherwise. That could be happening due the increase of negative tweets, which contributes significantly to the increase of the overall volume of tweets (positive and negative). Therefore, we assume that negative tweets can decelerate the growth rate of passengers.

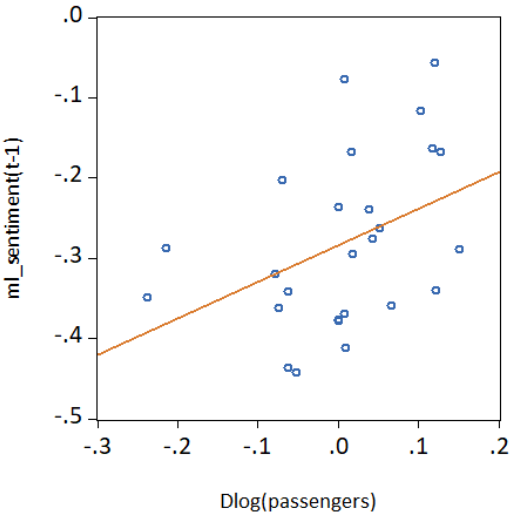
According to R-squared goodness-of-fit metric, among the predictors, VADER polarity has the best performance (32.3%). Volume of tweets explain the response variable by 25.5%. We also see that the volume of negative tweets as measured by VADER and Random Forest explain the growth rate of passengers by 27% and 26.7% respectively. The two percentages are very close to each other, although the metrics have been calculated using different methods, which means that there is an agreement between the two metrics. We observe the same phenomenon with the positive/negative ratio. R-squared for positive/negative ratio calculated by VADER is 21.1%, while calculated by our classification model is 20.1%.

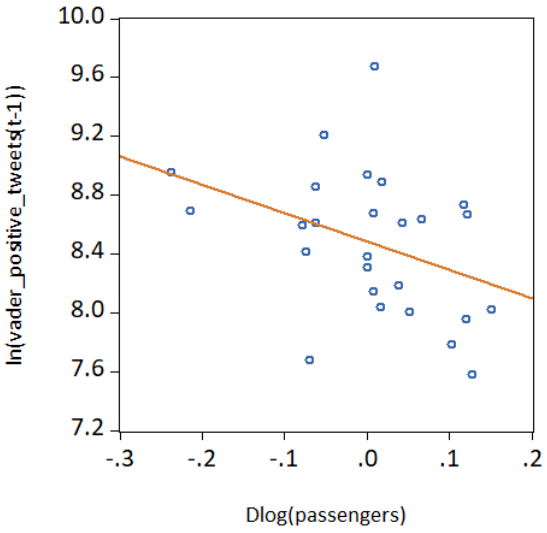
The following scatter plots (*Figures 14 – 22)* visually present the relationships between each predictor with the dependent variable, along with the regression line that best fits the data.

