February 2017

SSIX Big Data Technologies and Methods for Leveraging Social Sentiment Data in Multiple Business Domains

Keith Cortis [a], Waqas Khawaja [b], Ross McDermott [b], Laurentiu Vasiliu [c], Adamantios Koumpis [a], Siegfried Handschuh [a] and Brian Davis [b]

[a] *Universität Passau, Passau, Germany*
[b] *INSIGHT Centre for Data Analytics, NUI Galway, Ireland*
[c] *Peracton, Dublin, Ireland*

Abstract. Social Sentiment Indices powered by X-Scores (SSIX) aims to provide European SMEs with a collection of easy to interpret tools to analyse and understand social media users' attitudes for any given topic. These sentiment characteristics can be exploited to help SMEs operate more efficiently resulting in increased revenues. Social media data represents a combined measure of thoughts and views touching every area of life. SSIX will search and index conversations taking place on social network services, such as Twitter, StockTwits and Facebook, together with the most reliable and trustworthy news agencies, newspapers, blogs and industry publications. A statistical framework of qualitative and quantitative parameters called X-Scores will power SSIX. Classification and scoring of content will be done using this framework, regardless of language, locale or data architecture. The X-Scores framework will interpret economically significant sentiment signals in social media conversations producing sentiment metrics, such as momentum, breadth, topic frequency, volatility and historical comparison. These metrics will create commercially viable social sentiment indexes, which can be tailored to any domain of interest. By enabling European SMEs to analyse and leverage social sentiment in their discipline, SSIX will facilitate the creation of innovative products and services by enhancing the investment decision making process, thus assisting in generating increased revenue while also minimising risk exposure.

Keywords. sentiment analysis, social sentiment index, social media analytics, cross-lingual, social sentiment data, big social data, social media data, news data, data science,

1. Introduction

The emerging use of social media data as part of the investment process has seen a rapid increase in uptake in recent years, as by examined Greenfield [14]. The lag between Social Media Monitoring and Social Media Analytics - "Brand Analytics" and Finance specific analytics applications has narrowed. The Social Finance Analytics sector has built on the base developed by Brand Analytics and has evolved the ecosystem to focus on investment decision-making. The growth of trading specific social networks like StockTwits has also provided highly valuable structured social data on trading discussions, which was not accessible previously on general social media communities. This new data source has provided a vital pipeline of thoughts, words and decisions between people; connecting and interacting as never before. This collective pulse of conversations and emotional attitudes acts as a gauge of opinions and ideas on every aspect of society. Finance specific social media applications provide asset managers, equity analysts and high frequency traders with the ability to research and evaluate subtle real-time signals, such as sentiment volatility changes, discovery of breaking news and macroeconomic trend analysis. These data streams can be incorporated into current operating models as additional attributes for executing investment decision-making, with a goal to increase alpha and manage risk for a portfolio.

The European research project Social Sentiment Indices powered by X-Scores (SSIX) [1], seeks to assist in this challenge of incorporating relevant and valuable social media sentiment data into investment decision making by enabling X-Scores metrics and SSIX indices to act as valid indicators that will help produce increased growth for European Small and Medium-sized Enterprises (SMEs). X-Scores are time series metrics derived from Natural language processing (NLP) categorised by metadata and smart data. The different time series data streams mostly consist of volume and oscillator indicators; an example would be AAPL-VX, which is Volatility of sentiment X-Score for Apple Inc. SSIX will extract meaningful financial signals in a cross-lingual fashion from a multitude of social network sources, such as Twitter, Facebook, StockTwits and LinkedIn, and also authoritative news sources, such as Newswires, Bloomberg, Financial Times and CNBC news channel; transforming these signals into clearly quantifiable sentiment metrics and indices regardless of language or locale. Financial services' SMEs can customise SSIX indices enabling them to provide meaningful domain specific insight to design more efficient systems, test trading and investment strategies, better understand risk and volatility behaviour of social sentiment and identifying new investment opportunities.

SMEs can exploit the open source SSIX tools and methodologies to provide financial analytics services or alternatively resell custom SSIX Indices as valuable financial data products to third parties, thus leading to growth and increased revenue for SSIX industry partners within the consortium and beyond. Beyond the financial application, the SSIX approach and methodologies can have broader impact across geopolitical and socio-economic domains, generating multifaceted and multi-domain sentiment index data for commercial exploitation.

