LUND UNIVERSITY
School of Economics and Management

Master programme in Economic History

Detecting Important Swedish Innovations By Text Mining Articles From the 1970s

Mathias Johansson
Mathias.Johansson.743@student.lu.se

Abstract: This study explores the possibility of detecting otherwise overlooked important innovations by analysis of journalistic texts through naïve text mining techniques. The over 1,000 articles serving as the source data come from the SWINNO, a database constructed through the Literature Based Innovation Output method, where they serve as sources for information on innovations. The contents of the texts were processed through three distinct algorithms. The outputs were compared and cross-referenced with externally assessed major Swedish innovations in order to construct a simple binary classifier. The resulting classifier uncovered 220 previously overlooked important innovations, while including 70% of the references from the external sources.

Key words: Detecting innovation, Text-mining, LBIO

EKHM52
Master thesis, second year (15 credits ECTS)
August 2016
Supervisor: Josef Taalbi
Examiner: Faustine Perrin
Word Count: 14,000

Acknowlegmenets

I would like to thank everyone at the SWINNO project for the opportunity it has offered me over the past 17 months to work in the *capta* generating process. Would I have not been involved in the project I would, probably, never have gotten the idea for this paper and even if I somehow did I would not have had the unfettered access to the source data, and its keeper, that I now enjoyed.

Furthermore, I would like to thank my supervisor who's genuine interest and enthusiasm for my research topic has been a great pillar of support.

Thank you, Anjin-san for teaching me the secrets of the pickle.

Finally, I would like to thank my wife whose incessant support and encouragement enabled me in this egotistical pursuit of mine.

Table of Contents

List of figures

List of Tables

Glossary – Terms and Definitions

Bag-of-words	-	A model that disregards the order of words, maintaining the multiplicity.
Centroid	-	The mathematical center of a cluster
Cluster	-	A set of obervations grouped on mathematical similarity
Clustering Algorithm	-	An algorithm that clusters observations based on their mathematical similarities
Corpus	-	A collection of documents or texts
Dictionary	-	A set of unique words
Hierarchical clustering	-	An agglomerative clustering algorithm
k-means clustering	-	A clustering algorithm
LBIO	-	Literature Based Innovation Output, a bibliometric approach towards measuring innovation output
Lemmatization	-	The process of reducing words to their stems
Semi-supervised	-	A data-mining problem with a set of known outcomes that only covers a set of the sample
Sentiment analysis	-	An approach towards detecting the overal sentiment of a text.
Stem	-	A reduced form of a word, such as the singular form
Stemming	-	See lemmatization
Stopwords	-	A set of words and characters that is removed from the data before analysis as they add no information
Supervised	-	A problem where all the outcomes are known
Term document matrix	-	A quantitative representation of a corpus containing the frequencies of terms per document
Token	-	A word or character extracted from a stream of characters
Tokenization	-	The act of extracting tokens from streams of characters
Unsupervised	-	A problem where none of the outcomes are known

1. Introduction

> *"[W]hen you can measure what you are speaking about and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the state of science."*
> *(Lord Kelvin 1883, as quoted in (Swann, 2009, p. 34))*

Despite the extensive body of work that is being performed in innovation theory, many of its elements remain obscured. These opaque elements range from how to measure the phenomenon, the interplay of, as well as which, factors affect it, to its connection to economic growth (Fagerberg, Mowery, & Nelson, 2005) and innovations' cyclical behavior (Schumpeter, 1939; Silverberg & Verspagen, 2003). Within this spectrum of unsolved puzzles lies the problem at hand: the detection of *important*[1] innovations.

The problem of detecting and measuring which innovations are *important* originates in the issues of first detecting the innovations themselves. As the differences between innovations can be very qualitative, it is not surprising to see that even the detection of innovations is a problem with many different attempted solutions (Smith, 2005). Though one can, upon brief contemplation, easily identify a few technologies that have had a enormous impact on the world (steam engine, electricity, the microprocessor etc.), these technologies are not synonymous to individual innovations; "the steam engine" and "the computer" never existed (Verspagen, 2005). Rather, these technologies are a series of amalgamations of continuous innovations upon the basic, radical or major innovations that took decades to develop (Verspagen, 2005).

However, the detection of individual innovations that had a more moderate impact on the world, affecting a single industry rather than changing the world, typically requires a high level of specialized expertise or in-depth case studies, both of which were utilized by Wallmark and McQueen (1991) in creation of their list of 100 major Swedish innovations. The complexity of the task increases exponentially as the scope is widened to encompass an entire country rather than a single industry. As the phenomenon of innovation is largely qualitative they cannot be easily reduced to a quantitative comparison of their impact or actual novelty. Though there is a range of dichotomies assessing the contemporary aspects of the latter[2], this study will explore the prospect of detecting the former through articles from trade journals.

The source data for this exercise originates in one of the few gathering methodologies that measure actual innovation output, the Literature Based Innovation

[1] What this text means by *important* innovation is declared in section 2.4 on page13.

[2] Several innovation dichotomies are reviewed in section 2.3 on page 10.

Output (LBIO)[3], specifically the SWINNO database (Sjöö, Taalbi, Kander, & Ljungberg, 2014). Through analyzing a sample of the articles used as sources in this database, this study intends to achieve two things. Firstly, to detect *important* innovations without manually performing case studies, by applying three distinct, albeit naïve, algorithms that group articles based on the similarities in the occurring words. Secondly, in doing so, gauge whether the LBIO method can plausibly benefit from text- and data mining augmentation.

As the task of extracting digital texts from the image files of the articles, which are already gathered, is very time intensive, the study is limited to a single decade. There are several reasons for why the 1970s is a particularly suitable decade for this exercise. Firstly, it is the earliest decade in the database and the included innovations have, therefore, been given ample time to diffuse and leave their mark. Secondly, this decade intersects two external sources of major Swedish innovations (Taalbi, 2016; Wallmark & McQueen, 1991) that can be used to assess the outcomes. Thirdly, it is a particularly eventful period in Sweden's history, marking the end of a long economic growth and the beginning of a structural crisis that would have prolonged effects on the economy (Schön, 2009). Within this decade of upheaval the microprocessor arrived and became one of the major driving forces of innovation throughout the transition (Taalbi, 2014). And by the end of the decade innovation output surged, with a large share of highly novel innovations (Sjöö, 2014).

This study is divided into six parts. This introductory section will conclude after the research questions, aims, contributions and limitations have been clarified further. Section 2 will introduce concepts and definitions vital to the study. In section 3 the viability and origins of the data sources will be assessed. Section 4 describes the methods used, along with some unused options, to quantify the articles, a prerequisite of two of the methods. This section also explains how the three data mining techniques (sentiment analysis, k-means clustering and hierarchical clustering) will group the articles based on similar word frequencies, followed by how the reference data will be used to apraise the outcomes. Section 5 offers a walkthrough analysis of the model output towards the creation of a binary classifier for each technique used. This section culminates in the comparison of the three classifiers in order to elect one of them, and a short analysis of descriptive aspects of the resulting selection of *important* innovations. Finally, section 6 contains a brief discussion of the process of analysis and the results of the study and finishes with the conclusions and final thoughts.

[3] The problem of measuring innovation activity and various approaches to the problem are reviewed in section 3.1 on page 14.

1.1. Research Questions and Aims

Two of the fundamental assumptions for the LBIO method used in the compilation of the SWINNO database regard the editorial process. Editors are assumed to remain unbiased in what they report and possess enough expertise in their field to know what advancements are both relevant and interesting. If these assumptions are true, one could surmise that expectations of innovations based on this unbiased expertise would somehow be reflected in patterns in the articles' composition as well as the objects they choose to report on. Detecting and harnessing this pattern would then potentially enable the user to successfully identify extraordinary innovations.

1.1.1. Aims and Contribution of Study

Building upon one of the basic assumptions of the LBIO method, unbiased and expertise journalism, this study aims to explore the possibility of relying on trade journals to detect previously unnoticed *important* innovations from the 1970s. This will be done thorugh the creationg a classifier by comparing the presence of articles linked to major innovations[4] in groupings generated by text-mining algorithms. Each of these three algorithms (sentiment analysis, *k*-means clustering and hierarchical clustering) use different approaches to generate these groups, based on the word usages of the texts.

In a larger, more far reaching, perspective: by applying this limited number of simple techniques, this study aims to demonstrate that the text-mining toolbox offers viable options for extracting overlooked knowledge about innovations. Something that is conspicuously missing from the field, especially in light of the text-analysis intensive tasks of manually collecting data thorugh LBIO or case studies.

1.1.2. Research Questions

Though the focus of this exercise is to explore the application of rudimentary text mining techniques unto a subset of the journalistic texts that serve as the backbone for the SWINNO database, the ultimate reason for applying them serves an empirical purpose, as it allows for the absorbtion of overlooked knowledge. The examination of this dual nature is represented by two separate inquiries, where the second descriptive inquiry is conditional on an affirmative resolution of the first.

> *Can important innovations plausibly be identified by application of naïve and rudimentary text-mining on the articles reporting on them?*
>
> *How are the important innovations distributed among industries, firm and market novelty and artifactual and developmental complexity?*

[4] As assessed by two external sources: Wallmark and McQueen (1991) and Taalbi (2016)

1.2. Scope and limitations

The assumptions that serve as the foundation of this study are resting heavily on some of the fundamental assumptions of the LBIO method applied in gathering the SWINNO data. Whether they are accurate remains to be proven. Due to the time intensive task of generating the data, the study is limited to a single decade. Choosing the 1970s, the oldest decade recorded in SWINNO, gives the innovations ample time to diffuse and allows for a cross-reference with external data[5]. Furthermore, this decade is an eventful chapter in Sweden's economic history[6]. However, as there is nothing preventing journals from publishing long after innovations are diffused, there needs to be a cutoff point to prevent such articles from entering the analysis. As the this timeframe is unknown the innovations will be selected by the year of published articles rather than articles based on the innovations. This criterion also reduces the potential variance in terminology due to temporal factors.

Beyond the contents of the articles, one could look to their metadata for more information. An example of this is that, intuitively, one could easily expect the *important* innovations to have been written about more often. However, this is not necessarily the case as some of the selection by Wallmark and McQueen (1991) has not made it into the SWINNO database at all while one of the innovations with the most articles in the database is one that failed. Though this information by no means refutes such an approach, introducing this layer on top of the text contents would potentially add noise to the data and would complicate the application of the models.

[5] The subject of the external sources is discussed further in sections 3.4 and 4.3.

[6] Innovation activity and Sweden in the 1970s is reviewed in section 2.2.

2. Concepts, Definitions and Previous Research

> *"Technological change in the production of commodities already in use, the opening of new markets or of new sources of supply, Taylorization [sic] of work, improved handling of material, the setting up of new business organizations such as department stores – in short, any 'doing things differently' in the realm of economic life – all these are instances of what we shall refer to by the term of innovation. It should be noticed at once that that concept is not synonymous with 'invention'" (Schumpeter, 1939, p. 84)*

Before moving into some of the research related to the subject, it is pertinent to clarify what the term "innovation" means in this context. There are few restrictions on the forms of innovation; it can be both tangible and intangible, both implement new knowledge and reiterate age-old wisdom. Based on Schumpeter (1939), what matters is that the creation is *novel*, as that is the very definition of innovation (Smith, 2005), and that someone attempts to apply them in practice (Fagerberg, 2005), viz. they are introduced in a commercial setting. Only when both these criteria are fulfilled does something become an innovation.

2.1. Previous Research

With data collecting methods that rely heavily upon extracting information from texts, one could very well expect to see various attempts at automating part of this process. This does not seem to be the case. While there is quite a bit of research in the fields of the economics of innovation and data- or text mining both, little work seems to have been made to introduce even the most rudimentary data- or text mining techniques to innovation research. While searching for literature intersecting both fields, only a single case was uncovered: a dissertation by Hong (2013), who briefly analyzed the usage and co-occurrences of terms in the transcriptions of four interviews. However, the similarity with their research and this exercise ends there, as very different source material and techniques will be used here.