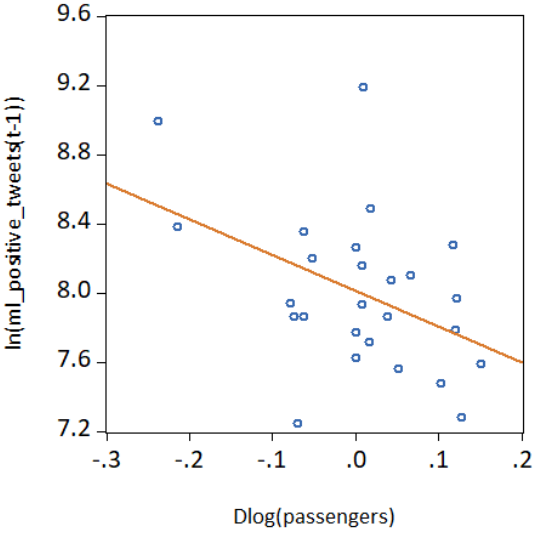
Figures 14 - 22

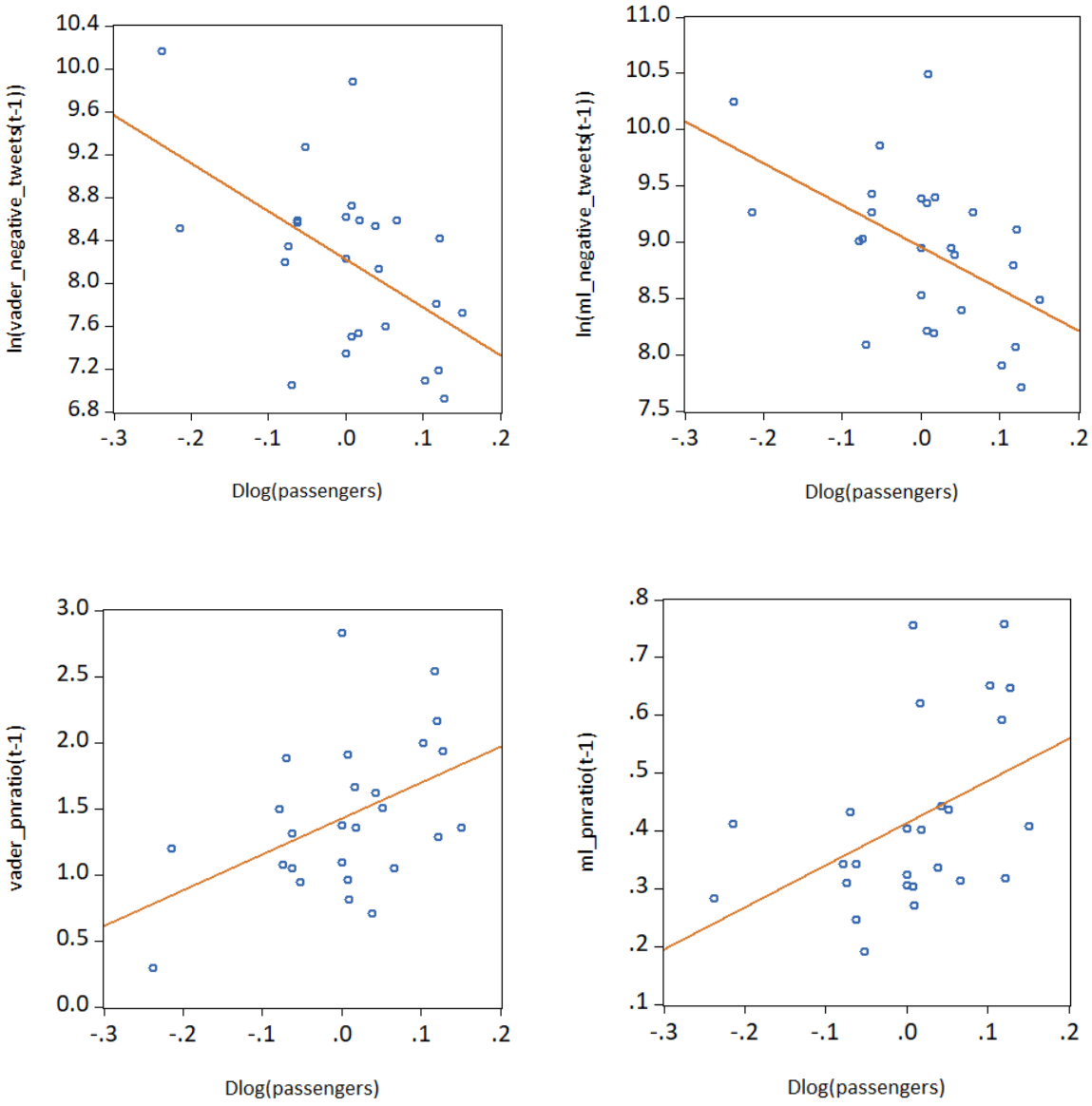



vader_polarity(t-1)
.3
.2
.1
.0
-.1
-.2
-.3
-.4
-.3
-.2
-.1
.0
.1
.2
Dlog(passengers)


ml_sentiment(t-1)
.0
-.1
-.2
-.3
-.4
-.5
-.3
-.2
-.1
.0
.1
.2
Dlog(passengers)


ln(vader_positive_tweets(t-1))
10.0
9.6
9.2
8.8
8.4
8.0
7.6
7.2
-.3
-.2
-.1
.0
.1
.2
Dlog(passengers)


ln(ml_positive_tweets(t-1))
9.6
9.2
8.8
8.4
8.0
7.6
7.2
-.3
-.2
-.1
.0
.1
.2
Dlog(passengers)



