Department of Economics and Finance *Chair: Economics and Business*

SENTIMENT ANALYSIS APPLIED ON EARNING GUIDANCE: A PRACTICAL APPROACH

Supervisor:

Bozzolan Saverio

Candidate:

Francesco Morabito

188591

Academic year 2017-2018

To my beloved Family.

Index

Introduction

Text mining is a new research area, which has set the goal of obtaining information from digital textual resources.
It is based on statistical, linguistic and machine learning techniques for unstructured data analysis such as text documents. In this paper an overview of the main text mining techniques is presented.
Subsequently, a sentiment analysis is performed using two different ontological vocabularies: the AFINN and the well-known SentiWordNet.

CHAPTER I:

THE WORLD OF BIG DATA

1.1 The Big Data

The term Big Data is part of the multitude of new concepts and meanings, brought by the ever-broader and intense use of internet and mobile applications, which are born and spread with the same rapidity of the means that generate them. Even Big Data, like the terms social network, RSS, Cloud, Geo tagging, are part of a new vocabulary, the result of the possibilities brought by the union of very advanced physical technologies, smartphones, tablets and notebooks, with software capable of make increasingly complex analysis and applications designed to be increasingly useful, social and easy to use.

Big Data refers to large and complex collections of datasets, available to companies, that require specific tools to be able to deal with the acquisition, management, analysis and visualization of the data of which they are composed. This complexity comes from many factors such as the potentially heterogeneous nature of data, from the fact that information can be of an unstructured type such as e-mails, images, comments taken from social networks or GPS data, and the possibility of coming from more sources such as the web, sensors of equipment used in the most varied uses, automatic calculators and many others.

The extraordinary amount of data that is produced every day by people and machines is not enough to describe Big Data: within these huge databases there is a considerable amount of noise, redundancy and meaningless data (*garbage*) which reduces the research potential, therefore it will be necessary to associate the Big Data with cleaning and optimization systems and procedures; secondly, without the capacity for analysis and interpretation, the information potential of a volume of data, even if very high, vanishes, so it will be necessary to develop a physical infrastructure and algorithms for the effective use of the collected data.

To help us to further describe Big Data we can also refer to the most commonly used model in scientific literature (Huang, Lan, Fang, 2015), which starts from "3V" by Doug Laney (2001), i.e. Volume, Velocity and Variety, adding to these a further V, Value, to update it on the basis of the evolution that has occurred in the last fifteen years:

Volume: the volume of data, i.e. the ability to obtain, store and access large volumes of data, so high that they cannot be managed and controlled by traditional databases. With Big Data the volume dimension is not anymore Gigabyte but, Terabyte, Petabyte, Exabyte.

Velocity: i.e. the speed at which data is generated and processed. The analysis of information must be carried out as quickly as possible, if not in real time. The speed must be a strong point in the study of its dataset since

this determines proportionally the value generated by the analysis activity itself, especially in cases where information becomes obsolete very quickly. Consider, for example, how important is the speed of information management in the financial sector and the advantages that this entails for those who obtain and interpret the information coming from the market first.

Variety: the variety of data that can be stored and analyzed. In the past, data analysis could only be used on data structured in tables or relational databases, in which data of different formats such as photos, conversations, audio and video data, documents, etc. could not be entered. With Big Data, now, data from different sources and types can be processed, in particular it is possible to combine structured and unstructured data, like the ones listed above, in the same data set. The latter, in fact, have now reached a decisive importance accounting for 80% of the data produced in the world and being the ones most generated by most of the customers to which the companies that produce consumer products are addressed.

Value: Big Data must be a source of value for those who use them as well as being a tool to perform specific analysis, this also thinking of all those production areas within companies where they can be used, for example to measure performance, to reduce production time or improve marketing or the service offered. In order to have valuable data, the information contained must first of all be truthful, secondly the data must be managed with transparency, so that all the stakeholders can gain an advantage, they must be useful for improving the performance of the offer in general and will have to allow to perform detailed analysis of the product, to be able to improve its characteristics by following or even anticipating the desires coming from market developments.

1.2 The sources of Big Data

The sources from which the data constituting a Big Data can come can be multiple and with different characteristics: they can be internal or external to the company, structured or unstructured, can possess different combinations of speed, variety and volume. In general, it can be said that these sources can be divided into three macro categories:

- **Person-to-person information** i.e. information that originates from human beings, therefore all the information that is usually exchanged through social networks, blogs, windows to leave comments is part of this category;
- **Personal-machine information**, are transactional data, which typically were taken even before the Big Data analysis, such as accounting records, customer databases, etc., and which are now represented by information collected through the fidelity cards, payment devices, smartphones and many others;

- **Machine-machine information**, is information collected by the sensors of some electronic devices and sent to other devices, such as sensors of the weather stations or for example the Rolls Royce turbine engine sensors for Boing 737 produce 20 terabytes of information for each hour of operation that are sent to the servers of the manufacturer that analyzes them to operate the so-called "preventive maintenance" that reduces costs and risks of failure.

1.3 The new big data business models

Big Data is still an object not completely defined as regards their function and application. Until now, the most intense use of Big Data has been that associated with customer-centric strategies, based on internal company data, especially those related to transactions and interactions with customers, while only now are increasing efforts in analyzing data from social networks. A European Commission research entitled "*Business opportunity: Big Data*" (European Commission, 2013), reports data from a 2012 IBM survey that identifies the opportunities Big Data can represent to strengthen or expand the offer of companies and consultants:

- Improvement of e-commerce and web market-place: in particular thanks to the greater possibility of creating personalized services;
- Enhancement of the tools for analysis and monitoring of business activities: therefore, to make its value chain more efficient, for example by optimizing logistics through GPS data.
- Big Data outsourcing, hosting and offered as a service: that is the opportunity, for the time being exploited by the largest IT companies and software developers, to have Big Data as its object of supply and the accessory services to these. Hence companies that are contracted by other companies to develop the IT system, data storage and analysis.
- Big Data consulting: in fact, by increasing the data collected and interpreted, work opportunities and opportunities for expanding the services offered to consulting companies increase.
- Remote monitoring by companies on their products and IT systems: previously we talked about preventive maintenance, but here we can also mention real-time monitoring of light and gas consumption data, etc.
- Improvement of the services offered to citizens by the public sector: in this context, Big Data offers immense opportunities, both in terms of cost reduction, through efficiency gains and improvements in supply, for example through the possibility of monitoring more effectively the development of initiatives undertaken or to better segment the population and the territory to create ad hoc services.