Innovation is an elusive phenomenon, the incidence and occurrence of which we might not be able to truly measure (Smith, 2005). This has led to a plethora of approaches simply for detecting, counting and measuring the economic act (ibid). Similarly there are various approaches towards dividing innovations on their contemporary or overall novelty. Both these subjects will be approached further in sections 3.1 and 2.3 with a clarification of what in this text is meant by an *important* innovation in 2.4.

2.2. Innovation Activity and 1970s Sweden

Schumpeter (1939) noted that innovations neither isolated events nor are their distributions random or uniform. Schumpeter had instead observed a tendency of innovations to bundle where many innovations followed “in the wake of successful innovations” (p. 100). Though it has been nearly a century since this observation was made, the innovation process remains an elusive phenomenon that is sometimes referred to as a “black box”, as we do not know what goes on inside it (Fagerberg, 2005; Jaruzelski, Staack, & Goehle, 2014; Ljungberg, 2004). As there are many interplaying factors involved in what entices, enables and induces firms to innovate (Lazonick, 2005; Lundvall & Borrás, 2005) and in spite of the numerous studies into this contingent process of innovation, it remains without a theoretical consensus (Oecd/Eurostat; Pavitt, 2005). With all these heterogeneities in the field, the importance of knowledge is often emphasized (Greve, 2007; Kuznets, 1973; Malerba, 2005; March, 1991; Mowery, 2005).

The 1970s was an eventful time for Sweden. It marked the end of a long period of remarkable economic growth and in the later half it was hit by an international crisis with prolonged effect on the economy and its industry was forced into a transformation (Schön, 2009). Amidst the restructuring, a development block of electronics was formed based on the microprocessor, which had appeared in the middle of the decade (ibid). The opportunities offered by the microelectronic revolution was one of the key forces behind Swedish innovating in the following decade and several notable innovations were put forth (Taalbi, 2014). By the end of the decade the outputs of innovations and radically new innovations, surged to levels not seen in the following decades (Sjöö, 2014).

2.3. Novelty Dichotomies

The inherent problems of measurement lie in the reality that “innovation is, by definition, novelty” (Smith, 2005, p.149). This quality is also what gives rise to the multitude of dichotomies that all start from the Schumpeterian definition of innovation that requires novelties to be introduced to the market in order to become *innovations*, before which they are merely inventions.

When Schumpeter (1939) introduced this approach to market news, he also categorized innovations on their novelty, relative to preceding innovations, into two categories: radical and incremental. The names of these two groups are intuitive as the former refers to innovations that bring something new and the latter are improvements of existing innovations (Fagerberg, 2005). However, this dichotomy is rather static, it does not consider that innovations can be combinations of these two aspects.

Henderson and Clark (1990) aim to, at least partially, solve the problems of this perspective in the case of products by analyzing technological change on two criteria: core concepts and linkages between core concepts and components. By use of these two axes they create a division of four categories, adding architectural and modular innovations, represented in Figure 1. In this view, *incremental* and *radical* are absolute states that challenge the old not at all or completely, respectively, and the new categories challenge the old in only one of the two aspects. Examples offered for these new categories include the architectural innovation of transforming from a ceiling fan to a portable fan, as it reinforces the core concept of moving air while components are linked in a new way. Their example of a modular innovation is that of the substitute of analog to digital phones. While their framework adds a layer of depth and complexity to Schumpeter's two categories, it tells us little of the innovations' novelty or importance to the rest of the world.

Figure 1. A framework for defining innovation.

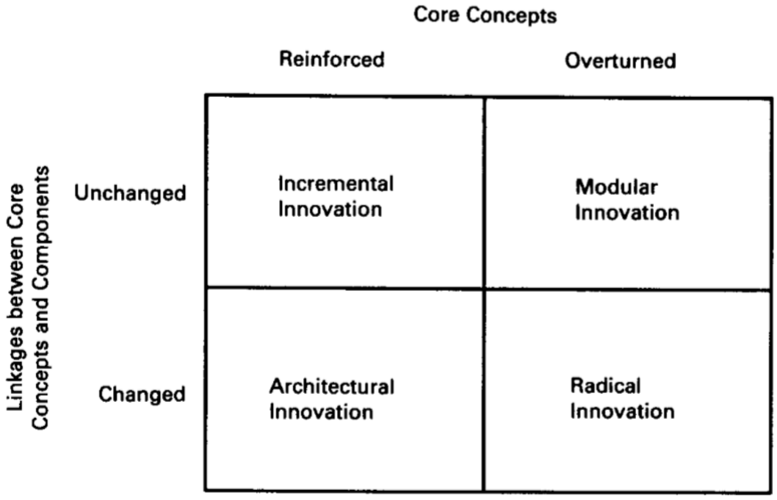


Figure 1 A framework for defining innovation (Henderson & Clark, 1990)

Few of the technologies and discoveries that have appeared have shaken the world and gave rise to a broad array of innovations. The form of these General Purpose Technologies (GPT) can range from the harnessing of a new material (eg: Bronze) or chemical reaction (eg: Electricity) to the application of new ideas (eg: Writing). According to Lipsey, Carlaw, and Bekar (2005), a total of 24 *transforming* GPTs have emerged throughout human history. Since the GPTs are defined through their novelty and impact, little doubt is left whether they were important. However, their importance is not necessarily derived from a single innovation's spread, but rather its reapplication in and influence on later innovations (eg: electricity has opened the possibilities to the internet).

The SWINNO database Sjöö et al. (2014) records, among many other things, a measure of the included innovations novelty. Instead of relying on existing dichotomies, it uses its own two relative measures in order to avoid recording the grey scale of novelty on a black-or-white scale. Firstly, they document whether the innovation was incremental, a major improvement or entirely new for the company. Secondly, they also document whether the innovation was new to the Swedish market or the world market. While this approach captures a different dimension of novelty, where the sources permits, it too fails to account for any impact these novelties might have yielded.

Wallmark and McQueen (1991) compiled a list of 100 Swedish major technical innovations from 1945 to 1980. Their selection process starts from screening annual reports of important technical innovations published by The Royal Swedish Academy of Engineering Sciences. Wallmark and McQueen's further criterion for an innovation to be included in their list was that the innovations should have been financially successful, resulting in a cutoff at $30 million (1980s monetary units) revenue attributed to the innovation by the entrepreneur. While they do admit that it would be ideal to focus on profits generated by the innovation, they were prevented from doing so, due to the complexity of estimating this value.

Instead of focusing on the innovation itself, other views focus on the businesses attitude towards knowledge in the process of innovation. In the framework by (March, 1991) the attitudes of entrepreneurial firms are somewhere on a continuous scale between exploration and exploitation. On the former absolute, the firm is entirely focused on applying exclusively new knowledge in their innovations. However, this does not imply that the innovation itself is radical, or particularly novel, just that they rely a great deal on knowledge creation in order to innovate. On the end of the spectrum, the firm relies entirely on existing knowledge in their entrepreneurship. While this approach does not dictate the novelty or the importance of the innovations, it helps illuminate the complex relationship between innovations already in their developmental stages.

Disruptive technologies is a term referring to the tendency of companies to market technologies in a new package, often worse than existing technologies, in order to reach new markets or market segments (Bower & Christensen, 1995). The focus of this perspective is less the technology itself and more its impact on the market, hence the name 'disruptive' (Nagy, Schuessler, & Dubinsky, 2015). In terms of Schumpeter's incremental or radical dichotomy, this perspective does not seem to fit, since that spectrum leaves little room for decrements in performance. This contrast is important to note since it shows that not all innovations are necessarily technologically superior to their contemporaries. Instead, they can be a downgraded or simplified versions that appeals to new customers.

In the variety of the reviewed innovation dichotomies, each focus in a different qualitative aspect of innovation novelty. Only two dichotomies approach the impact of the innovations, Lipsey et al. (2005)'s GPT and Wallmark and McQueen (1991)'s Major innovations. The former of which focuses on large transformative collections of innovations, the latter is based on expert opinion, case studies and financial results connected to the innovation.

2.4. Defining Important[7]

As the existing dichotomies dedicate no interest to the prevalence of influential innovations below a high threshold, a new dichotomy is here defined. Compared to the most basic dichotomy of radical versus incremental innovations, it might seem intuitive to assume that radical innovations, introducing never before seen elements, are likely to often be more *important* than incremental innovations, mere improvements of old innovations. However, as a radically new innovation might not be at all desired and a tiny improvement in performance could have been a long sought change enabling the expansion of new frontiers, this is not necessarily the case. Measuring the exact impact of an innovation on both the entrepreneur and its potential users requires the untangling of an uncountable number of events. The complexity of this task comes from the fact that innovations can be implemented, diffused and influential in a vast variety of ways, such as finding a new and unpredicted implementation for an old innovation (Maclurin, 1953). Therefore the definition of *important* remains rather open. Nevertheless, there are a few criteria that aid in the identification of *important* innovations: extent and mode of diffusion as well as market results.

Firstly, the innovation has to be implemented, by entrepreneur or user, before it can be shown to be *important*. Since the threshold between invention and innovation is merely market introduction, it does not prevent inconsequential ideas that have no real impact to become innovations. *Important* innovations have left their mark upon the market without necessarily shifting it; GPTs and the like are beyond the bounds of *important*. On a novelty scale, *important* innovations occupy the strata between world-shaking discoveries, as these need not be rediscovered, and humdrum. Due to the unpredictable nature of diffusion (Hall, 2005), *important* innovations can be spread out almost anywhere among the novelty dichotomies.

Secondly, for an innovation to have been important someone has to have generated a measurable level of revenue from it, in the case of Wallmark and McQueen (1991) the cutoff point was $30 million in 1980s currency. However, the financial benefactor does not have to be the inventor or entrepreneur since ideas can be copied, stolen or in the case of sold innovations: the benefits might be greater for the user than the entrepreneur. While both of these aspects can, to some extent, be quantified, at their cores they, just like innovations, remain very qualitative and are therefore difficult to reduce to a more concise set of criteria.

[7] In order to make a clear demarcation of this particular meaning of *important* and the regular meaning, I will refrain from using it in any other sense and clearly demark it with italics.

3. Data

3.1. Measuring Innovation Activity

The actual innovation output is something that is inherently difficult to measure, if possible at all (Smith, 2005). Due to these difficulties it has to be estimated through approximations and, as a result, multiple indicators have been used to approximate the output of innovation activities.

A commonly used approach is to rely on the records of patent offices (Beneito, 2006; Smith, 2005). This approach has the advantage that long time-series are available, however, the relationship between patents and innovations is not necessarily one-to-one (Basberg, 1987; Beneito, 2006; Verspagen, 2005). A patent does not imply any form of commercialization, but strictly the temporary monopoly to do so, and patents might therefore be an indicator of inventions rather than innovations (Basberg, 1987; Nelson, 2009). Due to the difference in marginal costs between first-time and repeat patentee, they might not display the same behavior towards patenting the same innovation (Kleinknecht, Van Montfort, & Brouwer, 2002). A further problem with such data is the potential influence of changes in Intellectual Property Rights (IPR) policy (Lerner, 2002) and what incentives policies enforce on the patenting behavior in the overall economy or different sectors (Granstrand, 2005; Hagedoorn & Cloodt, 2003).

As there is no requirement to commercialize a patented invention or to patent an innovation, the relationship between the two remains obscure (Hagedoorn & Cloodt, 2003), partially because not all innovations can be patented and partially because the former is an economic act and the second is the utilization of a legal tool that can be used not for the intent to commercialize, but rather to interfere with competition (Granstrand, 2005; Smith, 2005). For those seeking to protect their intellectual property, patenting might still not be their prime choice as it forces them to make technical details publicly available and can be costly, instead, they might resort to simple secrecy (Basberg, 1987).