5.3.2 Vector Autoregression (VAR)

5.3.2.1 Lag Length Selection

As discussed in section 4.3.3, we implement a Vector Autoregressive model, in order to examine the interdependencies between all variables that reflect Twitter metrics and Ryanair's performance. However, first, we investigate the optimal number of lags that will be used in our VAR models, with the help of the Schwarz information criterion (SIC). Our findings indicate that the optimal lag length is 1. Therefore, we employ VAR models, which include both the contemporaneous values of our variables and lagged values of one period (month) past.

5.3.2.2 Granger Causality

Table 5.3.2.2

Granger Causality Results

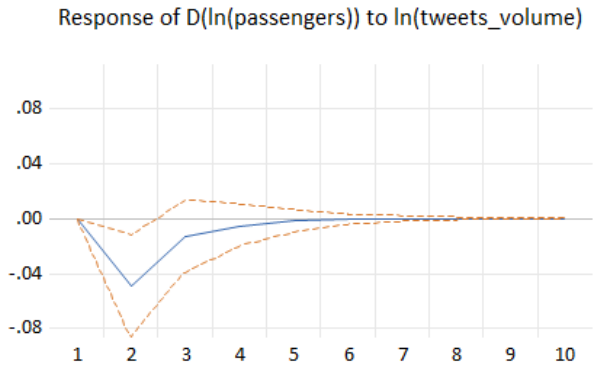
A		p-value	**B**		p-value
H_0	ln(tweets_volume) **/→** Δln(passengers)	0.0044***	**H_0**	Δln(passengers)**/→** ln(tweets_volume)	0.9390
	(vader_polarity) **/→** Δln(passengers)	0.0013***		Δln(passengers)**/→** (vader_polarity)	0.4310
	(ml_polarity) **/→** Δln(passengers)	0.0371**		Δln(passengers)**/→** (ml_polarity)	0.7205
	ln(vader_positive) **/→** Δln(passengers)	0.0415**		Δln(passengers)**/→** ln(vader_positive)	0.6203
	ln(vader_negative) **/→** Δln(passengers)	0.0034***		Δln(passengers)**/→** ln(vader_negative)	0.7721
	(vader_pnratio) **/→** Δln(passengers)	0.0145**		Δln(passengers)**/→** (vader_pnratio)	0.6572
	ln(ml_positive) **/→** Δln(passengers)	0.0153**		Δln(passengers)**/→** ln(ml_positive)	0.8283
	ln(ml_negative) **/→** Δln(passengers)	0.0035***		Δln(passengers)**/→** ln(ml_negative)	0.8674
	(ml_pnratio) **/→** Δln(passengers)	0.0187**		Δln(passengers)**/→** (ml_pnratio)	0.6770

Note: **/→** implies does not Granger cause. *** = 1% significance level; ** = 5% significance level; * = 10% significance level.

As described in section 4.3.3, Granger Causality tests the Null Hypothesis that variable x_t does not Granger cause variable y_t. Table 5.3.2.2 presents the Granger Causality results obtained from the Vector autoregression (VAR) analysis. Results suggest that Granger Causality is unidirectional. P-values in panel (A) indicate that all Twitter metrics are statistically significant predictors of future passenger growth rate, as the Null Hypothesis is rejected in any case. On the contrary, p-values in panel (B) suggest that there is no Granger Causality directed from the growth rate of passengers to Twitter metrics. As it is observed, in panel (A), the Null Hypothesis is more strongly rejected in the case of VADER polarity, VADER negative tweets, tweets volume and Random Forest negative tweets, where the p-value is less than 1%.