The objectives of the SSIX project are to:

1. **Develop the "X-Scores" statistical framework**, which will analyse metadata from indexed textual sources to capture the signature of social sentiment, gener-

[1] http://ssix-project.eu/

ating a sentiment score. Statistical methods will include regression, covariance and correlation analysis. These X-scores will be used to create the custom SSIX Indices.

2. **Create an open-source template for generating custom SSIX indices** that can be tailor-made with domain specific data parameters for specific analysis objectives, such as Economics, Trading, Investing, Government, Environmental or Risk profiling.
3. **Create a powerful, easy to implement and low latency "X-Scores API"** to distribute the raw sentiment data feed and/or custom SSIX Indices that will allow end users to easily integrate SSIXs sentiment data into their own systems.
4. **Enable end users to do cross-lingual target and aspect oriented sentiment analysis** over any significant social network using user defined dedicated SSIX Index.
5. **Enable various public/private organisations and institutions to create a SSIX Index** and integrate them with their proprietary tools in an easy to use manner.
6. **Explore the domain of SSIX Indices and X-Scores beyond its primary focus of Finance applications.** Research has shown there is a positive correlation between social media sentiment and a financial securities performance, but it is more difficult to measure a broad topic such as, welfare of a region or community. X-Scores will seek to provide metrics which can filter out the noise and provide real quantifiable data, which can give insight via a custom SSIX Index into domains diverse as Education (SSIX-EDU), Media trust (SSIX-MEDIA), Economic sociology (SSIX-ECOSOC), Security (SSIX-SEC) and Health (SSIX-HLTH).
7. **Empower and equip SMEs within the emerging Big Data Financial News sector** to better compete with established industry players via technology transfer involving stable, mature and scalable open source semantic and content analysis technologies.
8. **Trigger, nurture and maintain a SSIX and X-Scores commercial ecosystem within and beyond the project lifecycle.**
9. **Pierce language barriers with respect to untapped and siloed multilingual financial sentiment** content by harvesting cross-lingual Big Social Media and News Data.

By number crunching news text and social networks data feeds regarding a company, product or various financial products (such as, stocks, funds, exchange-traded funds (ETFs), bonds etc.) in a mathematical and statistical way, our approach will allow investors and traders to combine SSIX generated indices with their own proprietary tools and methodologies. We envisage empowering the end-user, such as financial data providers, financial, institutions, investment banks, wealth management houses, asset management professionals, online brokers, professional traders and individual investors with the ability to make more informed and better and safer financial decisions. Finally, SSIX could help in identifying unwanted or dangerous trends that could be signalled to financial regulators in advance in order to take appropriate measures, potentially preventing unhealthy and toxic trading behaviour, thereby safeguarding economic growth and prosperity.

The remainder of the paper[2] discusses related work in Section 2 and the SSIX Architecture in Section 3. Information about SSIX Templates and Scenarios is provided in Section 4, whereas Section 5 discusses Big Social and News Data Management for SSIX. Details about Natural Language Processing Services and Analysis is presented in Section 6. Business Case Studies for SSIX are discussed in Section 7, before providing some concluding remarks in Section 8.

[2]This work is an update and extension of the paper titled "Social Sentiment Indices Powered by X-scores" presented at the ALLDATA 2016 conference [12]

2. Related Work

2.1. Sentiment Analysis on Financial Indices

In [11], Bormann defines several psychological definitions about feelings, in order to explain what might be meant by "market sentiment" in literature on sentiment indices. This study is very relevant to SSIX, since it relates short and long term sentiment indices to two distinct parts of sentiments, namely emotion and mood; and extracts two factors representing investor emotion and mood across all markets in the dataset.