Similarly, the input into the innovation process is sometimes approximated through spending on Research and Development (R&D) and offers the longest available time series (Smith, 2005). Based on the funds or man-hours dedicated to R&D, this indicator allows for as long a period of study as such data has been recorded. However, not all companies formally dedicate their resources to R&D (Kleinknecht et al., 2002). This measure also neglects the potential impact of tax and policy incentives placed on this activity. Furthermore, for companies that maintain development of multiple projects simultaneously, untangling which input is connected to which output can prove difficult (Kleinknecht et al., 2002).

A third way to gather information on innovation activities is through Community Innovation Surveys (CIS) (Kleinknecht et al., 2002; Smith, 2005; Swann, 2009). This rather flexible approach allows for questions to target some of the qualitative information missing from the other approaches. However, this method has s few issues as well. Firstly, the surveys are typically not sent out to small companies. Secondly, the gathered information is self-reported and one must therefore trust that the companies are not distorting the truth in their answers, if they answer at all. Thirdly, since the data has to be collected close to the year it concerns, one cannot use this method to investigate periods before the earliest survey (Oecd/Eurostat, 2005).

Expert opinions, panels or interviews make up a further source for detecting innovations (Kleinknecht et al., 2002; Smith, 2005; Swann, 2009), which is the foundation of Wallmark and McQueen (1991)'s 100 major Swedish innovations. One of the earliest databases to rely on this methodology was created by the Science Policy Research Unit (SPRU) at the University of Sussex and played an important role in the development of the discipline (Fagerberg, 2005). The database covered over 4,000 innovations commercialized in 1945-1983 (it was discontinued in 1984) by multiple means of collection, including a panel of circa 400 experts (Smith, 2005; Swann, 2009). Though this overall approach does not limit the kind of innovations one can find, it can suffer greatly in respect to costs. An approach that can still rely on expert opinions without the panels is the Literature Based Innovation Output (LBIO) approach, which is based on the reviewing of independent literature for identification of actual innovations, as opposed to approximations through patents or R&D.

Among the multiple datasets collected through the LBIO method the approaches differ slightly, some simply rely on the new product announcements (Coombs, Narandren, & Richards, 1996; Gerben van der, 2007; Hagedoorn & Cloodt, 2003; Kleinknecht et al., 2002; Smith, 2005) and the database utilized in this study, SWINNO, relies upon all the edited texts that pertain to domestic innovations (Sjöö, 2014; Taalbi, 2014). A key requisite of this method is the existence and detection of independent trade journals. Once journals have been identified the only thing limiting the length of a study is the duration in which the journals have been published. In this approach there is no non-response problem as firms need not be contacted for the relevant information, though it is assumed that firms have some incentive to make their innovations public (Link, 1995).

Each of the above-mentioned approximations of innovation activity has their own set of strengths and weaknesses. Amid this range of data forms, the 'true' innovation output is still unknown, so verifying the methods is a problem yet to be solved. With the discrepancy between the various approaches, the choice of indicator can have a direct effect upon the results (Kleinknecht et al., 2002), as the correlation between the different estimators vary across industries (Hagedoorn & Cloodt, 2003). In light of this insight, the reliance on the SWINNO data, and through it LBIO, it is certain that each object in the dataset represents an actual innovation.

3.2. The SWINNO Database

SWINNO (Sjöö et al., 2014) is a database of Swedish innovations, which was constructed and used by Sjöö and Taalbi for their dissertations (Sjöö, 2014; Sjöö et al., 2014; Taalbi, 2014). The original data was gathered through use of LBIO by screening 15 journals over a 38-year period, resulting in a collection of almost 4,000 innovations and cites over 6,000 articles as used sources. The following overview of the gathering process focuses on the most relevant elements to this study and further details on the entire process can be found in Sjöö et al. (2014), (Sjöö et al., 2014) and Taalbi (2014).

In order for innovations to be entered into the database it was not enough for an innovation to simply be mentioned in an article, three criteria had to be fulfilled, in addition to the prerequisite of being Swedish (Sjöö, 2014). Articles had to mention both (1) a commercial interface and a (2) commercializing agent, both these in order to avoid the recording of mere inventions. (3) Only innovations with explicit information regarding the form of the novelty are included, to avoid innovations that are minor increments.

While this selection process prevents the recording of inventions and miniscule incremental innovations or lateral differentiations, it also introduces bias to the data (Sjöö, 2014). By focusing only on the innovations that successfully are introduced to the market there is a success bias. In addition, the abstract and complex nature of some services sometimes makes it very difficult to discern their novelty and as a consequence there is a product bias in the data.

The filters and biases imposed by the LBIO method applied in the construction of the SWINNO database do not pose great problems to this study. The first two criteria ensure that only innovations, and no mere ideas or inventions, are even considered for recording into the database. The third criterion reduces this set of candidates further by ensuring that the innovations are indeed new. None of these constraints contradict the definition of *important* innovation and therefore do not explicitly exclude any *important* innovations from entering the data. Sjöö (2014) uses the analogy of an iceberg to describe the success bias in the data, proposing that the innovations included in the database are sufficiently remarkable to float above the surface and subsequently be detected by trade journals.

A lot of different data is recorded regarding the different innovations; most are simply recordings of the information divulged in the sources and some are based on assessments based on the texts. One example of the latter, which pertains to this study, is the five-digit product code based on the 2002 SNI (Svenskt Näringslivsindelning) definitions which, based on the first two digits, allocates the innovations into 22 industries (Sjöö, 2014). The other variables of interest are all of the former type; they simply record the data from the article. The first pair is the novelty variables that record the relative novelty of the innovation to the firm (incremental improvement, major improvement or totally new) and relative novelty to the market (new to domestic market or new to world market). The second pair refers to the complexity of the innovation on two axes, artifactual and developmental, by three ordinal steps; ‘low’, ’medium’ or ‘high’.

3.3. The Corpus

The source data for this exercise taken from SWINNO are the articles used as sources in the database's construction. One of the key assumptions of the LBIO method is that the journalists selecting and writing about these innovations are knowledgeable within their field. Ergo, if they claim that something is new and interesting to a certain field, it is so[8].

SWINNO contains over 8,500 unique journalistic texts[9] out of which 1,462 were published in the period of interest. When restricting the corresponding innovation sample to products and processes that reportedly have entered the market, the number of articles is further reduced to 1,288. The characteristics of these texts vary greatly, text length spans from but a paragraph to several pages and any article may focus on a single innovation or regard a group of them. Nearly all the used articles have been digitalized, the vast majority are scanned pages but in more recent years some publishers have chosen to provide their journals digitally. All the texts have to be digitalized as well before any text mining software or method can process them. The most viable option to achieve this is through Optical Character Recognition (OCR) software[10]. Due to the potential inaccuracies introduced by these programs (Croft, Harding, Taghva, & Borsack, 1994) it is generally advised to avoid using them if there are other options (Weiss, 2005), which is not the case here. Due to the time intensity of extracting digital texts from the source-files the analysis has been limited to a single decade rather than the entire period covered by the database.

There are two practical reasons for why this single decade is the 1970s, which is the earliest period of recorded data in the database. Firstly, because of the varying length of diffusion of innovation it is reasoned that an older dataset increases the chances that included innovations have been given ample time to leave their mark. Secondly, the period in question overlaps with to lists of major innovations compiled by Wallmark and McQueen (1991) and Taalbi (2016), allowing for an attempt to verify the models.

The articles used in SWINNO are categorized into five different groups depending on the contents and journal section of the article. Two of these types are, as a rule, related to and write exclusively about a single innovation, Innovation Focus and Product News. As the retreival of only the sentences and paragraphs pertaining to the particular innovations would require manual extraction, and leaving them in implies a lot of noise, the article types with less focus will not be included in this analysis, nor will articles linked to multiple innovations. After removing these articles 1,061 (82.38%) out of the original 1,288 articles remain for analysis, pertaining to 936 (79.39%) of the original 1,179 innovations. The remaining articles are unevenly divided between Innovation Focus (870) and Product News (191).

[8] Allowing for lapses in judgement.

[9] As of 2 Feb 2016, which has been updated to include material published 2008-2014.

[10] In this case: FineReader for OSX.

Table 1 Articles and Innovations in the Corpus

	All	Innovation Focus	Product News	Sample
Innovation Focus	1062	870		870
Overview	61			
Other	84			
Fair	45			
Product News	207		191	191
N/A	3			
Articles	1462	870	191	1061
Innovations	1329	767	169	936

Before the texts are introduced to the models, at least some of the mistakes introduced by the OCR software needs correcting. Due to the varying quality and resolutions of the scans of the articles the software was sometimes unable to recognize that certain characters are not relevant to the text. A prime example of this is the end line hyphenation of words, which many times were not corrected by the OCR software, resulting in some peculiar terms. Another example is the tendency of some journals to indent the first row of a new paragraph with a geometric shape that then became the first character of a word. However, due to the variety of the adjacent characters and acceptable uses of hyphens the task is more complicated than simply removing all the hyphens and special characters. A relatively simple script[11] was created to conservatively deal with multiple different patterns that occurred with varying frequency. Once the texts have been cleaned, they become ready for processing.

3.4. The Reference Data

Wallmark and McQueen (1991) compiled 100 case studies of Swedish innovations that they categorized as *major*, that were commercialized the between 1945-1980. They started with a set of 176 innovations constructed from the annual reports of The Royal Swedish Academy of Engineering from 1945-1975 (they do not clarify the detection process for the final years). Though they do not clarify the criteria, they mention that, through time, innovations were added and removed from this list before the final selection of 100 innovations. The final selection criteria were: to be novel enough for a “meaningful” patent, the possibility to identify the innovators and financial outcome; only the 100 innovations with the higher annual turnover attributed to them were selected. The authors admit that this last criterion is arbitrary. They also make it clear that their selection, and their initial set of innovations, does not necessarily contain the topmost major innovations of the period. Rather, they admit they can make no such claim and maintain that they have selected 100 innovations from that stratum.

[11] The used Python (2.7.11) script is available upon request.

Table 2 Reference Articles and Innovations

	Joint	*W&M*	*Overlap*	*Taalbi*
Articles	54	15	7	46
Innovations	30	9	5	26

Out of these 100 innovations 27 pertain to the period covered by SWINNO and 20 of these are recorded in the database. Furthermore, 17 out of these 20 were commercialized in the 1970s, though seven of them are not connected to an article written in the 1970's. Out of the remaining 10 innovations one cites only a single article, which is excluded from this analysis as its source files are incomplete. This leaves only nine innovations to act as a reference for detecting *important* innovations, represented by 15 articles.

The second group of references comes from a set of 151 innovations from the original SWINNO database, picked out by Taalbi (2016). Just as for the compilation by Wallmark and McQueen (1991), there is no claim that these are the top 151 most influential innovations of the period or the original sample. The selection process is based on the author's expertise within the area, many inclusions of which are supposedly fairly obvious to the initiated, rather than a rigorous empirical method. As such it is not as reliable as the preceding reference category, but for the intents and purposes of this paper they are deemed sufficiently so.

Out of this selection of 151 innovations 26 are included in the dataset used, represented by 46 articles. The combination between the two sets of references allows them to jointly cover 30 (3.21%) innovations and 54 (5.10%) of the articles used in the sample with five innovations appearing in both of them.

4. Methods

In data mining, problems where the outcomes are known are called *supervised* problems. The problem at hand, however, is *unsupervised* since the outcome is not known (Basu & Davidson, 2009; James, Witten, Hastie, & Tibshirani, 2013); in fact, estimating this outcome is the goal. By introducing the reference data from Wallmark and McQueen (1991) and Taalbi (2016) containing approximations for the outcomes for a small fraction of the data that can be used for verification of the models, it then becomes a *semi-supervised* problem.

A notable issue with the explorative approaches applied here is that there are few guidelines in how to apply them. In combination with the numerous forms and variants in the modeling and data preparation, many arbitrary decisions need to be made and any road taken will leave plenty of openings to criticism. In most of these cases the road that implied fewer assumptions will be taken, in the spirit of Occam's razor. Therefore the term document matrix data will not be transformed after its creation and the clustering algorithms will be applied in their basic forms. In contrast, as there is no direct interpretation of the cluster algorithms, the results sometimes require some manipulation before they can be interpreted.