5.3.2.3 Impulse Responses

Figures 23 -31



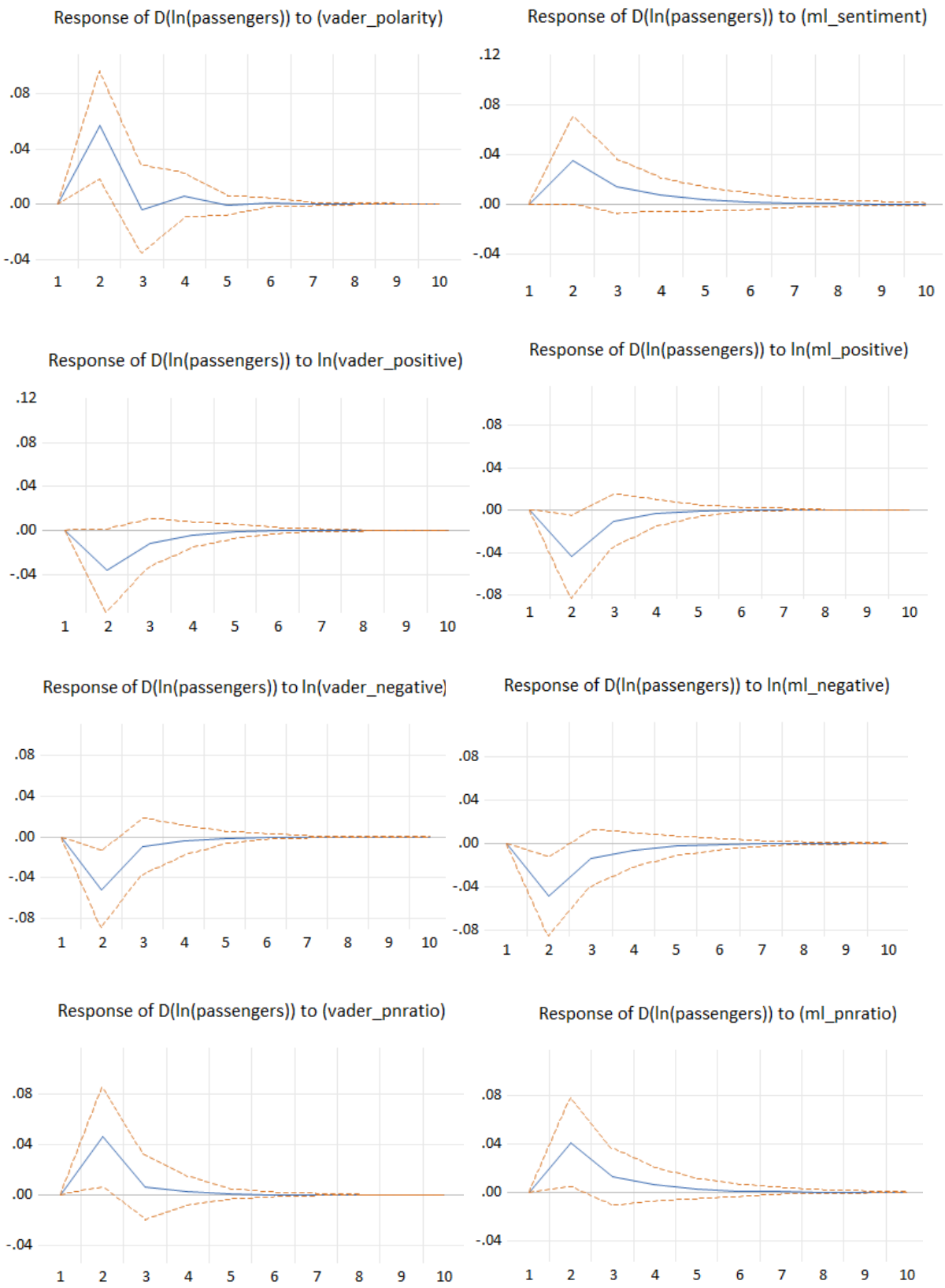


Figure 23 depicts the passengers' growth rate response to a shock in the volume of tweets. Specifically, the impact of the shock is negative, and we can observe that the shock in passengers shows up with one period lag. That is, if a shock is applied on tweets volume in period 1, the negative impact on the

growth rate of passengers will become visible in period 2. After period 2, we see that impact of the shock decreases and it finally disappears in period 5. The same pattern is observed in *Figures 26 -29,* when the shock is applied on the number of positive and negative tweets. The result is robust when both VADER and machine learning classification approaches are considered.