From a business and management point of view, the opportunities offered by the effective exploitation of Big Data are realized through, first of all, different forms of control of the data creation and management process and later, through the adoption of particular business models or "*profit patterns*" to realize profit. Regarding the forms of control, a company can develop its Big Data Business by creating a partnership with another company, can develop contractual relationships with a consulting agency or can develop the whole process within the own corporate structures and functions. Usually the big companies tend to develop the system internally or, in some cases, they resort to third parties and consulting agencies in the initial phase of system development, in order to reduce the costs related to the design risks and to make the faster implementation, and then conclude the relationship once the infrastructure and the Big Data management procedure become fully operational. Also in this case the typical advantages and disadvantages of working in an autonomous or associated way are valid; therefore, developing the new IT system will allow companies to obtain advantages deriving from economies of scale, greater strategic control and greater profit potential, while a partnership will make the development process less risky and faster, as well as bringing the company into contact with the knowledge and data of specialized partners that could make access to the market more effective and profitable.

Some emblematic examples of business generated by the exploitation of Big Data, derived from the same research, we find them within the sectors in which the investments and profits deriving from Big Data are greater, i.e. financial sector, telecommunications and large retail. For example the Australian National Bank transfers, once made anonymous, the data in its possession concerning 21 million electronic transactions to a joint venture created with a data analysis agency that then sells them to third parties. Or, another example, the great Tesco distributor who, with his division specializing in the analysis of Big Data Dunnhumby, manages to manage and study millions of data from the points of sale and individual fidelity cards, which he then sells to the same companies that they form the assortment, such as Unilever, Nestlé, Philip-Morris. In 2012 the Dunnhumby generated £ 53 million in profit for Tesco.

1.4 Limits and criticalities of Big Data

The expansion of Big Data adoption by companies is slowed by some critical issues and problems generated by the qualities that each Big Data should possess and by the frictions that may be caused by the state regulations for the protection of personal data and to respect for privacy. Big Data, in order to be considered "reliable" and therefore to be able to represent a decision-making and analysis tool, must possess some very precise structural characteristics: it must possess complete data, it must be as up-to-date as possible, they it be

coherent with each other and technically complying with the rules on which Big Data is based, must be accurate and must not contain duplications.

Regarding the critical issues related to the ethical and legal aspects of the processing of personal data or data that for their particular characteristics can be considered sensitive, the European authorities have framed the problem through some rules. In fact, in May 2014, the European Court of Justice, following the sentence "Google Spain vs AEPD and Mario Costeja Gonzalez", established that each person has the right to request search engines to remove, that is to deindexes, some results that they concern and do not have a public interest value. This rule does not yet resolve many doubts concerning the possibility of processing personal data, in fact it has not been specified what is meant by "public interest", moreover the standard refers to search engines and not to the plurality of administrators of web domains, this means that a piece of information can only be deindexed, so it is possible to prevent search engine to find that piece of information through out the web but it is not possible to delete it from the web site in which it was stored.

1.5 Towards Text Analysis

The analysis of textual data, within the new discipline of Big Data, is one of the most important horizons, in terms of volume and relevance of the information obtainable, and is, in fact, one of those fields in which researchers and businesses are currently focusing its efforts. This interest stems from the fact that, if we can say that companies are already more or less equipped with systems and methods to analyze non-textual data or, however, can find these tools from the market, the same can not be said for what relates to textual data. Obviously this delay is understandable, they first developed the tools to analyze the data already held by the companies, i.e. the internal data that are in a structured and numerical form, plus the value of textual data has acquired real importance only in the last few years thanks, in particular, the generalized adoption of smartphones and the massive entry of social networks in everyday life. The goal of companies today lies in being able to interpret and extract from this huge amount of data, generated every day, useful information for their activities; think, for example, of the usefulness that could be obtained by companies producing consumer products to understand how their offer is perceived, or the activities that produce services in detecting which parts of the service are most often mentioned and which are not. In general, practically all sectors can benefit from the analysis of text data.

CHAPTER II

SENTIMENT ANALYSIS: AN OVERVIEW

2.1 Introduction

This chapter presents a general description of the techniques and problems related to text analysis and gives some historical notes on the development of research in this field. Subsequently, we proceed with a more in-depth discussion, describing the preprocessing techniques necessary for the construction of a set of quantitative data starting from the written texts.

Text analysis can be defined as the analysis of information contained in documents written through data mining methods. The type of information analysis may vary, ranging from systems that automatically classify documents based on content, to search for documents containing specific information, up to applications that find a precise information requested by a user, allowing to catalog it and describe it in a structured form. The areas of study related to text analysis cover various problems, including statistical methods in use, an algorithmic part and an important part of linguistics.

The key issues that the subject of text analysis research is:

- ✓ the construction of a data set starting from unstructured information present in the written text;
- ✓ supervised classification of documents based on content;
- ✓ searching for information based on specific requests from a user;
- ✓ unsupervised classification and organization of documents;
- ✓ automatic extraction of information;
- ✓ the choice of algorithms and their evaluation.