The FIRST project [2] provides sentiment extraction and analysis of market participants from social media networks in near real-time. This is very valuable towards detecting and predicting financial market events. This project is relevant to SSIX, since the tool consists of a decision support model based on Web sentiment as found within textual data extracted from Twitter or blogs, for the financial domain. The relationship between sentiment and trading volume can provide the end-user with important insights about financial market movements. It can also detect financial market abuse, e.g., price manipulation of financial instruments from disinformation. Unlike SSIX, only social networking services are used for extracting and analysing sentiment, whereas the developed tool cannot be easily customised to support media sources, target specific companies or select the required language. In this respect, SSIX provides a template methodology and source code to create in a consistent manner the sentiment index for any type of financial product and financial derivatives. Also the outcome is easily integrated within other analytics tools as a data stream with values between 0 and 100 that will define the ranges of that specific sentiment.

Mirowski et al. [20] presents an algorithm for topic modelling, text classification and retrieval from time-stamped documents. It is trained on each stage of its non-linear multilayer model in order to produce increasingly more compact representations of bags-of-words at a document or paragraph level, hence performing a semantic analysis. This algorithm has been applied to predict the stock market volatility using financial news from Bloomberg. The volatility considered is estimated from daily stock prices of a particular company. On a similar level, in [19] the authors present StockWatcher through a customised, aggregated view of news categorised by different topics. StockWatcher performs sentiment analysis on a particular news messages. Each message can have either a positive, negative or neutral effect on the company. This tool enables the extraction of relevant news items from RSS feeds concerning the NASDAQ-100 listed companies. The sentiment of the news messages directly affects a company's respective stock price. SSIX, will extract meaningful financial signals from multilingual heterogeneous (microblogging and conventional) content sources and not just news items.

Gloor et al. introduces a novel set of social network analysis based algorithms for mining unstructured information from the Web to identify trends and the people launching them [13]. This work is relevant, since the result of a three-step process produces a "Web buzz index" for a specific concept that allows for an outlook on how the popularity of the concept might develop in the future. A possible application of this system might be for financial regulators who try to identify micro- and macro-trends in financial markets, e.g., showing the correlation between fluctuations in the Web buzz index for stock titles and stock prices. Similarly, the Financial Semantic Index estimates the probability that on a particular day, an article in the financial press expresses a positive

attitude towards financial markets. This is measured through the emotional tone of the mentioned article [21]. It is relevant to SSIX, since it provides a certain viewpoint of the media environment the market participants consume. In the case of SSIX, it targets to transform the extracted information into multiple clearly quantifiable social financial sentiment indices regardless of language and data format. This will improve the trading and investment accuracy through the combination of various fundamental and technical parameters together with sentiment ones.

2.2. Cross-lingual mining of information

The MONNET project provides a semantics-based solution for integrated information access amongst language barriers [3]. MONNET is relevant for SSIX, since one of its major innovations is the provision of cross-lingual ontology-based information extraction techniques for semantic-level extraction of information for text and (semi) structured data across languages by using multilingual localised ontologies. It provides real-life applications that demonstrate the exploitation potential in several areas, such as financial services. In fact, one of the project's use-cases deals with searching and querying for financial information in the user's language of choice. On the other hand, it focused on cross-lingual domain, thus failed to target other important aspects, e.g., mining the extracted information. SSIX will help identify unwanted/dangerous trends that could be signalled to financial regulators in advance, in order to potentially prevent unhealthy trading behaviour. Hence, SSIX indices can be used as 'early warning' signals for traders, investors and regulator agencies, such as European Central Bank, EU states national banks and rating agencies.

TrendMiner, another European project [7], presents an innovative and portable open-source real-time method for cross-lingual mining and summarisation of large-scale social media streams, such as weblogs, Twitter, Facebook, etc. One high profile case study was a financial decision support (with analysts, traders, regulators and economists). In terms of novelty, a weakly supervised machine learning algorithm is utilised for automatic discovery of new trends and correlations, whereas a cloud-based infrastructure is used for real-time text mining from stream media. This project is relevant to SSIX given that it provides several multilingual ontology-basedsentiment extraction methods.

The main goal of the LIDER project [8] is to create a Linguistic Linked Data (LLD) cloud that is able to support content analytics tasks of unstructured multilingual cross-media content. This will help in providing an ecosystem for a new Linked Open Data based ecosystem of free, interlinked and semantically interoperable language resources (e.g., corpora, dictionaries, etc.) and media resources (e.g., image, video, etc.). It also aims to make an impact on the ease and efficiency with which LLD is exploited in processes related to content analysis with several use cases in multiple industries within the areas of social media, financial services and other multimedia content providers and consumers. One limitation is that LIDER aims to make an impact on the LOD cloud and not to further transform any extracted signals into clearly quantifiable social sentiment indices, as in the case of SSIX. Such indices are targeted to any equities, stock indices or derivatives.