4.1. From qualitative to quantitative data

In order for quantitative models to be applied on the data the corpus has to be transformed into a *dictionary*, the list of terms used, and subsequently a matrix. The first step is *Tokenization*, where instances of words, phrases and special characters are separated from the text to be counted and entered into the *dictionary*. As a lot of different words tend to be used, the *dictionary* needs to be reduced through a series of steps. *Lemmatization* standardizes the tokens, reducing the number of types and increasing their frequencies by reducing terms to their *stems*. A potential problem with this step is that these algorithms do not differentiate between words that are spelled the same, but have different meanings or different forms of the same word[12] (Weiss, 2005). The former problem can be dealt with through *inflectional stemming*, separating words on meaning based on a created *dictionary*, the latter by *stemming to a root*, condensing and counting words by their core meanings.

After the above transformations, the *dictionary* will contain counts for every single word and special character, as these tend to be quite many it needs further reduction, for which there are multiple approaches. Local *dictionaries* can be kept for different types of texts, in this exercise this could be done by article type or journal. This particular approach does not apply here, as the goal is to compare all the texts with each other. *Stopwords* is a list of words that can be removed from the dictionary without loss of data, such as: "a", "the" or words that appears frequently in most texts (Weiss, 2005).

[12] The noun "bark" is not the same as the verb "to bark", while being essentially the same as the plural form "barks" and "flew" is the past tense of "fly".

Once the dictionary has been reduced by the above methods the *term document matrix* can be created, in which columns represent terms and rows represent documents. The elements of the matrix represent the frequency of the corresponding term in the corresponding document. This approach does not take into account the sentence structure or even the order of words; instead the texts are simply treated as *bags of words* from which terms are counted (Zhai & Aggarwal, 2012). Since little information is gained by the exact count of words and phrases used multiple times, especially in this case where the length of texts vary greatly, instead of recording the exact count, it will simply record the integer 2. Thereby, each element of the matrix will contain 0, 1 or the standardized value of 2 (Weiss).

There are also various purely quantitative methods, of varying complexity, of feature reduction in the *term document matrix* that aim to reduce noise without loss of relevant information. However, as one has to arbitrarily select the number of features that should remain after the reduction (Zhai & Aggarwal, 2012), these will not be applied here.

4.2. Classification Through Clustering

There is a plethora of approaches and methods available to those who whish to categorize objects. However, this paper only has the potential to test a very limited sample of this cornucopia, due to the *semi-supervised* nature of the problem. In order to keep things simple, three of the most rudimentary techniques in their basic forms will be applied, starting with a very simple and naïve model requiring only the terms used and a list to compare them to. The following two clustering models are slightly more, both of these latter methods will use the term document matrix as input because they require purely quantitative data. The clustering methods used are overall easy to understand as they both can be summarized accurately in a few sentences. Nevertheless, they can still be difficult to apply as choosing the number of clusters is "a notoriously hard key problem in cluster analysis" (Hennig, 2014, p. 112).

4.2.1. Sentiment Analysis by Lexicon

The first suggested approach is by far the humblest one methodologically; singling out the texts with a positive sentiment by counting the number of positive and negative phrases used in each text, dividing into three groups: positive, neutral and negative. By ignoring the syntax of the texts some information is lost and Benamara, Cesarano, Picariello, and Subhamanian (2005) showed that adjective-adverb combinations are more accurate than the naïve approach, in analyzing 200 news articles. Still Ding, Liu, and Yu (2008) (as cited in Dadoun and Olsson (2016)) have shown that even the naïve version yields sensible results.

Creating a *lexicon* with all the positive and negative terms and phrases available in Swedish is a laborious task. Instead of wasting time on this a readily available lexicon will be applied. The lexicon used is based on the openly available Affective Norms for English Words lexicon created by Nielsen (2011), a manually created list of 2477 phrases, each with a score in the range [-5;5]. This lexicon has been translated into Swedish, tested and again been made publicly available by Gustavsson (2016).

Though the terms in this *lexicon* are attributed with magnitudes, the polarities will also be used to create two different groupings based on this algorithm.

4.2.2. *k*-means Clustering

Clustering techniques are used to group observations by the similarities in their quantitative features, as such *k*-means clustering is the first approach that relies entierly on the term document matrix. Regular *k*-means clustering relies on Euclidean[13] distance to determine similarities between observations and relies on an iterative process to group observations. Each of the *k* groups' *centroid* is calculated, each iteration, to be the point that minimizes the squared error of its group. Observations are then assigned to the group of their closest centroid. These two steps are repeated until the algorithm can no longer reduce the errors by reassigning observations between clusters, or a preset maximum. This whole process is repeated a few times, since each observation is randomly classified at the start, and the solution(s) with the lowest error, distances to assigned *centroid*, is chosen. Being a rather simple approach it requires very little of the applier, except assigning a value to *k* (James et al., 2013; Weiss, 2005; Zacharski, 2015; Zhai & Aggarwal, 2012). In the case of two-, and sometimes three-, dimensional data, an expectation of the number clusters can be generated from an ocular inspection of the data. However, in this case there number of features is so large that the visual inspection is virtually impossible.

While the goal of this exercise is to distinguish the *important* innovations from the rest, it does not logically follow that $k = 2$. One of the key expectations is that the journalist's expectations or exuberance of *important* innovations is somehow reflected in their choice of words. However, due to the variety of words at their disposal one cannot reasonably assume that the same words would be used in each case. Even if all the texts pertaining to the less *important* innovations fall into the same cluster, the *important* innovations could still be spread out at different distances from this cluster. Not knowing how many clusters to look for means that this number has to be arbitrarily selected. Weiss (2005) suggests that one could test for this by applying the model with an increasing number of clusters until additional clusters offers no corresponding decrease in variance.

4.2.3. Hierarchical Clustering

Hierarchical clustering is a bottom-up approach that sequentially pairs clusters of observations, based on their similarity, until a single cluster remains. The resulting clustering is then divided top-down to reach the number of clusters required. Since the sequence of aggregations is tracked one does not have to select the number of clusters one is searching for (James et al., 2013; Weiss, 2005; Zacharski, 2015; Zhai & Aggarwal, 2012) *Hierarchical* clustering differs itself from *k*-means particularly in that the user does not have to specify the number of clusters the model should look for before applying the algorithm to the data (James et al., 2013). While this greatly

[13] While Euclidean distance is the default, and what is applied here, any formula for calculating the distance between two points in n-dimensional space can be used.

simplifies the application of the algorithm, the method raises another problem that can greatly affect the results and is, as of yet, without a consensus solution: how to link clusters. Three common solutions are *average-*, *complete-* and *single-linkage* (Weiss, 2005; Zacharski, 2015; Zhai & Aggarwal, 2012). The first is self-explanatory: the average distance between all the points in two clusters are used as the distance between the clusters. The latter two models are opposites in that the former relies on the two furthest points and the latter the two nearest points, to represent the distance between clusters. Out of these three approaches *average-* and *complete-linkage* tends to create the more balanced clusters (James et al., 2013) and therefore these two linkages will both be used in two different models. Furthermore, a third linkage method, *ward*, relying on the variance between groups, is known to create even more balanced clusters (Pedregosa et al., 2011), and will also be used.

Similarly to *k*-means clustering, it relies on Euclidean distance[14] between clusters and it too is an iterative process. Each iteration, the two closest clusters are joined into a larger cluster, with single observations treated as clusters of one. This is repeated until only a large cluster remains. By keeping track of how the agglomeration is performed any number of clusters can be chosen and tested without having to rerun the model, since the outcome will always be the same.

4.3. Evaluation of the Models: Verification

The actual accuracy of any models cannot be calculated; due to not knowing beforehand exactly which innovations have been important and not knowing any that were not. However, by relying on external assessments of major innovations, particularly those of Wallmark and McQueen (1991) but also from a selection made by Taalbi (2016), the model outputs can be evaluated. It is well worth noting that Wallmark and McQueen's list focuses exclusively on innovations in the field of engineering and the innovations recorded in SWINNO are not limited to any one discipline, though it is somewhat biased towards products. Being aware of this bias creates a caveat that is not easily dealt with without further data. In the event that no services are suggested as being important by a model, it could point towards none of the tested services being *important* or simply because services are discussed, by the journalists, in a very different manner from the engineering-based innovations from the reference list.

As there are two separate selections, they could be applied in four ways (not counting ignoring them altogether). In evaluating the models, the references will be applied both jointly and individually. In the cases of minor disagreements between two lists, the edge is awarded to Wallmark and McQueen (1991) as their selection is far more rigorous. Since the overlap between the two selections is only seven article and five innovations of the sample, it is rather small to be relied upon to find other important innovations.

[14] See note 6.

The two, and subsequently created four, sets of references will be referred to frequently in the following section as their presence in the different clusters and groups will be used to compare these groups. In order to make the reading of these comparisons less strenuous, each of the subsets will receive their own handle. The selection by Wallmark and McQueen (1991) referred to as *W&M*, the selection by Taalbi (2016) as *Taalbi,* the joint selection of these *Joint* and the small overlap as *Overlap*. It follows that these aliases will be utilized in the tables and figures as well.

A particular problem with relying on these data is that they did not record innovations that made no impact. This means that there is no clear way of detecting false positives in the output. Unlike an econometric regression, one will not be able to state a level of significance or assess the probability that the models are indeed yielding the desired results, as such statistical tools do not apply under these conditions.

The references will be applied both to select a single classifier to represent each of the algorithms and to then select the classifier that will be applied to the sample. In a perfect world, every single model will point out exactly the same innovations and will include every single reference innovation while excluding a large portion of the sample. However, the world tends to resist such simplicities and the best one can realistically hope for is to have a high level of overlap between each model that includes a majority of the reference innovations and still leaves a reasonable portion without the label *important*.

5. Applying the Models

5.1. Creating the Term Document Matrix

After the tokenization and stemming, processed by the Natural Language Toolkit (Bird, Loper, & Klein, 2009), the first set of tokens that were prevented from entering the term document matrix contained various special characters that have no valuable interpretation. Still, over 45,000 unique tokens remained, though the vast majority of these appeared only once, were single characters or contained non-alphanumeric characters in such a way that their original meaning was impossible to distinguish, leaving just over 14,000 features for further reduction. Due to the arbitrary nature of selecting a number of features, as would be required by any of the purely quantitative approaches to the problem of reduction, coupled with the need to manually check terms for errors introduced by the OCR, the feature selection was subjected to manual inspection. Manually sifting though the terms facilitated manual removal of unintelligible strings of characters, inflectional stemming, merging words of similar meaning and stemming to a root, stemming words to their core meaning, to be done simultaneously. Furthermore, names of people, places and companies were removed in this stage, as these are inconsequential to innovation *importance*. After this reduction 6,427 terms remained, in comparison to 1,061 articles.[15]

5.2. Sentiment Classification

Out of the three methods applied here the sentiment classification is the only one that does not require the term document matrix. Instead the scores are calculated by comparing each word in the texts to the lexicon to retrieve the corresponding score. The virtue of the sentiment classification lies in its inherent simplicity of comparing the number of positive and negative terms. In relying on an existing lexicon that included magnitudes, two different approaches have been used. The first approach relied simply on the polarity of the lexicon's term, counting the positive and negative against each other and then dividing the texts into the three groups based on the resulting polarities. The second approach relied on the magnitudes of the terms before classifying the texts into the same categories by the texts' overall polarity.

At a first look at Table 3, there are three things that are of particular relevance to the text sentiment classifications. Firstly, both approaches are overwhelmingly classifying articles as positive and more negative than neutral. Secondly, a similar pattern holds through the Innovation Focus articles and a there is a disproportionate concentration of neutral Product News articles. Thirdly, the distributed of the references between the article types, and between the different groups, is close to the expected values. Thusly this model performs only marginally better than what could

[15] As it does not fit here, the term document matrix is graphically viewable in Appendix A, where each coloured square represents an element from the matrix.

be expected from simply picking 54 articles at random. However, this ignores that this selection represents 30 innovations with varying number of articles.