Figure 24 illustrates the response of the dependent variable, when a shock is applied on the sentiment, obtained from VADER analysis. In that case, we observe a positive effect, meaning that an increase in sentiment on Twitter is followed by an increase in the passengers' growth rate. Once again, there is a one period lag between the time that a shock is applied and the time that its impact starts to appear. After period 2, we observe a decline in the effect of the shock and in period 5, it dies out. *Figure 25* and *Figures 30 - 31* present a similar behaviour in the response to a shock.

Each independent variable is closely connected with the sentiment that people express in Twitter about Ryanair. Positive or negative sentiment is triggered by events that happen in the real world and affect the experience of passengers in regard to Ryanair as a brand. These experiences are then reflected in Twitter, and in our research, are expressed in the form of our independent variables. Thus, impulse responses describe the impacts of shocks in the independent variables, hence in real world events and experiences, upon the growth rate of Ryanair passengers.

6. Conclusions and Future Work

In this chapter, we summarise our findings on the relationship between Twitter and Ryanair's business performance. Next, we discuss the limitations of the study and the future perspective in this area. Last, we present the final concluding observations on this thesis.

6.1 Summarising the findings

The purpose of the study was to investigate the relationship between Twitter sentiment and Ryanair's number of passengers, per month. We aimed at contributing to this area, by employing different methods of sentiment analysis and by explaining how variables related to Twitter affect the sales of a low-cost airline company. First, we performed sentiment analysis on the tweets in order to extract information about the sentiment that they conveyed. We employed a lexicon-based methodology, using the VADER lexicon. Moreover, we created a machine learning classification model, using the Random Forest classifier. Both methods classified the tweets into three categories; positive, negative and neutral and assigned to them a sentiment score from -1 to +1. Next, we examined the relationship of tweet sentiment and growth rate of Ryanair's passengers, with the use of simple linear regression. In addition, we employed Vector Autoregressive models, in order to examine the interdependencies between the variables. We performed Granger Causality analysis in order to infer exogeneity of the

independent variables and Impulse Response analysis, to assess the reaction of the dependent variable, in changes in Twitter.

The research questions that we tried to address through our research were the following:

1. *Can the sentiment and volume of Twitter messages (tweets) have an impact on Ryanair's monthly traffic numbers?*
2. *Can tweets be used as predictor variables?*

The regression analysis results indicate that Twitter variables are highly correlated with the future passengers' growth rate. The high R-squared goodness-of-fit metric shows that each independent variable explains the growth rate of passengers to a large extent. Moreover, Granger Causality analysis results indicate the exogeneity of Twitter metrics, as a unidirectional Granger causality from Twitter metrics to passengers' growth rate is discovered. Passengers' growth rate reacts significantly to shocks in Twitter, as resulted from Impulse Responses analysis. Taking these findings into consideration, we can accept that the sentiment and volume of tweets have an impact on Ryanair's monthly traffic numbers. Additionally, all above three inferences imply that tweets can be used as predictors in future sales forecasting models.

6.2 Limitations

Although we consider our attempt to answer the research questions successfully, it is a matter of fact that we faced some serious limitations during our research, which affected our findings. These limitations are further described, as follows:

- *The exclusion of non-English tweets*

In this study, only tweets written in English were collected and analysed. Therefore, all non-English tweets about Ryanair were excluded. That constitutes a serious limitation of our research, as Ryanair flies throughout Europe and carries millions of passengers that do not speak English, but do tweet about the airline carrier. (Beleveslis et al., 2019)

- *Quantity of Twitter Data*

The present study has not been conducted in collaboration with Ryanair, therefore the collected data do not represent the total amount of tweets that have been published during the span of research. Moreover, due to computer storage capacity, the size of our dataset was further limited. Thus, it would be appropriate to bear in mind that our findings might be influenced. (Oikonomou & Tjortjis, 2018)

- *Selection of Twitter as a social media platform*

In our analysis, social media data from only one platform was considered, namely Twitter. However, consumers use a variety of social media and other online platforms daily, in order to review different products or services and share their experiences from them (Rousidis et al., 2019, Koukaras et al., 2019, Koukaras and Tjortjis, 2019). That poses an important limitation to our study, as we understand that tweets about Ryanair do not represent the total online user generated content about the brand. Hence, our results might be affected by the selection of Twitter.