2.2 The historical background

Before starting, it is interesting to present a brief reference to the history of Text Mining, by retracing the fundamental stages of research. Historical background Studies based on text mining began in the 1960s, and then slowly developed and exploded in recent times, thanks to the enormous availability of texts in electronic format available in archives and on the web. The fundamental stages of research in this area can be summarized as follows:

1960s: The first studies of the 1960s focused on indexing and searching for documents. Problems similar to classification are treated, which, given a set of categories, provide for automatically assigning a category to a given document.

1970s: In the seventies, the idea of a mathematical description of written texts began to be disseminated. This description allows to perform the first studies on cluster analysis applied to documents. Therefore, different definitions of similarity between written texts are affirmed, which allow the implementation of the main data mining algorithms.

1980s - 1990s: Subsequently, the dissemination and popularity of studies in the field of artificial intelligence also involves text mining and, more generally, the linguistic field. Multiple algorithms typically treated by artificial intelligence texts are applied, for example, to the assignment of documents to categories, but also to deal with problems related to linguistics such as extraction of prefixes or logical analysis or of the period. Computational linguistics also sees a great development, and consequently the potential of text mining increases.

1990s – today: The last few years have seen the multiplication of studies and applications of text mining. The dissemination of internet and information technology has made available a large amount of texts in electronic format. In fact, one of the main problems in the '60s and '70s was the lack of availability of data sets, linked to the difficulties in digital coding of texts. It is therefore necessary to have valid tools to deal with the information available in the texts, making them effectively exploitable. Hence the need to develop effective search engines, filters to highlight relevant documents, systems to trace the main information present in a document and applications to support customer services. As often happens, the initial push for research in the sector is mainly due to public funding, often linked to military projects. In this way numerous conferences have been called, each addressing a specific theme. The most relevant are the MUC (Message Under standing Conference), the TREC (Text Retrieval Conference), the CoNLL (Conference on Natural Language Learning) and the ACE (Automatic Context Extraction). Each conference poses a specific problem, which the participating researchers face independently, trying to propose effective and innovative resolution techniques.

2.3 Data preparation

Each data mining analysis requires a data preparation phase. This phase is even more important in the case of text mining. Not only is it necessary to have documents in a standard format, but a more detailed preliminary treatment of the texts will lead to greater efficiency of the algorithms used and will make possible more in-

depth analyzes. In this section, therefore, we see the main phases that characterize the preparation and construction of the data set, intended as a collection of documents.

Depending on the software used, there are several accepted formats for input documents. In general, a text without formatting (.txt) is always usable. However, often the information contained in the formatting and in the structure of the text could be important. For example, the string of characters that make up the title will be very important to identify the topic covered by the text, as well as the subject field in an e-mail. Words highlighted in bold may indicate key concepts, as often information about the date and the author may be useful. Among the various formats available, the use of XML as a standard for text mining has spread. The XML is based on the division of the text into fields, which can give indications about the content (for example the title fields, author, summary, etc.) or the format of the font (style, size, etc.). Most of the choice is due to the fact that almost all editors allow the export or automatic conversion of the text into XML, and the same applies to the export of e-mails (think of Thunderbird or Claw-mail) or of web pages. Moreover, unlike a text-only format, XML allows to preserve the structure of the text, removing the information related to the formatting, which is often useless for the purposes of a text mining analysis. But the use of XML goes beyond the simple input format. Usually, Natural Language Processing software provides output information by adding XML tags in the text. For example, an application for logical analysis can indicate subject and object complement through appropriate tags.

2.3 Processing and stemming

The language can be considered and evaluated in a very effective way to interpret and decode in all its complexity, since the text is analyzed from the point of view of sentiment is required some basic steps of pre-processing in order to reduce the number of terms within the vocabulary that will be taken into use.

So, first of all, in the process of analyzing a text, we must try to reduce the text into a quantitative datum, so it can be treated by a statistical model. The preliminary phase of the process will filter the texts that this has been done for each text deals with the study, which is in such a way as to be a set of texts, corpus, useful and easily interpretable by the algorithm and to avoid problems of overfitting, typical of models that adapt to the observed data using too many parameters.

This phase is essential and necessary for all the analysis processes. Next phase to this is that of the "Stock Exchange" or the "Processing", now used in almost all the studies, which provides to dissolve the linguistic ties of the text, i.e. to mitigate the information related to the order with which words follow each other, turning

the corpus into a deconstructed set of unigrams. This phase may seem simpler, but in reality, a simple list of words, not ordered, is enough to identify a general sense.

You translate phrases or texts into "lots of words", a lot of untidy words, formed by unigrams, in their majority, but also by bigrams or trigrams, as if to analyze for analysis not too thorough, such as measuring feelings, an in-depth specification of these does not bring great increases in the quality of the analysis. These last ones are pairs or treble of words that acquire a sense in themselves and in particular, with reference to the analyzes that are sought; for example, the words "five-star-movement", "White-House" and many others assume a specific value precisely because they stand side by side with each other and take on a different meaning if read differently.

In the pre-processing phase of the text, tokenization must also be carried out; with this we mean the process of subdividing the characters into minimal units of analysis by an automatic analyzer, these minimum units are called tokens. Usually, by splitting the texts into tokens, we refer to the spaces that separate one word from the other and is not complicated for all those languages that use space as a natural separator. In each there are some general criteria to be respected during the tokenization phases of any text document:

- Heuristic criteria to decide whether dates, addresses, links and decimal separators constitute a single token;
- Choose how to interpret words with accents, elisions or cripples;
- Consider the effects of punctuation on the use of spaces, especially for the use of brackets and quotation marks;
- Consider the presence of differences, even Marquis, between words with which begin with a capital letter or not (for example, Trapani and drills) A further step, besides the tokenization, is the stemming;

Once we have "broken" the links between the words of a text, we can simplify the same words, then the vocabulary that constitutes the phrase, through the "stemming" technique. This aims to reduce the single words to the roots of which they are composed by removing the final part, in such a way as to reduce the number of terms that will result in the final data set. Therefore, the stemming serves to simplify the identification of words that refer to the same basic concept by reducing them and making them correspond to their common root; for example the words "family", "families", and "familiar" that all derive from the common root *famil*, which becomes the style. Or Martin Porter, the author of the famous Porter-stemmer, in one of his articles that most influenced and still influences this type of research, makes the example of the "connect" style, valid for grouping the words "connected", "connecting", "connection", " connections ". So the stemming is an approximation, based on the needs of the researcher, of what in linguistics is called lemmatization. In reality,

between a stem and a lemma there is an important difference that is the fact that a stylema may not always be the same for a given group of words, but may depend on the characteristics of the research, while a lemma is always and only one given a set of words, because it corresponds to the true morphological root of these.