The two approaches are quite consistent, 92.74% overlap, and they jointly pick out 821 (77.38%) articles as positive and these represent 715 (76.39%) unique innovations. However, this excludes two innovations with five articles each, out of which only a single article was classified as neutral rather than positive. It is therefore not enough to simply rely on the perfect overlap of the classifiers.

Table 3 Sentiment Analysis Results by Article Types Compared to (*Joint*)

		Magnitudes						
		Negative		Neutral		Positive		
Polarities	Negative	112	(3)	16	(1)	8		All Articles
	Neutral	13	(1)	51	(1)	30		
	Positive	5	(1)	5		821	(47)	
		Negative		Neutral		Positive		
	Negative	81	(2)	13	(1)	7		Innovation Focus
	Neutral	9		25	(1)	22		
	Positive	5	(1)	5		704	(43)	
		Negative		Neutral		Positive		
	Negative	31	(1)	3		1		Product News
	Neutral	4	(1)	26		8		
	Positive	0		0		117	(4)	

Relying on the underlying categories one can easily take the classification one step further, using the polarities of these labels to reach the innovation's overal sentiment. By this approach, represented in Table 4, the number of innovations labeled as *important* change to just over 700 and the different approaches labeled all of *W&M* while missing the same three observations from *Taalbi*[16]. Worth noting is that these three innovations are connected to three Innovation Focus articles and a single Product News. Furthermore these selections are picked out better than one would expect from a random assignment.

Table 4 Sentiment Analysis of Innovations Compared to References

Selected By	Unknown	Taalbi	Overlap	W&M	Total
Neither	172	3	0	0	175
Both	690	17	5	4	716
Magnitudes	36	0	0	0	36
Polarities	8	1	0	0	9
Sample	906	21	5	4	936

Overall, it seems as if this very simple classification is able to identify a lot of *important* innovations in the sample. However, seeing that it picks out a suspiciously

[16] These four innovaitons are visible in Appendix B

high portion, over 70%, its selection is difficult to justify. Looking further into the distribution of the sentiment scores of the innovations, Figure 2, one can observe that the overall distributions of standardized scores are fairly similar to one another; skewed so that a large portion of the observations have positive scores but close to zero. As the concentration of the references data is more skewed towards positive than the overall data, it is possible that a threshold classifier would be of use.

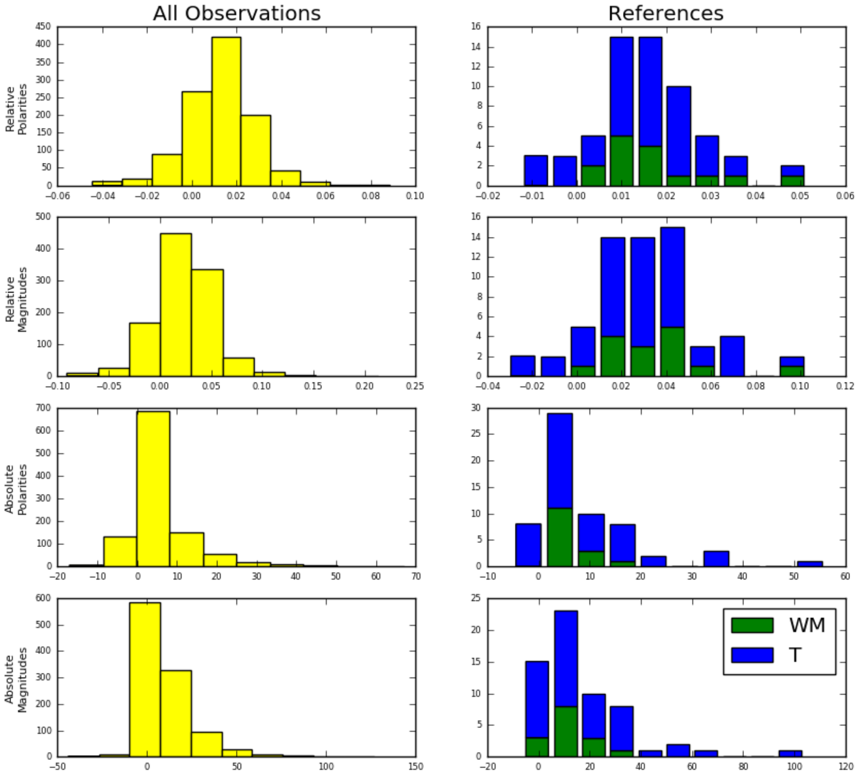


Figure 2 Sentiment Analysis Scores Histograms

The three pairs of Receiving Operator Characteristics (ROC) curves in Figure 3, compared to *Joint*, each represent a specific approach towards ranking the innovations through the two sentiment scores, hence six lines. The methods are: the plain sentiment score of each text, the scores relative to the number of words in each text and the polarity of text relative to the number of words in it. Furthermore, the scores are calculated in ascending order based on these rankings so that all unknowns are treated as negative cases. And since several of the unknowns have higher scores than any of the references the true positive rates are low from the start. The scores of the innovations with multiple articles are the arithmetic mean of those articles' scores.

Quite surprisingly, the best performing of the three approaches is the polarity of the innovations relative to the number of words. Though that is not saying much since it barely performs better than random guessing and the rest are performing worse than random guessing. No matter which level one would choose as the cutoff, there would be a problematic amount of false negatives, viz. innovations from the references that were not selected as *important*. As there is no viable threshold visible in this data, the seemingly best classifier to represent this algorithm is that of having a positive sentiment by both variants, even though it selects a very large portion of the sample.

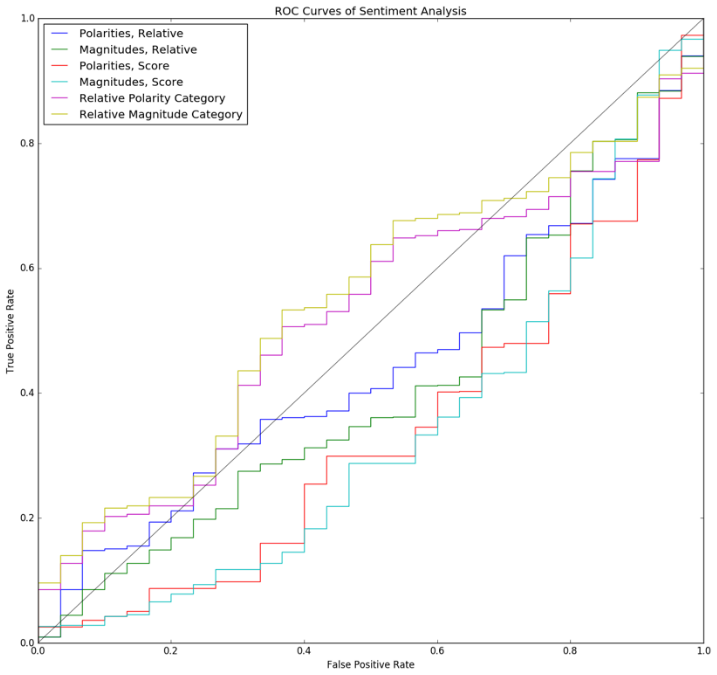


Figure 3 Sentiment Scores ROC

5.3. *k*-means Clustering

A key problem with the *k*-means approach lies in its sensitivity random selection of the initial groupings. This problem can partially be dealt with by allowing for a large number of iterations[17]. A further problem is selecting a value for *k*. Though there is no predetermined standard for choosing a value for *k*, in line with Weiss' suggests approach, choosing *k* based on the minimum cluster variance, the *k* should be selected as 14, as visible in Table 5. In this application 14 is not a good value, since it entails a large number of miniscule clusters and the decrease in variance is so small that it is difficult to imagine that it makes up for this problem. With the goal of avoiding clusters of one, the highest number of clusters that is a viable option is six, though as the goal is not to maximize the number of clusters but to use them to detect *important* innovations, all the lower number of clusters will also be used.

On inspection Table 6, containing the distribution of the articles connected to the reference innovations among the clusters, it appears immediately that they are not concentrated into the larger clusters. Quite the reverse, their concentrations are higher in the smaller, though not smallest, clusters. With this strange distribution none of the models is particularly useful, as is, because one would either have to collect all clusters that contain a reference and remain with almost the entire sample, reject the

[17] For this exercise the algorithm was allowed to run for a maximum of 1000 times and the algorithm was iterate 20000 for each *k*.

smaller clusters though their concentration of references are higher or focus on the higher concentrations to reject most of the references.

Table 5 *k*-means Clustering Statistics

k	Model Variance	Smallest	Median	Largest
2	171.908	206	530.5	855
3	169.141	75	299	687
4	167.935	38	185.5	652
5	167.023	10	121	605
6	166.285	1	56	667
7	165.502	1	52	679
8	164.952	1	40.5	580
9	164.494	1	19	627
10	163.399	1	21.5	599
11	163.229	1	11	558
12	162.963	1	3	576
13	162.546	1	3	413
14	161.605	1	3	478
15	161.624	1	3	616

Table 6 *k*-means Clusters Compared to References

	Group 1	Group 2	Group 3	Group 4	Group 5
k: 2	855	206			
W&M	10	5			
Taalbi	26	20			
Overlap	6	1			
k: 3	687	299	75		
W&M	8	6	1		
Taalbi	21	17	8		
Overlap	4	3	0		
k: 4	652	289	82	38	
W&M	7	6	2	0	
Taalbi	19	14	9	4	
Overlap	3	4	0	0	
k: 5	605	253	121	72	10
W&M	6	5	2	2	0
Taalbi	17	10	11	7	1
Overlap	2	5	0	0	0

As it is clear that no single clustering excludes enough observations to make a justifiable classifier, one could then instead create a binary label to separate observations that are collocated with references from those that are not and rely on the intersection between each *k*'s label. A problem with this approach, in this instance, is that the composition of the larger clusters are fairly similar between outputs, hence few observations are being removed through this intersection. A further problem is that the concentration of references is higher in the smaller clusters, though are then given the same importance as lower concentrations through this intersection. As at least one label from *Taalbi* is present in every group some of them have to be lost in order to reduce the sample. If one disregards *Taalbi*, relying only on *W&M*, still only 38 observations can easily be excluded. Even by the very stringent criterion of relying on *Joint*, a mere 203 observations are excluded, out of which 23 have references. After this inspection it becomes evident that simply assigning a binary classifier based on the collocation of reference articles is rather problematic.

By calculating the share of references per cluster and then calculating the arithmetic mean of the shares that each article is associated with, each article gets a score on a continuous scale between 0 and 1. By further assigning these values to the innovations, by arithmetic means, the innovations can too be directly connected to such a score. This roundabout way is the equivalent of calculating the share of the total number of collocated reference articles in the total number of collocated articles. By applying this for each of the four reference groupings, four different scores for each innovation can be created. Figure 4 below contains the histograms of each of these scores, divided between labeled and unlabeled values. The number above each subplot refers to the number of observations with that particular labeling. The texts above this refers to the source of the grouping, from all 30 reference innovations down to the five overlapping references. From an ocular inspection of the distributions it appears that any one of them could be used to create a classifier by sacrificing a few reference innovations in the process.