- *Aggregated Data*

Both tweets and traffic data were aggregated to monthly data. Therefore, based in our analysis, we could not draw conclusions about patterns in the daily sentiment in tweets, as well as the daily volume of tweets and how these patterns affected the number of Ryanair's passengers.

6.3 Future Improvements

- *Alternative social media platforms and websites*

As mentioned before, there are a lot of different social media platforms, as well as popular websites that passengers use, in order to share their experience from an airline company and rate the service that they have been provided with. Some major airline review websites include Tripadvisor[4], Skytrax[5], AirlineRatings[6] and Trustpilot[7], while Facebook and Instagram are also used daily by millions of passengers. Thus, we understand that there is a wealth of online sources, other than Twitter, that could be used to extract opinions and sentiments towards a brand.

- *Alternative Twitter metrics*

Our study was focused on researching how specific Twitter metrics influence the growth rate of Ryanair passengers. Such Twitter metrics include the volume of tweets, the sentiment that they convey, the volume of positive and negative tweets, as well as the ratio positive/negative tweets. However, we did not take into account alternative metrics, such as the number of Retweets that each tweet had or how influential a person who tweets about the brand is.

[4] https://www.tripadvisor.com/Airlines
[5] https://www.airlinequality.com/
[6] https://www.airlineratings.com/airline-passenger-reviews/?l=R
[7] https://www.trustpilot.com/review/

- *Sentiment Analysis*

In this study, a lexicon-based, as well as a machine learning approach was used, in order to perform sentiment analysis. In both cases, a polarity score between -1 and +1 was assigned to each tweet. However, in the end, each tweet was classified as either positive, negative or neutral. As a future improvement, tweets could be classified according to specific emotions, such as happy, angry, sad and surprise.

6.4 Conclusions

The aim of this dissertation was to investigate the importance of opinion sharing in Twitter, in the business performance of Ryanair airline company. Specifically, 382.425 tweets relevant to Ryanair were collected and classified into three categories, according to the sentiment that they conveyed. The categories were the following; positive (tweets with sentiment score equal or greater than +0.2), negative (tweets with sentiment score equal or smaller than -0.2) and neutral (tweets with sentiment score between -0.2 and +0.2). Sentiment analysis was performed with the use of two separate methods; the VADER lexicon-based sentiment analysis tool and a machine learning classification model, based on Random Forest classifier. The model was trained on a pre-labelled dataset obtained by Kaggle and achieved an accuracy score of 92.5%. A Statistical Analysis followed, in order to investigate the relationship between metrics that represent Twitter and the growth rate of Ryanair's traffic, measured in monthly number of passengers. A simple linear regression analysis showed that all Twitter metrics significantly affect the growth rate of Ryanair's passengers with a time lag of one month. Of all nine independent variables that were examined, four of them have a positive impact on the dependent variable, while the rest have a negative effect. It is interesting to note that according to R-squared goodness-of-fit metric, metrics that have been calculated using VADER lexicon, explain the dependent variable in a remarkably higher degree than metrics measured by the machine learning classification model. In particular, tweets sentiment, as calculated by VADER lexicon achieved the best performance of 32.3%, according to R-squared goodness-of-fit metric. Next, a Vector Autoregressive model was employed, in order to examine the dependencies between all variables that reflect Twitter metrics and Ryanair's growth rate of passengers. According to our results, three important conclusions can be drawn. First, analysis underlines a significant relationship between sentiment in Twitter and business performance of a low-cost airline company. Second, Twitter sentiment affects exogenously the business performance. Last, the response of sales on a shock in Twitter does not last for more than one period. The contribution of this study is twofold. First, this study contributes to the field of

Sentiment Analysis, by suggesting a classification model based on Random Forest classifier, with a significant performance. In addition, this study confirms the importance and the power of online opinion sharing in Twitter, in the business performance of Ryanair. The results highlight the need of social listening, on the part of Ryanair, in order to understand the perception of consumers towards the brand and the way that the company's performance is affected by this perception. Moreover, taking into account the findings of this study, we believe that Ryanair can act proactively upon negative sentiment in Twitter, in order to protect its brand name, as well as its economic prosperity.