There are several algorithms of stemming, more or less aggressive in cutting words, but one of the most widely used is the algorithm of Martin Porter, which is usually used because it adopts a moderate approach to the simplification of words. Or the Snowball algorithm, which is currently one of the most used, and which can boast of supporting 15 different languages.

Another phase of preprocessing is the cleaning and homogenization of the text. Homogenization occurs by modifying special characters present in the text as uppercase letters, while with the cleaning it is possible to modify or eliminate punctuation; words that are functional only for grammatical and syntactic purposes, such as articles, suffixes, prefixes, propositions etc.; words too common and too uncommon within the analyzed corpus, thus eliminating the words that appear in 99% of the texts or only in 1% (these thresholds are at the discretion of the researcher, Iacus, Curini and Ceron talk about 90% and below 5%) .

There are numerous other difficulties in the application of Sentiment Analysis related above all to the distortions of the human language itself, as for example, the expansion of the letters (write "Helooo" instead of "Hello"), the typo of the author, who many software can solve through self-correction systems; the use of word separators can confuse programs in the interpretation of the meaning of the presence or absence of a stem. Even the double negatives can create difficulties, in fact currently few systems can detect them. These are the grammatical difficulty, so the texts cannot be extremely complex from a syntactic point of view; the presence of photos or links in the text, the presence of self-generated spam messages; as well as emoticons, irony and sarcasm.

All these problems find solution in human intervention during the process of machine interpretation in such a way as to decode all the possible nuances contained in the texts, although this may not always be possible since, sometimes, even for the operator can be difficult to grasp the meaning of texts or particularities of the highly ambiguous language.

Preprocessing and stemming allow you to create a matrix having lines for each text, for columns each style identified and, in each row-column intersection a 1 or a 0 according to whether the style is present or not. In some cases, it is used and used to enter the number of times that the style appears, but it has been seen that this different solution does not bring any particular improvements to statistical reliability or more information. From this table we can obtain, therefore, the representation of each document i $(i = 1, ..., N)$ as a vector $Si =$

(s1, s2, s3, ..., sn), indicating the presence or absence, within the document, of a word deriving from the relative style.

This matrix, called the steam matrix, is the starting point of every text analysis model; usually this kind of tables has at most 400-500 styles, but in general much less, and can also contain hundreds of thousands if not even millions of texts and then rows. For each steam table one can calculate the density, Sparsity index, that is the percentage ratio of the presence of zeros with respect to the one inside the same.

2.4 Alternative processing methods

The path just described is not the only solution to transpose a text into a set of numerical data. In fact, depending on the needs of the research can be eliminated different elements and to a more or less aggressive. An interesting example of different preprocessing of the text comes from one of the first quantitative analysis searches of the text. This, carried out by Mosteller and Wallace (1963), had as its objective to identify who among Alexander Hamilton, James Madison and John Jay was the author of some articles belonging to the Federal Paper, and of which most of these were known by whom of the respective three politicians had been written. So the two researchers, being more interested in the form and style of the documents than the concepts expressed, have made reference and examined the words useful to the syntactic aspect such as prefixes, suffixes, articles, propositions and anything else that normally they would be eliminated, excluding the words useful to make sense, in terms of meaning, to the text.

Other differences could be given by the count of how many times the single stylus appears in the document, as already mentioned, or by understanding the possessive pronouns in the analysis, or pondering the importance of the words based on their rarity in the text.

Sentiment classification systems are usually formulated in such a way as to understand whether the opinion, review or evaluation we are studying is positive or negative and, often, this becomes particularly easy in the case of evaluations present on the web when they are accompanied by a scale of values in which the user must indicate his opinion numerically, usually formulated through a scale from 1 to 5. In cases where you do not want to insert a scale that goes from "very negative" to "very positive", there is a problem of classification of the text, that is we must understand if the words that make up the sentence we are reading can make us consider the same as corresponding to positive or negative feelings.

This case, i.e. being able to assign each document to a category of opinions, corresponds to most of the research in which methods of content analysis are applied, in particular in fields such as marketing and politics. This type of investigation aims to be able to deduce which category the single document belongs to, how the texts are distributed in their entirety or both of this information.

The most effective method for achieving these objectives is to proceed with a human-based analysis, in which, after having identified the classification rules for the texts examined, a trained operator specifically reads each document and assigns it to the appropriate category. Obviously, this solution is the most expensive in absolute terms, both for economic and time expenditure, which is why automated methods of analysis are often applied, able to reduce costs by reducing human intervention in the analysis process. Moreover, an analysis of human origin can also present the risk of reducing objectivity, especially if we are talking about ethics or politics, so precautions must be taken such as the rotation of two or more operators in the analysis process, the Cross-coding of texts and sample checking of coded texts by an external supervisor.

The automated methods can be divided first of all based on the presence or absence of predetermined categories in which to insert the observations; in the first case the methods can be of the type Dictionary methods or Supervised Learning methods, in the second case they will be of the Clustering type.

The first of the two is composed of methods that use the calculation of the frequency with which appear the "keywords", previously determined, to identify which opinion category the document belongs to. The second group, on the other hand, is made up of methods that replicate manual decoding of the text using a calculator. For these it will be necessary, first of all, for an operator to analyze a subset of the totality of the documents, subset, to form a predetermined classification scheme of the texts, training set; subsequently, the latter will be used to program the computer that will perform the analysis of the remaining documents.