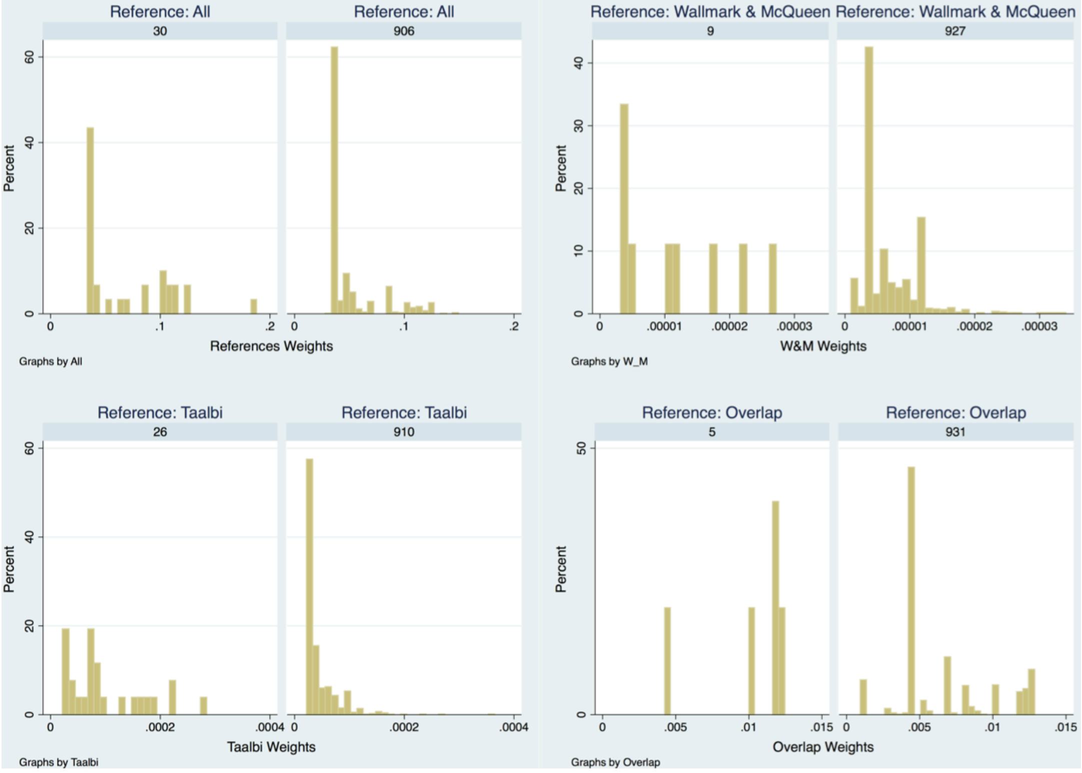


Figure 4 *k*-means Clustering Weights Histogram

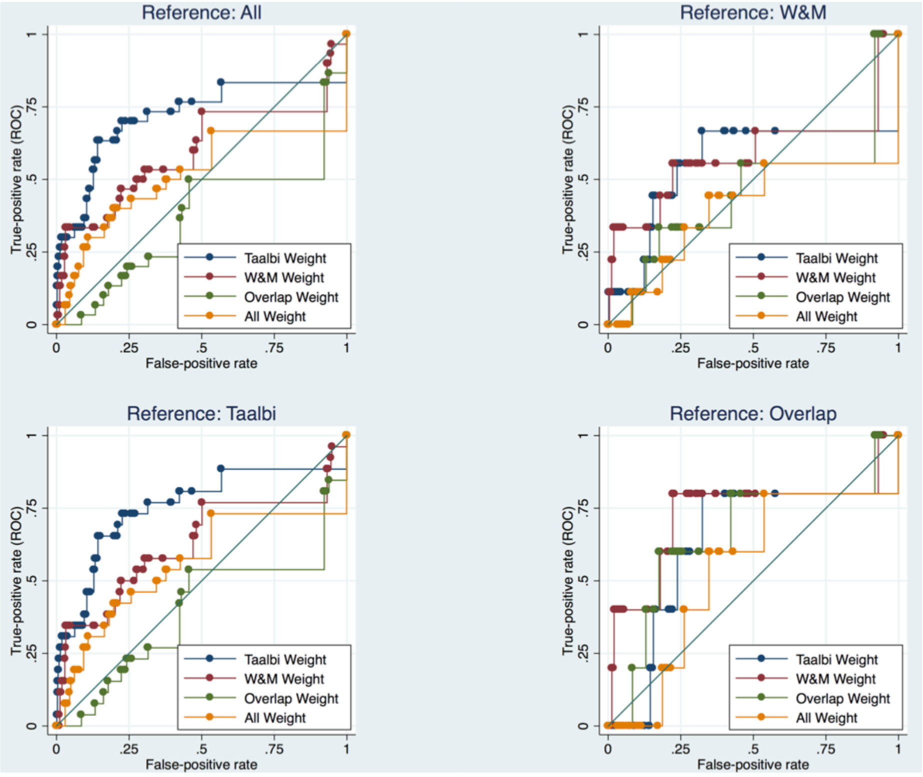


Figure 5 *k*-means Clustering Weights ROC

Figure 5 contains the four ROC graphs, each with a different reference for the ‘true’ labels. While there are visible similarities within each column pair it looks as if the overall best classifier is based on the *Taalbi* weight score. What then remains is to arbitrarily select the threshold for the scores that allows for acceptable levels of true positive rate TPR, false positive rate FPR and number of innovations that can be considered *important*.

In order to avoid false positives, which cannot be tested for, the number of *important* innovations should be kept to a minimum. As to not ignore the reference innovations, the inclusion of both groups should be maximized. Judging by the ROC curves for the *Taalbi* weight, this implies that the FPR should be close to 0.25. The unlabeled and labeled values included in three adjacent thresholds are displayed in Table 7. Out of these the middle threshold contains less than a quarter of the sample innovations, and a majority from either reference group and will therefore be used as the classifier from this method.

Table 7 *k*-means Clustering Thresholds Compared to References

Unknown	*Taalbi*	*Overlap*	*W&M*	*Joint*
190	16	2	4	18
204	16	3	5	18
215	16	3	5	18
936	21	5	9	30

5.4. Hierarchical Clustering

A considerable issue of applying hierarchical clustering on the term document matrix is that there are few guidelines and yet fewer rules to follow in its application. Fewer still are the clear interpretations and conclusions one can reach from applying them. A further obstacle is the fact that clusters of one observation are still technically clusters, although such miniscule clusters are problematic to use. In this exploration it is particularly so as directly defy the reliance on external data for categorizing groups of observations. As the goal to rely on the references to validate the model and select clusters to label as *important*, models with clusters of single observations cannot be relied upon.

On this remark, Table 8 shows that both complete- and average linkages created numerous small clusters, making their divisions very uneven and therefore particularly difficult to base any conclusions on. Since these two linkages agglomerated singular observations in such late stages and most observations are gathered in a single large cluster both linkages are unusable in this instance.

Table 8 Hierarchical Clustering Statistics

Linkage	k	Smallest	Median	Largest	Variance
Compete	2	1	530.5	1060	280370.25
Average	2	1	530.5	1060	280370.25
Ward	2	313	530.5	748	47306.25
Ward	3	59	254	748	84086.89
Ward	4	7	153	748	86335.69
Ward	5	7	52	748	77715.76
Ward	6	7	65.5	748	67172.47
Ward	7	1	52	748	61403.67

In contrast, the *ward* linking created more evenly sized clusters that better facilitate themselves to be compared with the references. From Table 9Table 8 it appears that the ward linkage not only created more even clusters, it also spread the references out among these clusters. It is not particularly strange that the references are almost proportionally distributed between the clusters at the higher levels, though it is problematic that this pattern persists throughout the higher number of clusters. Even more problematic for this approach is the biggest cluster, making up over 70% of the sample is not split even at the lowest level. Quite interestingly, there is a lack of an overlap between the *W&M* and *Taalbi* in four of the six clusters. Due to the even concentration of references between the clusters there is no way of constructing a label without consciously excluding a fair amount of the references or including the vast majority of the sample. As the relationship between observations and clusters vary greatly compared to *k*-means clustering, reapplying the same approach will not yield as clean results, however they are applicable.

Table 9 Ward Compared to References

k=6				*k*=5				*k*=4				*k*=3				*k*=2			
n	*W&M*	T	O	*n*	*W&M*	T	O	n	*W&M*	T	O	n	*W&M*	T	O	n	*W&M*	T	O
748	4	17	4	748	4	17	4	748	4	13	4	748	4	17	4	748	4	17	4
146	1	4	3	225	1	11	3	254	1	14	3	254	1	14	3	313	4	22	3
79	1	7	0																
29	0	3	0	29	0	3	0												
52	3	7	0	52	3	7	0	52	3	7	0	59	3	8	0				
7	0	1	0	7	0	1	0	7	0	1	0								

n: cluster size k: number of clusters T: *Taalbi* O: *Overlap*

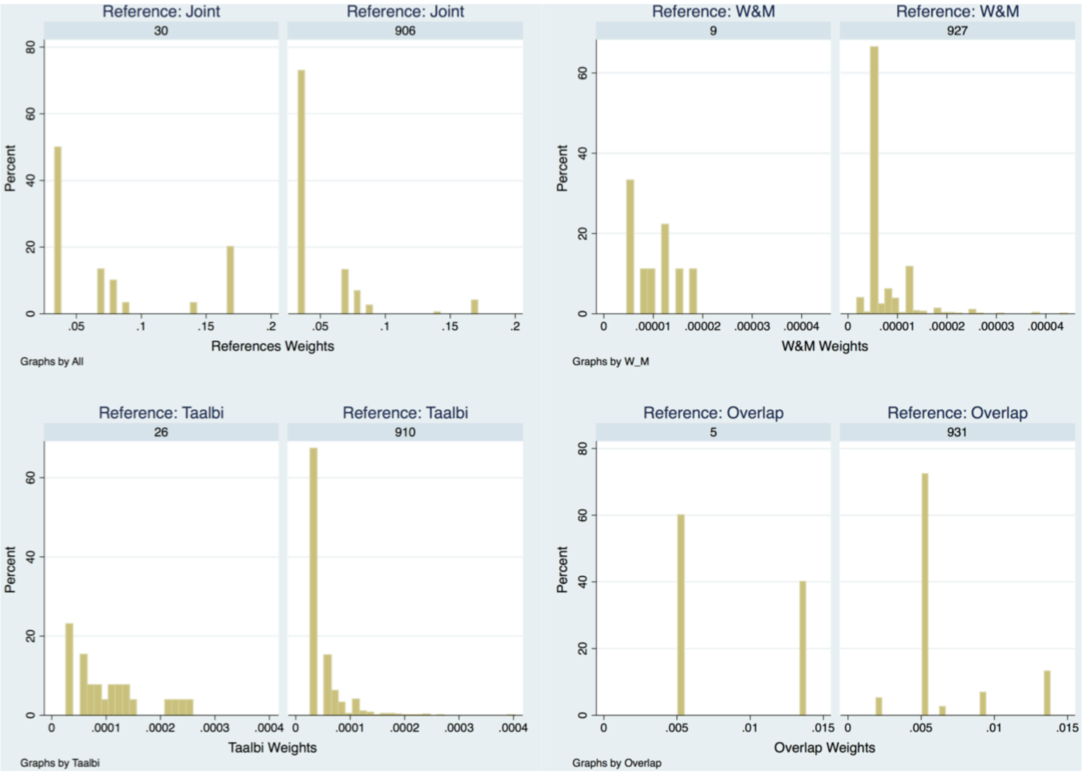


Figure 6 Hierarchical Clustering Weights Histogram

The four scores are calculated in the same way as it was in *k*-means clustering; an innovation's score is the share of collocated reference articles in of all the collocated articles and have their histograms displayed in Figure 6. The influence of the core cluster of 748 observations is visible in each of the histograms and the only group that seems to lend itself to replicating the *k*-means classifier is again based on *Taalbi*. However, the ROC curves in Figure 7 show that the *W&M* weights can possibly be applied as well, again with a FPR around 0.25.

The size and frequencies from the reference groups for the top contenders for thresholds, for both the *Taalbi* and *W&*M weights, are visible in Table 10. Based on the idea to minimize the selection while maximizing the frequencies of the reference groups, the best *Taalbi* threshold is the middle one as the first fails to encapsulate enough references and the latter adds two labeled at the cost of 35 unlabeled innovations. In selection the *W&M* threshold the cost of a larger sample has to be weighed against benefit of including more references, and there is no obvious choice. The additional five observations include two references (40%), the slightly larger group will be relied upon. However, as the goal is to end up with a single classifier per method, the threshold of choice is the one based on *W&M*, as it offers a smaller selection with as many references.