7. References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 492-499). IEEE Computer Society.

BBC 2017 https://www.bbc.com/news/business-41414414

BBC 2018 https://www.bbc.com/news/uk-45932027

Beleveslis, D., Tjortjis, C., Psaradelis, D., & Nikoglou, D. (2019, September). A Hybrid Method for Sentiment Analysis of Election Related Tweets. In *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)* (pp. 1-6). IEEE.

Bing, L., Chan, K. C., & Ou, C. (2014, November). Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In *2014 IEEE 11th International Conference on e-Business Engineering* (pp. 232-239). IEEE.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Brandwatch 2014 (https://www.brandwatch.com/blog/travel-agencies-adaptations-and-violins/)

Brooks, C. (2019). *Introductory econometrics for finance*. Cambridge university press.

Cheung, C. M., Lee, M. K., & Rabjohn, N. (2008). The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet research*, *18*(3), 229-247.

CNN 2017 https://money.cnn.com/2017/11/07/technology/twitter-280-character-limit/index.html

Davis, A., & Khazanchi, D. (2008). An empirical study of online word of mouth as a predictor for multi-product category e-commerce sales. *Electronic markets*, *18*(2), 130-141.

Fan, Z. P., Che, Y. J., & Chen, Z. Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. Journal of Business Research, 74, 90-100.

Goldsmith, R. E., & Horowitz, D. (2006). Measuring motivations for online opinion seeking. *Journal of interactive advertising*, *6*(2), 2-14.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.

Hudson, S., & Hudson, L. (2017). *Customer service in tourism and hospitality*. Goodfellow Publishers Ltd.

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, *60*(11), 2169-2188.

Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business horizons*, *54*(2), 105-113.

Kaggle https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

Koukaras, P., Tjortjis, C., & Rousidis, D. (2019). Social Media Types: introducing a data driven taxonomy. *Computing*, 1-46.

Koukaras, P., & Tjortjis, C. (2019). Social media analytics, types and methodology. In *Machine Learning Paradigms* (pp. 401-427). Springer, Cham.

Lassen, N. B., Madsen, R., & Vatrapu, R. (2014, September). Predicting iphone sales from iphone tweets. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference* (pp. 81-90). IEEE.

Laurikkala, J. (2001, July). Improving identification of difficult small classes by balancing class distribution. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 63-66). Springer, Berlin, Heidelberg.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415-463). Springer, Boston, MA.

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

Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering* (Vol. 1, No. 1, pp. 61-67).

Oikonomou, L., & Tjortjis, C. (2018, September). A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter. In *2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM)* (pp. 1-8). IEEE.

Rousidis, D., Koukaras, P., & Tjortjis, C. Social Media Prediction: A Literature Review.

Ryanair 2019 https://investor.ryanair.com/traffic/

Schivinski, B., & Dabrowski, D. (2016). The effect of social media communication on consumer perceptions of brands. *Journal of Marketing Communications*, *22*(2), 189-214.

Sims, C. (1980). Macroeconomics and Reality. *Econometrica,48*(1), 1-48. doi:10.2307/1912017

Statista 2019 https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M., & Zadrozny, W. (2018, July). Causality Analysis of Twitter Sentiments and Stock Market Returns. In *Proceedings of the First Workshop on Economics and Natural Language Processing* (pp. 11-19).

The Guardian 2019 https://www.theguardian.com/business/2019/aug/21/ryanair-makes-high-court-attempt-to-stop-pilot-strikes

Thelwall, M., Buckley, K., Paltoglou, G., 2011. Sentiment in Twitter events. J. Am.Soc. Inf. Sci. Technol. 62 (2), 406–418.

Tzirakis, P., & Tjortjis, C. (2017). T3C: improving a decision tree classification algorithm's interval splits on continuous attributes. *Advances in Data Analysis and Classification*, *11*(2), 353-370.

Wan, Y., & Gao, Q. (2015, November). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 1318-1325). IEEE.

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Computers in Human behavior, 27(2), 634-639.