2.5 Clustering techniques

Among the unsupervised text classification techniques, we find the most classical techniques deriving from data analysis, such as that of the Cluster Analysis. This, like the other text mining tools, aims to identify regularities in the data, and therefore regularities within the texts.

The Cluster Analysis, in particular, is based on the possibility of giving a value to dissimilarity by measuring the distance between objects that you want to classify, or rather to divide into groups the most homogeneous as possible according to this predefined distance.

A measure of dissimilarity d between two objects *A and B, i.e. d (A, B),* is a value number 0 when it is calculated for the same element, i.e. $d(A, A) = 0$, it is always non-negative, i.e. $d(A, B) \geq 0$, and is also

symmetrical, i.e. *d (A, B) = d (B, A).* If the triangular inequality is also satisfied, ie d (A, C) ≤ d (A, B) + d (B, C), then dissimilarity can be called distance.

If there is such a measure of dissimilarity *d*, the clustering algorithms will be able to proceed in an agglomerative or dissociative manner, that is, starting respectively to aggregate in an ordered way the data most similar to them or taking into consideration the whole data group and separating the farthest from one another.

Once the distances between data have been calculated, there are many ways to tell if these are near or far between. For example, by calculating the minimum dissimilarity between an element and all the groups of the whole, or the distance of the element from the center of gravity of the whole. These different possibilities of distance calculation generate a plurality of different results, depending on whether the distance has been calculated in one way or another. In this situation the best solution is to compare all the different clustering solutions and to adopt the one that provides the most robust results.

In any case, once the Clusters are generated, it is necessary to look inside each of them to find out in what some groups are similar semantically inside them and for what reasons they differ from the others.

Cluster Analysis can also be partly supervised, for example by specifying the number of groups to be formed, or by forcibly shifting the elements of one group into another in order to force the algorithm to learn a better classification. Alternatively, or in parallel, the texts to classify also texts for which classification is known in order to help the clustering technique to discriminate between the various texts of the corpus can be introduced among the texts.

Dictionary methods: The dictionary methods are the most intuitive and the easiest to apply among all automated methods. These use the degree with which "keywords" appear within a single text in order to assign it to a category or to be able to assess to what extent this belongs to one category of opinion rather than another. In this case, therefore, we are in the presence of automatic tagging, i.e. a calculator, through an ontological dictionary and a set of rules provided a priori, which associates each text with a semantic content rather than another. This method assumes greater utility and reliability when the theme of which the texts speak is well circumscribed and when the set of the steam and of keywords is very small (for example in the parliamentary speeches). In addition, the dictionary allows you to perform highly automated, repeatable and applicable operations to all the texts of a corpus of any length. You can also use these methods to understand the "tone" of newspaper articles, but they could be applied to an infinite number of other fields, including the interpretation of web comments. In order to measure the tone of the articles of a newspaper, a dictionary of keywords must be created, to which a score, a score or weight is assigned that gives a positive or negative value, and then the frequency must be calculated. Sometimes these words appear and which ones. Therefore, for each word m (m = 1, ..., M) a weight of s_m will be associated; for example, $s_m = -1$ if the word is

assigned negative value, or $s_m = 1$ if positive value is assigned. If $N_i = \sum_{m=1}^{M} W_{im}$ words are used in a document *i*, then the dictionary method can measure the relative tone for each document:

$$t_i = \sum_{m=1}^{M} \frac{s_m W_{im}}{N_i}$$

Approximately continuous measurements can also be used to determine the text tone, but the simplest method is to assign texts with $t_i > 0$ to a positive tone category and insert texts with $t_i < 0$ within the negative tone group.

The dictionary method, to work well, must be accompanied by a score as close as possible to the meaning that the words used assume in the particular context. The scores must be developed tailored to each specific new analysis that you want to do. In case you use a dictionary with relative scores for a research for which it has not been specially designed, there is a risk of producing errors. Other errors in the analysis of texts by dictionary can occur if they are to be applied to languages other than those for which they were formulated, or for corpus composed of texts in different languages, as can be the comments on the websites, or if there are metaphors, word games or other aspects of the language in the texts that cannot be integrated into an ontological dictionary. So, you have to pay a lot of attention in the use of dictionaries and you must always evaluate the results they provide, whether pre-formulated dictionaries are used or that specially created dictionaries are used, because even the latter can produce unexpected results and not completely correct.

Supervised Learning Methods: With the Dictionary Methods, researchers must identify words and terms that define distinct classes as the first step in the process. This can cause inefficiencies in the application of methods to the interpretation of data of real origin, and in particular if they are used in contexts for which they have not been specifically designed.

The Supervised Learning Methods can constitute an alternative tool for the assignment of the various documents to the categories previously identified by the operator. The supervised learning techniques are part of the machine learning techniques and aim to educate an information system in such a way as to allow it to solve different types of tasks autonomously, starting from a base of ideal examples that are provided during the training phase. These are based, therefore, on the idea of simulating the work done manually by a human operator thanks to the use of an algorithm and, therefore, of a calculator. The algorithm will be based on the training set, so as to "learn" the calculator how to separate the documents of a corpus into categories or how to insert them in the most suitable predetermined categories. We can say that all the supervised learning algorithms start from the assumption that, if we provide the system with a sufficient number of examples, this

will accumulate an experience enough to allow it to create a function *f (ha)*, i.e. the inductive hypothesis that at each input data associates a hypothetical correct answer, suitable to approximate the function *f (hb)*, the objective function that associates the correct response desired by the user to each input data. If we insert data not present in experience E in the system, the function *f (ha)* should sufficiently approximate the function *f (hb)* so as to obtain sufficiently satisfactory answers for the user.