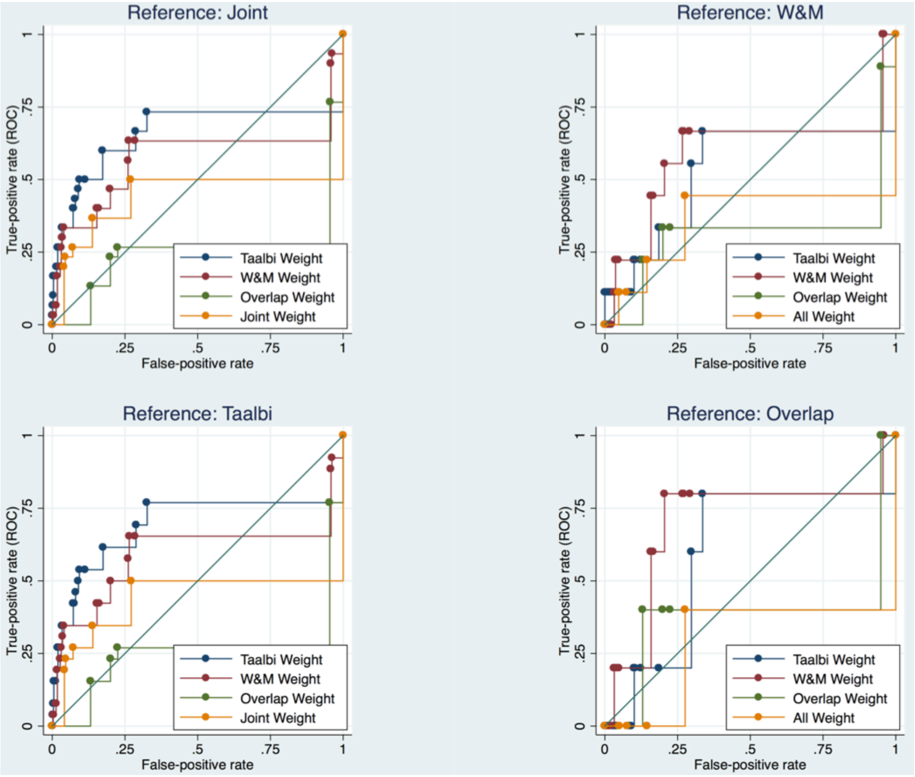


Figure 7 Heirarchical Clustering Weights ROC

Table 10 Hierarchical Clustering Thresholds Compared to References

Weight	Unknown	*Taalbi*	*Overlap*	*W&M*	*Joint*
W&M	181	13	4	5	14
	236	15	4	6	17
	239	17	4	6	19
	868	23	5	9	27
Taalbi	157	16	1	3	18
	260	18	3	5	20
	295	20	4	6	22
Sample	906	26	5	9	30

5.5. Overlap – Selecting the Classifier

A final attempt of trying to assess the models is to compare which innovations the different methods have jointly selected by intersecting them. Based on each of the three classifiers a binary label has been constructed, where 1 means *important*. These labels are compared and intersected in Table 11, with their corresponding counts of innovations. As the sentiment classification is constructed by a very different method it could well have been the case that the small group of unlabeled innovations not chosen by it would have been the ones mainly selected by the other two. However, the alignment is rather high as its intersection with the other classifiers reduces their selection by just over 10%. In contrast, the largest reduction happens at the intersection of these two models, though they maintain a relatively high share of reference innovations despite this.

While the composition of unlabeled and labeled innovations vary between the three elected models, the classifier constructed based on *k*-means clustering is clearly the best choice of the three, as it offers the smallest selection of unlabeled innovations and contains a relatively large amount of reference innovations. As it simultaneously selects a fair amount of the same innovations and reference innovations from the other two classifiers, it is the prime candidate for detecting *important* innovations.

Table 11 Classifier Labels Comparison

Label	Unknown	*Taalbi*	*Overlap*	*W&M*	*Joint*
S	690	22	5	9	26
K	204	19	3	5	21
H	236	15	4	6	17
S*K	183	17	3	5	19
S*H	208	13	3	5	15
K*H	142	14	3	5	16
Sample	906	26	5	9	30

S: Sentiment Analysis **K: *k*-means clustering** **H: Hierarchical Clustering**

5.6. Distributions of *important* innovations

With the created classification, 225 innovations are labeled as *important*. In order for the distribution of *important* innovations among the variables of interest to have a meaning, one also has to consider the overall distribution of the same variables. Out of the five SWINNO variables of interest (industry, artifactual complexity, developmental complexity, novelty to the firm and novelty to the market) four are organically compared in pairs; complexities and novelties.

Overall, there is very little difference between the entire sample's distribution and that of the *important* innovations. The only noticeable difference is that the *important* innovations are less skewed towards low complexities along either variable, visible in Table 12. Similarly the distribution of *important* innovations is noticeably more skewed towards totally new to firm and new to the world market, Table 13 ,though several observations are missing values and this might change when these are found.

Table 12 Innovation Complexities

Developmental:	**n/a**	**High**		**Medium**		**Low**		**Total**	
Artifactual	**Sample**	**Sample**	**Important**	**Sample**	**Important**	**Sample**	**Important**	**Sample**	**Important**
n/a	5 0.53%	4 0.43%	1 0.44%	14 1.50%	4 1.78%	2 0.21%	0 0.00%	**25** **2.67%**	**5** **2.22%**
High	1 0.11%	57 6.09%	24 10.67%	139 14.85%	33 14.67%	1 0.11%	0 0.00%	**198** **21.15%**	**57** **25.33%**
Medium	0 0.00%	27 2.88%	11 4.89%	346 36.97%	86 38.22%	101 10.79%	20 8.89%	**474** **50.64%**	**117** **52.00%**
Low	1 0.11%	5 0.53%	1 0.44%	106 11.32%	31 13.78%	127 13.57%	14 6.22%	**239** **25.53%**	**46** **20.44%**
Total	**7** **0.75%**	**93** **9.94%**	**37** **16.44%**	**605** **64.64%**	**154** **68.44%**	**231** **24.68%**	**34** **15.11%**	**936**	**225**

Table 13 Innovation Novelties

Firm:	**n/a**		**Totally new**		**Major improvement**		**Increment**		**Total**	
Market	**Sample**	**Important**	**Sample**	**Important**	**Sample**	**Important**	**Sample**	**Important**	**Sample**	**Important**
n/a	39 4.17%	2 0.21%	137 14.64%	52 23.11%	421 44.98%	89 39.56%	145 15.49%	9 4.00%	**742** **79.27%**	**152** **67.56%**
New to Swedish market	4 0.43%	0 0.00%	19 2.03%	3 1.33%	31 3.31%	11 4.89%	1 0.11%	0 0.00%	**55** **5.88%**	**14** **6.22%**
New to world market	0 0.00%	0 0.00%	106 11.32%	44 19.56%	32 3.42%	15 6.67%	1 0.11%	0 0.00%	**139** **14.85%**	**59** **26.22%**
Total	**43** **4.59%**	**2** **0.21%**	**262** **27.99%**	**99** **44.00%**	**484** **51.71%**	**115** **51.11%**	**147** **15.71%**	**9** **4.00%**	**936**	**225**

Table 14 Innovation Across Industries

SNI	Industry	Sample		Important	
15+16	Food, beverages & tobacco	20	2.14%	4	1.78%
17+18	Textiles & apparel	6	0.64%		
19	Leather & footwear	2	0.21%		
20	Wood and wood products	13	1.39%	4	1.78%
21	Pulp and paper	14	1.50%	3	1.33%
22	Coke and refined petroleum products	1	0.11%		
23	Printing and publishing	2	0.21%	1	0.44%
24	Chemicals, chemical products and man-made fiber	26	2.78%	9	4.00%
25	Rubber and plastics	61	6.52%	12	5.33%
26	Non-metallic mineral products	17	1.82%	1	0.44%
27	Basic metals	26	2.78%	9	4.00%
28	Fabricated metal products except machinery and equipment	61	6.52%	12	5.33%
29	Machinery and equipment	339	36.22%	81	36.00%
30	Office machinery and computers	49	5.24%	17	7.56%
31	Electrical machinery and apparatus	55	5.88%	12	5.33%
32	Radio, television, and communication equipment and apparatuses	40	4.27%	7	3.11%
33	Medical, precision and optical instruments, watches and clocks	107	11.43%	31	13.78%
34	Motor vehicles, trailers and semi-trailers	36	3.85%	5	2.22%
35	Other transport equipment	17	1.82%	6	2.67%
36	Other manufacturing	11	1.18%	1	0.44%
72	Computer and related activities (software)	4	0.43%	1	0.44%
74	Technical consultancy and testing	19	2.03%	8	3.56%

At a quick glance of Table 14 there are, seemingly, few interesting observations to be made on the industry distributions as it generally follows a similar distribution across the groups. And the few industries with a higher concentration of *important* innovations have very few observations in the sample viz. the distribution is very close the expected values from a random selection. Furthermore, there is no indication that the classifier excluded services, which is not as encouraging as their exclusion would have been discouraging.

6. Discussion

The present study investigates the viability of augmenting innovation theory with text mining. This is done by applying three naïve algorithms to studdy patterns in articles from trade journals, used in the LBIO based SWINNO database, in order to create a classifier that detected several previously overlooked *important* innovations.

In the pure output of each model the tendency was to find the reference data grouped with large groups of unknown outcomes. Selecting some of these groups as *important* is not a viable option as the references were spread out and large selections remained after the different groups were intersected. Though one cannot say for certain that these selections were wrong, as the 'true' outcome is unknown, having a large portion of unknowns is problematic to interpret. As the selection of innovations that enter the SWINNO database is akin to the tip of the iceberg of innovation activity (Sjöö, 2014), it may simply be that most of them are, in fact, *important*. Though Occam's razor would suggest that this simple approach does not work as a classifier.

There are several plausible explanations for why the pure models produced results that were unusable as classifiers. Firstly, the errors introduced by the OCR could very well have led to the removal of terms that could have led to different cluster compositions. Secondly, as the reduction of the dictionary (stemming to a root, inflectional stemming and removal of stopwords) was performed manually by sifting through over 14,000 stems, it is possible that something was excluded or joined incorrectly, though they were checked multiple times in order to prevent this. Thirdly, as the bag of words approach does not capture syntax it also fails to capture the assumed presence of the journalists' exuberance. Finally, the algorithms were not originally created with the purpose to construct classifiers per se; they simply generate groups based on mathematical similarities.

Even by relying on methodologically simplistic methods to create continuous variables based on the cluster, a classifier could be constructed that managed to capture 70% of the innovations from Wallmark and McQueen (1991) and Taalbi (2016), adding 204 previously undetected *important* innovations. As this study is, to the author's knowledge, the first classifier of this kind, there is nothing to compare this performance to.

While there are several potential problems with this way of constructing a classifier: from the untested assumption that *importance* is reflected in the texts to the way of selecting the model to base the classifier on, the biggest issue has not yet been examined: reproducibility. As there is a random element in the k-means approach there is no guarantee that the same results can be reached again. By allowing for a large number of iterations this influence has been reduced, though it cannot be completely eliminated without running the algorithm based on every possible starting position. This random element is further reduced by relying on several levels of clusters in the creation of the final classifier.

Though in the very brief space dedicated to some of the descriptive characteristics of *important* innovations did not reveal much, some expectable features were found;

the *important* innovations are more skewed towards higher complexities and radical novelty, relative to the entire sample. However, it is a bit surprising to see that the *important* innovations are almost perfectly following the industry distribution of the entire sample, with a few industries with relatively few innovations generating much more than their share. This overall raises some interesting questions regarding the propensity to generate *important* innovations in different industries.

As the model outcomes are not verifiable,as of yet, the evaluation of their reliability will have to wait. Still, by capturing 70% of the reference innovations while including only 24.01% of the entire sample used, it performs quite well. As the majority of these innovations were also picked out by the other approaches, coupled with the skewness towards higher complexities and novelties among the *important* innovations, the results are well within what can reasonably be expected of the classifier.