The AnnoMarket project has delivered a cloud-based platform for unstructured data analytics services, in multiple languages [4]. This text annotation market is delivered via annomarket.com and has been in public beta as of April 2014. The services being offered

can be adopted and applied for many business applications, e.g., large-volume multilingual information management, business intelligence, social media monitoring, customer relations management. It includes several text analytics services that would be of benefit to the SSIX project. Similarly, OpeNER will provide a number of ready to use tools in order to perform some NLP tasks (entity mentions detection and disambiguation, sentiment analysis and opinion detection) that can be freely and easily integrated in the workflow of SMEs [6]. This project aims to have a semi-automatic generation of generic multilingual (initially for the English, French, German, Dutch, Italian and Spanish languages) sentiment lexicons with cultural normalisation and scales through the reuse of existing language resources. FREME, an "Open Framework for e-services for multilingual and semantic enrichment", is a project designed to bring understanding to multilingual digital content by addressing the whole content value chain from content creation to publishing [9]. Another project - MixedEmotions, develops innovative multilingual multi-modal Big Data analytics applications for emotion analysis across multilingual text data sources, audio/video signal input, social media and structured data [10]. SSIX goes beyond text analysis on unstructured data, since an "X-Scores" statistical framework will be implemented to capture the signature of social sentiment from indexed textual sources. These scores will help create custom SSIX Indices that can be tailored for a particular domain depending on specific data parameters. This will provide a meaningful insight to drive trading, investment decisions and strategies, and create new investment opportunities.

3. SSIX Architecture

3.1. Lambda Architecture

The Lambda architecture essentially has three layers named as the batch, serving and speed layers. Figure 1 provides an overview diagram of this architecture and the following is a description of the basic components [1].

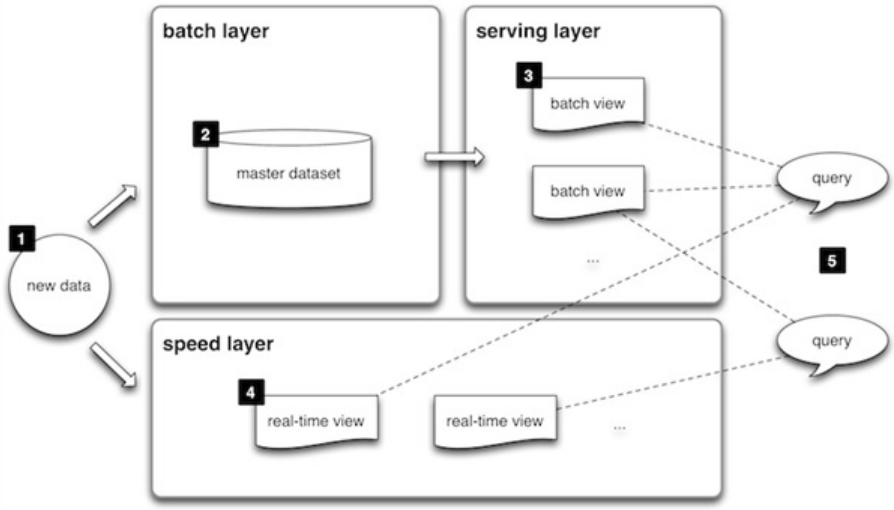

Figure 1. Overview of the Lambda Architecture

- **New Data**: All data entering the system is dispatched to both the batch layer and the speed layer for processing.
- **Batch layer**: This layer has two functions: (i) managing the master dataset, an immutable, append-only set of raw data, and (ii) to pre-compute arbitrary query functions, called batch views.
- **Serving layer**: This layer indexes the batch views so that they can be queried ad-hoc with low latency.
- **Speed layer**: This layer compensates for the high latency of updates to the serving layer, due to the batch layer. Using fast and incremental algorithms, the speed layer deals with recent data only.
- **Queries**: Last but not least, any incoming query can be answered by merging results from batch views and real-time views.