Taking advantage of these methods we can obtain two advantages in particular: first of all, it is necessary, for each research that is carried out, to create a specific algorithm specifically designed on the topic that you want to study and format on the training set related to the corpus in question. Researchers will have to develop a specific system of words and meanings specially formulated for the study. So, this need to customize tools to the research area, obliges to avoid the problem of applying an instrument to an inappropriate field, or not entirely appropriate, as could happen with dictionary methods. Secondly, these methods are much easier to validate since they themselves can provide clear statistics on the performance of the analysis, so as to be able to immediately realize if the results found have a certain significance or not.

The use of supervised methods for the classification of texts is a growing trend, rapidly expanding and evolving, but all supervised learning methods have three common points:

- Training set: the construction of a set is always necessary of texts or documents on which to program and test the model that will then be used for the actual analysis.
- Learning: phase in which the designed model is updated or "learn" the relationships between the characteristics of the individual documents and their categories of membership that make up the training set; the information obtained from this passage will be necessary to place the documents that make up the corpus taken into consideration in the most suitable categories.
- Validation: the final phase that all the supervised learning methods have in common is the validation of the final results, which also includes the placement of the documents for which an adequate placement has not been found.

CHAPTER III:

STUDY CASE: SENTIMENT ANLISYS APPLIED TO EARNIG GUIDANCE

3.1 Introduction

In the previous chapters it was seen what the methodologies of approach to sentiment analysis are according to the main schools of thought. In this chapter we will focus more concretely on the practical application of sentimental analysis to earnings guidance txt files. Python was chosen as the development platform for the code, both for the large number of packages available and for its extreme ease of use. The script was compiled with Geany.

3.2 Getting and fromatting data

Data on earning guidance are obtained from *Factivia.* The raw text is extracted from “Press Release Newswire” and “Dow-Jones Business News” from the North-America region.

Using a python script that arrange one earning guidance per row I created a .cvs file that we will use as data frame for the sentiment analysis’s script.

```
review,sentiment
"DETROIT (AP)--Kmart Holding Corp. (KMRT) said Monday that it had strong profitability and cash generation in November and December as the discount retail
"Kmart expects net income, excluding any asset sales and bankruptcy-related expenses, to be about $two hundred and fifty million for the two-month period.
"OKLAHOMA CITY (Dow Jones)--Sonic Corp.s' (SONC) first-quarter earnings grew twenty-five%, driven by a sharp jump in same-store sales that reflects increa
"The franchiser and operator of drive-in restaurants also projected second-quarter earnings and revenue to be slightly below and within Wall Street's esti
"ATLANTA (Dow Jones)--Oxford Industries Inc.'s (OXM) fiscal second-quarter net income rose thirty-three% due to a double-digit increase in sales and an im
"The private-label clothing maker also expects third-quarter and fourth-quarter earnings per share to both increase from year-earlier totals.",positive
"NEW YORK (Dow Jones)--Tarragon Corp. (TARR) backed its fourth-quarter earnings guidance and issued an outlook for two thousand and five that projects ear
"In a press release Wednesday, Tarragon, which builds residential homes in high-density urban locations, reiterated its two thousand and four earnings tar
"CHICAGO (Dow Jones)--Medical device maker Medtronic Inc. (MDT) said Wednesday it sees fiscal year two thousand and five sales rising at the low end of a
"During a mid-quarter conference call, the company also said it expects fiscal-year per-share earnings growth in the fourteen% to sixteen% range, and thir
"CAMDEN, N.J. (Dow Jones)--Campbell Soup Co. (CPB) will increase list prices of selected soups and other domestic retail products to partially offset cost
"The food company also backed its fiscal two thousand and five earnings guidance of a five% to seven% increase from a year ago, excluding restructuring ch
"SANTA BARBARA, Calif., Jan. five /PRNewswire-FirstCall/ -- Superconductor Technologies Inc. the global leader in high- temperature superconducting (HTS)
"STI expects net revenues for the fourth quarter ended December thirty-one, two thousand and four to be approximately $four.one million, compared to $seve
"GAITHERSBURG Md. (Dow Jones)--Privately held Panacos Pharmaceuticals Inc., which is in the process of merging with V.I. Technologies Inc. (VITX), receive
"Fast-track status is designed to expedite development and approval of new drugs that may have the potential to improve treatment for serious or life-thre
"QUINCY, Mass. (Dow Jones)--J. Jill Group (JILL) lowered its fourth-quarter earnings guidance and now expects to be below the year-ago result, as a projec
"PLEASANTON, Calif. (Dow Jones)--Ross Stores Inc.'s (ROST) December same-store sales rose two%, bucking an expected decline, due to strong sales of access
"NEW YORK (Dow Jones)--Holiday sales were ho-hum for most retailers, as price-conscious shoppers put off purchases until the last minute and scoured store
"NEW YORK (Dow Jones)--DoubleClick Inc. (DCLK) boosted its revenue and earnings guidance for the fourth quarter and full year to reflect better-than-expec
"FAIRPORT, N.Y. (Dow Jones)--Constellation Brands Inc.'s (STZ) third-quarter net income rose seventeen%, helped by growth across the company's branded win
"DALLAS, Jan. six /PRNewswire-FirstCall/ -- Tuesday Morning Corporation today reported that net sales for its fourth quarter ended December thirty-one, tw
"SAN DIEGO, Jan. six /PRNewswire-FirstCall/ -- WD-forty Company today reported net sales for the first quarter ended November thirty, two thousand and fou
"DENVER, Jan. six /PRNewswire-FirstCall/ -- M.D.C. Holdings, Inc. today announced that it received orders, net of cancellations, for two,six hundred and s
"INDIANAPOLIS (Dow Jones)--WellPoint Inc. (WLP) adjusted its guidance lower for the fourth-quarter to account for the successful repurchase of surplus not
"NORCROSS, Ga. (Dow Jones)--Immucor Inc.'s (BLUD) fiscal second-quarter earnings rose twenty-two%, lifted by price increases and robust sales of its Galil
"ATLANTA (Dow Jones)--Video Display Corp. (VIDE) put fiscal third-quarter earnings at more than double a year earlier as revenue rose more than fifteen%.A
"NEW YORK (Dow Jones)--IVillage Inc. (IVIL) bought Healthology Inc., a producer of physician-generated health and medical information on the Internet, for
"GOLETA, Calif. (Dow Jones)--Deckers Outdoor Corp. (DECK) boosted its fourth-quarter earnings and sales projection, citing demand for its Ugg brand, an im
```

Every data frame is composed by the text extracted from the *Factiva* database and the corresponding opinion about that data input that will be used as validation after that the sentiment analysis is performed.