There are two key directions that future research could take based on what has been found in this study. Firstly, the descriptive patterns observed in the *important* subset could be explored further, looking at more variables and attempt to explain the distribution across industries, as this is rather peculiar.

The other direction is to dig deeper into the toolbox of text mining. Though I maintain that I have shown some of the benefits of intersecting the fields, the problem of finding *important* innovations in general is not solved. There is a range of techniques available that could be applied at various stages in the process, potentially amending the results. Futhermore, as this study focused exclusively on articles from the 1970s, there are four more decades of article available in SWINNO alone.

Further techniques are available that could potentially be applied to the LBIO method, from the selection of relevant articles to extracting the data. Though I cannot say for sure that this is going to happen, I hope that this paper is an early step in that direction as this would allow for a broader data collection.

References

Primary Sources

Basberg, B. (1987). Patents and the Measurement of Technological Change: A Survey of the Literature. *Research Policy, 16*, 131-141.

Basu, S., & Davidson, I. (2009). Constrained Partitional Clustering of Text Data: An Overview. In A. N. Srivastava & M. Sahami (Eds.), *Text mining. [Elektronisk resurs] : classification, clustering, and applications* (pp. 155-212). Ich: Boca Raton : CRC Press, c2009.

Benamara, F., Cesarano, C., Picariello, A., & Subhamanian, V. (2005). Sentiment Analysis: Adjectives and Adverbs are Better Than Adjectives Alone. Retrieved from http://oasys.umiacs.umd.edu/oasysnew/papers/icwsmV2.pdf

Beneito, P. (2006). The innovative performance of in-house and contracted R&D in terms of patents and utility models. *Research Policy, 35*, 502-517. doi:10.1016/j.respol.2006.01.007

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*: O'Reilly Media Inc.

Bower, J. L., & Christensen, C. M. (1995). DISRUPTIVE TECHNOLOGIES (Vol. 73, pp. 172-172): Harvard Business School Publication Corp.

Coombs, R., Narandren, P., & Richards, A. (1996). A literature-based innovation output indicator. *Research Policy, 25*, 403-413. doi:10.1016/0048-7333(95)00842-X

Croft, W. B., Harding, S. M., Taghva, K., & Borsack, J. (1994). An Evaluation of Information Retrieval Accuracy with Simulated OCR Output. *Symposium on Document Analysis and Information Retrieval.*

Dadoun, M., & Olsson, D. (2016). *Sentiment Classification Techniques Applied to Swedish Tweets Investigating the Effects of translation on Sentiments from Swedish into English.* KTH, KTH Royal Institute of Technology. Retrieved from http://kth.diva-portal.org/smash/get/diva2:926472/FULLTEXT01.pdf

Fagerberg, J. (2005). Innovation: A Guide to the Literature. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 1-26): Oxford : Oxford University Press, 2005.

Fagerberg, J., Mowery, D. C., & Nelson, R. R. (2005). *The Oxford handbook of innovation*: Oxford : Oxford University Press, 2005.

Gerben van der, P. (2007). Issues in measuring innovation. *Scientometrics, 71*(3), 495-507.

Granstrand, O. (2005). Innovation and Intellectual Property Rights. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 266-290): Oxford : Oxford University Press, 2005.

Greve, H. R. (2007). Exploration and exploitation in product innovation. *Industrial & Corporate Change, 16*(5), 945-975.

Groot, B. d., & Franses, P. H. (2005). Cycles in basic innovation. *Technological Forecasting and Social Change, 76*(8), 1021-1025.

Gustavsson, A. (2016). sentiment-swedish

. Retrieved from https://github.com/AlexGustafsson/sentiment-swedish

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: is there an advantage in using multiple indicators. *Research Policy, 32*(8), 1365-1379. doi:10.1016/S0048-7333(02)00137-3

Hall, B. H. (2005). Innovation and Diffusion. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 459-484): Oxford : Oxford University Press, 2005.

Henderson, R. M., & Clark, K. B. (1990). Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms*,* 9.

Hennig, C. (2014). How Many Bee Species? A Case Study in Determining the Number of Clusters. In M. Spiliopoulou, L. Schmidt-Thieme, & R. Janning (Eds.), *Data Analysis, Machine Learning and Knowledge Discovery* (pp. 129-150): Springer.

Hippel, E. v. (1988). *The sources of innovation*: New York : Oxford U.P., 1988.

Hong, S. (2013). *Innovation in New Zealand: A Firm-Level Analysis.* (Doctor of Philosophy Dissertation), University of Canterbury. Retrieved from http://hdl.handle.net/10092/7659

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning. [Elektronisk resurs] : with Applications in R*: New York, NY : Springer New York : Imprint: Springer, 2013.

Jaruzelski, B., Staack, V., & Goehle, B. (2014). The Global Innovation 1000: Proven Paths to Innovation Success. *Strategy and Business*(77), 2-16.

Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). THE NON-TRIVIAL CHOICE BETWEEN INNOVATION INDICATORS. *Economics of Innovation & New Technology, 11*(2), 109.

Kuznets, S. (1973). Modern Economic Growth: Findings and Reflections. *American Economic Review, 63*(3), 247-258.

Lazonick, W. (2005). The Innovative Firm. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 29-55): Oxford : Oxford University Press, 2005.

Lerner, J. (2002). patent Protection and Innovation Ove 150 Years. *NBER Working Papers*. Retrieved from http://www.nber.org/papers/w8977 doi:10.3386/w8977

Link, A. N. (1995). The Use of Literature-Based Innovation Output Indicators for Research Evaluation, 451.

Lipsey, R. G., Carlaw, K., & Bekar, C. (2005). *Economic transformations : general purpose technologies and long term economic growth*: New York ; Oxford : Oxford University Press, 2005.

Ljungberg, J. (2004). Technology and Human Capital in Historical Perspective: An Introduction. In J.-P. Smits & J. Ljungberg (Eds.), *Technology and Human Capital in Historical Perspective. [Elektronisk resurs]* (pp. 1-21). Basingstike: Palgrave Macmillan Ltd. 2004.

Lundvall, B.-Å., & Borrás, S. (2005). Science, Technology, and Innovation Policy. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 599-631): Oxford : Oxford University Press, 2005.

Maclurin, W. R. (1953). The Sequence From Invention to Innovation and its Relation to Economic Growth.

Malerba, F. (2005). SectoralSystems of Innovation. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 380-406): Oxford : Oxford University Press, 2005.

March, J. G. (1991). Exploration and Exploitation in Organizational Learning, 71.

Mowery, K. B. D. C. (2005). Innovation Through Time. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 349-379): Oxford : Oxford University Press, 2005.

Nagy, D., Schuessler, J., & Dubinsky, A. (2015). Defining and identifying disruptive innovations. *Industrial Marketing Management*. doi:10.1016/j.indmarman.2015.11.017

Nelson, A. J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy, 38*, 994-1005. doi:10.1016/j.respol.2009.01.023

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *eprint arXiv:1103.2903*. Retrieved from

Oecd/Eurostat. (2005). Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data (3rd Edition ed.). Paris: OECD Publishing.

Pavitt, K. (2005). Innovation Process. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 86-114): Oxford : Oxford University Press, 2005.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Schön, L. (2009). Technological Waves and Economic Growth - Sweden in an International Perspective 1850-2005. *Circle Electronic Working Paper series*.

Schumpeter, J. A. (1939). *Business Cycles. A Theoretical, Historical and Statistical Analysis of the Capitalist Process.* (Vol. 1). New York: McGraw-Hill Book Company Inc.

Silverberg, G., & Verspagen, B. (2003). Breaking the waves: a Poisson regression approach to Schumpeterian clustering of basic innovations, 671.

Sjöö, K. (2014). *Innovation and transformation in the Swedish manufacturing sector, 1970-2007*: Lund : Department of Economic History, School of Economics and Management, Lund University, 2014 (Lund : Media-tryck).

Sjöö, K., Taalbi, J., Kander, A., & Ljungberg, J. (2014). SWINNO: A Database of Swedish Innovations, 1970-2007. *Lund Papers in Economic History*. Retrieved from

Smith, K. (2005). Measuring Innovation. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 148-177): Oxford : Oxford University Press, 2005.

Srivastava, A. N., & Sahami, M. (2009). *Text mining. [Elektronisk resurs] : classification, clustering, and applications*: Boca Raton : CRC Press, c2009.

Swann, G. M. P. (2009). *The economics of innovation : an introduction*: Cheltenham : Edward Elgar, 2009.

Taalbi, J. (2014). *Innovation as creative reponse: determinants of innovation in the Swedish manufacturing industry, 1970-2007*: Lund : School of Economics and Management, Department of Economic History, Lund University, 2014.

Taalbi, J. (2016). *Major Swedish Innovations*. [data file] unpulished dataset, cited with permission.

Verspagen, B. (2005). Innovation and Economic Growth. In J. Fagerberg, D. C. Mowery, & R. R. Nelson (Eds.), *The Oxford handbook of innovation* (pp. 487-513): Oxford : Oxford University Press, 2005.

Walker, R. M., Jeanes, E., & Rowlands, R. (2002). Measuring Innovation – Applying the Literature-Based Innovation Output Indicator to Public Services. *Public Administration, 80*(1), 201.

Wallmark, J. T., & McQueen, D. H. (1991). One hundred major Swedish technical innovations, from 1945 to 1980. *Research Policy, 20*, 325-344. doi:10.1016/0048-7333(91)90093-6

Weiss, S. M. (2005). *Text mining. [Elektronisk resurs] : predictive methods for analyzing unstructured information*: New York : Springer, 2005.

Zacharski, R. (2015). *A Programmer's Guide to Data Mining: The Ancient Art of the Numerati* (pp. 395). Retrieved from http://guidetodatamining.com/assets/guideChapters/Guide2DataMining.pdf

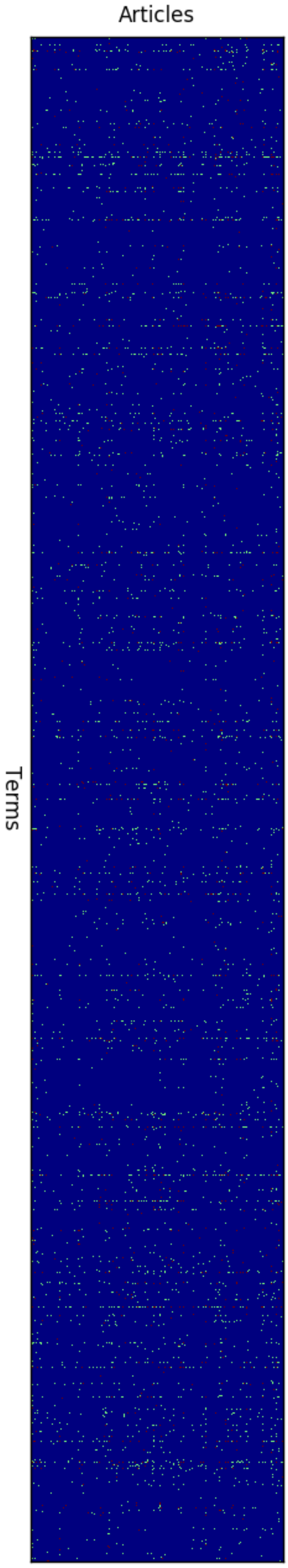
Zhai, C., & Aggarwal, C. C. (2012). *Mining Text Data*: Dordrecht, Netherlands : Springer.

Secondary Sources

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *Proceedings of the 1008 International Conference on Web Search and Data Minig*, 231-240.

Appendix A – The Matrix

The values are encoded into color where ones and twos are represented by green and red squares respectively.



Appendix B

Reference Innovations Missed by Sentiment Analysis

Innovation	Year of Commercialization	Type	Journal	Year
Inpro-metoden	1977	3	Livsmedel I Fokus	1977
Nucon	1972	1	Ny Teknik	1972
MHU (Material Hardening Unit)	1971	1	Ny Teknik	1971
	1971	1	Verkstäderna	1972