3.3 Text normalization

Subsequentially to the creation of the data frame we will normalizing and standardizing our text data using some built-in functionalities in the NLTK library (Natural Language Toolkit) in order to expanding contractions, removing unnecessary HTML characters, tokenization, removing stopwords, special characters, and lemmatization.

3.4 Processing

Now the Text data is ready to be processed. In order to evaluate the sentiment of every data entry we will use two different ontological dictionaries:

- AFINN: is a manually labeled by Finn Årup Nielsen in 2009–2011 list of English words rated for valence with an integer between minus five (negative) and plus five (positive);
- SentiWordNet: is a lexical resource for opinion mining that assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity. It has a Web-based graphical user interface, and it is freely available for research purposes. The development of the resource is based on the quantitative analysis of the glosses associated to synsets, and on the use of the resulting vectoral term representations for semi-supervised synset classification. Positivity, negativity, and objectivity are derived by combining the results produced by a committee of eight ternary classifiers.

Using the function afn.score it will be possible to evaluate the AFINN score for every entry in our data frame; in order do to so the following for loop is created:

```
## Predict sentiment for test dataset
print ("Predict sentiment for test dataset")

sentiment_polarity = [afn.score(review) for review in test_reviews]
predicted_sentiments = ['positive' if score >= 1.0 else 'negative' for score in sentiment_polarity]
```

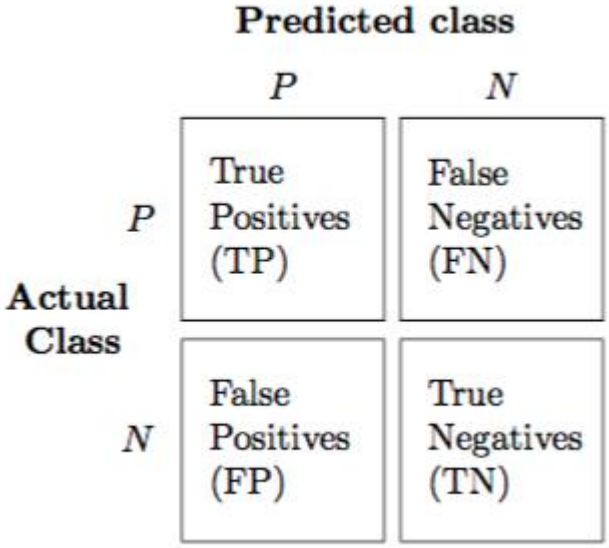
While instead using the custom function analyze_sentiment_sentiwordnet we will get the SentiWordNet evaluation for every entity in the data frame.

```
def analyze_sentiment_sentiwordnet(review,  verbose=False):

    # tokenize and POS tag text tokens
    taggedtext = [(token.text, token.tag_) for token in tn.nlp(review)]
    score = scoreneg = count = obj_score = 0
    # get wordnet synsets based on POS tags
    # get sentiment scores if synsets are found
    for word, tag in taggedtext:
        ss_set = None
        if 'NN' in tag and list(swn.senti_synsets(word, 'n')):
            ss_set = list(swn.senti_synsets(word, 'n'))[0]
        elif 'VB' in tag and list(swn.senti_synsets(word, 'v')):
            ss_set = list(swn.senti_synsets(word, 'v'))[0]
        elif 'JJ' in tag and list(swn.senti_synsets(word, 'a')):
            ss_set = list(swn.senti_synsets(word, 'a'))[0]
        elif 'RB' in tag and list(swn.senti_synsets(word, 'r')):
            ss_set = list(swn.senti_synsets(word, 'r'))[0]
        # if senti-synset is found
        if ss_set:
            # add scores for all found synsets
            pos_score += ss_set.pos_score()
            neg_score += ss_set.neg_score()
            obj_score += ss_set.obj_score()
            token_count += 1

    # aggregate final scores
    fscore = posscore - negscore
    normfinalscore = round(float(fscore) / token_count, 2)
    FS = 'positive' if normfinalscore >= 0 else 'negative'
    if verbose:
        norm_obj_score = round(float(obj_score) / token_count, 2)
        norm_pos_score = round(float(pos_score) / token_count, 2)
```

```
norm_neg_score = round(float(neg_score) / token_count, 2)
    # to display results in a nice table
    sentiment_frame = pd.DataFrame([[final_sentiment, norm_obj_score, norm_pos_score,
                                     norm_neg_score, norm_final_score]],
                                   columns=pd.MultiIndex(levels=[['SENTIMENT STATS:'],
                                                                 ['Predicted Sentiment', 'Objectivity',
                                                                  'Positive', 'Negative', 'Overall']],
                                                         labels=[[0,0,0,0,0],[0,1,2,3,4]]))
    print(sentiment_frame)

  return final_sentiment
```

3.5 Model Performance Evaluation

We will be evaluating our models based on precision, recall, accuracy, and F1-score. Additionally, we will be looking at the confusion matrix and detailed classification reports for each class, that is, the positive and negative classes to evaluate model performance. We can define the following performance measures as:



Standard confusion matrix

- **Confusion matrix**: “confusion matrix, that is known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised and unsupervised machine learning. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions.”
- **Accuracy: “**Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.”
- **Precision:”** Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.”
- **Recall: “**Recall is the ratio of correctly predicted positive observations to the all observations in actual class.”
- **F1-Score**: “F1-Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if the data have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.”

We have taken a total of 600 reviews out of the 1980 to be our training dataset and we will evaluate our models and test them on the remaining reviews. This is in line with a typical 70:30 separation used for training and testing dataset building.

3.6 Model Results

```
Evaluate model performance
Model Performance metrics:
------------------------------
Accuracy: 0.7227
Precision: 0.7385
Recall: 0.7227
F1 Score: 0.718

Model Classification report:
------------------------------
              precision    recall  f1-score   support

    positive       0.68      0.85      0.75       991
    negative       0.80      0.59      0.68       989

   micro avg       0.72      0.72      0.72      1980
   macro avg       0.74      0.72      0.72      1980
weighted avg       0.74      0.72      0.72      1980


Prediction Confusion Matrix:
------------------------------
                Predicted:
                  positive negative
Actual: positive       844      147
        negative       402      587
```

AFINN model results

```
Model Performance metrics:
------------------------------
Accuracy: 0.6833
Precision: 0.6871
Recall: 0.6833
F1 Score: 0.6817

Model Classification report:
------------------------------
              precision    recall  f1-score   support

    positive       0.66      0.75      0.70       991
    negative       0.71      0.61      0.66       989

   micro avg       0.68      0.68      0.68      1980
   macro avg       0.69      0.68      0.68      1980
weighted avg       0.69      0.68      0.68      1980


Prediction Confusion Matrix:
------------------------------
                Predicted:
                  positive negative
Actual: positive       748      243
        negative       384      605
```

SentiWordNet model results

According to the results of the two models, the model that can better classify the text extracted from the earning guidance files is the model that makes use of the AFINN ontological dictionary. The AFINN model is able to classify 72% of the earning guidance correctly (recall). Moreover, of all the elements indicated by the system, 73% corresponds to a correct classification (precision). The model that uses the SentiWordNet dictionary showed a total accuracy of 4 percentage points lower than the AFINN model. This discrepancy is probably due to the different construction of the internal vocabulary and the fact that this vocabulary was mainly developed to classify textual data more similar to reviews and opinions than to sentences extrapolated from economic-financial contexts. In fact, by observing this type of document we can notice that aspects or words that usually have a negative connotation can, in some cases, take on a positive meaning; therefore words like "cancer", "cost", "crude" or "taxes" can correspond to positive effects if referring to hospitals, crude oil and gas trading agencies or to some public bodies. Obviously, as negative positive words can be confused, the inverse can also occur, that is, that words are not recognized as negative and that they are assigned to this positive or neutral character. This is one of the biggest critiques make by data scientists to the use of ontological vocabularies like the one used in the project.

Conclusion

The Sentiment Analysis aims to decipher, in a statistical way, the feelings, opinions and tones of the comments that refer to a company, a product, an event and much more in order to produce mathematical and objective parameters, not influenced by the human factor, to evaluate their performance in terms of happiness, positivity or, more simply, correspondence with consumer expectations. In the pages that make up this work we began with the description of Big Data, ie the environment in which the theory on Sentiment Analysis is placed, then move on to the vision of some methods and classification models that represent the conceptual basis of the functioning mechanisms. analysis of texts. Finally, we carried out a sentiment analysis on 1980 texts extracted from various earning guidance estimates comparing the results of two of the most famous ontological dictionaries available open source: AFINN and SentiWordNet.

Bibliography

https://www.nltk.org/

https://www.python.org/

https://scikit-learn.org/stable/

- Agarwal, B. e Mittal, N. Prominent Feature Extraction for Sentiment Analysis. Springer, 2016.

- Benedetto, F. e Tedeschi, A. «Big Data Sentiment Analysis for Brand Monitoring in Social Media Streams by Cloud Computing.» In Sentiment Analysis and Ontology Engineering, di Shyi-Ming, C. e Witold, P. 341-377. Springer, 2016.

- Bonzanini, M. «Stemming, Lemmatisation and POS-tagging with Python and NLTK.» Marco Bonzanini. 26 Gennaio 2015. https://marcobonzanini.com/2015/01/26/stemminglemmatisation-and-pos-tagging-with-python-and-nltk/.

- Brownlee, J. «Machine Learning Mastery.» Supervised and Unsupervised Machine Learning Algorithms. 16 Marzo 2016. http://machinelearningmastery.com/supervised-andunsupervised-machine-learning-algorithms/.

- Ceron, A., Curini, L. e Iacus, S.M. Social Media e Sentiment Analysis, L'evoluzione dei fenomeni sociali attraverso la rete. Springer, 2014. - D'Andrea, A., Ferri, F., Grifoni, P. e Guzzo, T. Approaches, Tools and Applications for Sentiment Analysis Implementation In International Journal of Computer Applications (0975 – 8887) Volume 125 – No.3, September 2015, 26 -33. 2015

- Esuli, A. e Sebastiani, F. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC 2006, 417-422. 2006 - «Facebook ha fatto parlare tra loro due bot, e questi hanno parlato una nuova lingua.» www.ilpost.it. 1 Agosto 2017. http://www.ilpost.it/2017/08/01/intelligenza-artificialeinventare-nuovi-linguaggi/.

- Haddi, E., Liu, X. e Shi, Y. «The Role of Text Pre-processing in Sentiment Analysis.» Sience Direct. 2013. http://www.sciencedirect.com/science/article/pii/S1877050913001385.

- Internet World Stats. 13 Luglio 2017. http://www.internetworldstats.com/stats.htm.

- Liu, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

- Medhat, W., Hassan, A. e Korashy, H. «Sentiment analysis algorithms and applications: A survey.» Sience Direct. Dicembre 2014. http://www.sciencedirect.com/science/article/pii/S2090447914000550