The Lambda architecture has been built to provide robustness, fault tolerance, scalability, generalisation, extensibility and ad-hoc queries [18]. The presence of two layers allows maintaining real-time processing as well as the ability to process large amounts of data in batches. However, this adds complexity as there is a need to maintain two separate code bases one for each layer. One solution is to use the same processing framework e.g., Apache Spark[3] in both layers to eliminate the need for separate code bases. In addition

[3]http://spark.apache.org/

to managing the layers, there are challenges in displaying merged results from the output of both layers as well.

3.2. Overview

The Lambda Architecture was chosen as the basis for the SSIX architecture, mainly because of its excellent suitability for Big Data systems. The Lambda Architecture advocates storage of complete raw data that shields against human error. Storage of immutable raw data and the ability to reprocess it is particularly useful for a research and development project when the exact scope is not clear in the beginning and processing algorithms are very likely to change. These changes can be reapplied to raw data completely.

SSIX aims to provide a one minute feed that can be managed by using speed layer and moving heavier tasks to batch layer. This is in addition to other uses of batch layer like reprocessing raw data in cases such as back-testing of SSIX indexes. The SSIX platform is being developed around financial domain but should be applicable to other domains such as education, politics, product reviews. Lambda architecture is highly generalisable, hence its systems can be adapted to different domains and datasets. SSIX also borrows from some concepts of the Kappa Architecture[16] –designed to address some of the complexity problems associated with the Lambda architecture–, such as storing data in Apache Kafka[4] and maintaining the same layer for speed and batch processing for easier implementation and maintainability reasons which are highlighted in detail below.

The SSIX Platform is divided in four main tiers:

1. Ingestion
2. Processing
3. Platform
4. Storage

These tiers are explained in detail in Section 3.4. The ingestion tier performs its operations of data collection, filtering and refinement and data is put to Apache Kafka which is a distributed publish subscribe messaging service. SSIX currently stores its raw data in persistent storage outside of the batch layer but the Kafka topic can be expanded to store raw data to an agreed point. This takes away the ability to reprocess all of the raw data but gives the advantage of simpler implementation which is promoted by the Kappa architecture. SSIX architecture is flexible to allow for storing of complete raw data if it is dependent on the user case requirements.

SSIX uses Apache Spark as its distributed computation engine. SSIX leverages the flexibility of Lambda architecture to maintain same code bases for its speed and batch layers where the execution of jobs is controlled by a custom layer manager that maintains the workload of both these layers. The batch layers can also be invoked by query manager to satisfy certain ad-hoc queries or where reprocessing of the data is required. Again, the reprocessing is dependant on the timestamp up to which the raw data is available. A change in code base or processing logic may also trigger reprocessing of the raw data through batch layer. Both layers have Natural Language Processing and X-Scores calculation engines running within them. An overview of inner working of these components is provided in Section 3.4 below. The results from both speed and batch layers

[4] Apache Kafka is an open-source message broker, it provide a unified, high-throughput, low-latency platform for handling real-time data feeds - http://kafka.apache.org/

are stored in ElasticSearch[5]. These results are then used by the query manager to satisfy API requests. In addition to the core components, the SSIX infrastructure is also helped by many utility components. One of these utility components is the layer manager which manages the workload between the speed and batch layers. Another example, is an execution manager that ensures timely execution of both layers at required intervals. There is also a results health manager that ensures consistency of data in ElasticSearch after a batch job is finished. That would mean the removal of unnecessary real-time views or inconsistent results after reprocessing.

3.3. Architecture Diagram

Figure 2 presents the overall Lambda architecture of the SSIX platform with components from the functional tiers coloured according to the legend below. The ingestion tier acts as a starting point and provides input data for the SSIX platform. It is important to note that although the ingestion layer performs different types of processing on the data it collects, its output is treated as raw data by the SSIX platform. This data is fed to a Kafka queue and is later picked up by the speed and batch layers for Natural Language Processing and X-Score calculations. The output from both these layers in stored in an instance of ElasticSearch. SSIX exposes its services in form of a REST API. Query requests from this API are handled by the query manager that uses calculated data in ElasticSearch for query results. Different shapes in the following diagram are color coded according to their processing tier except for 'Queries' that are external to the SSIX platform.

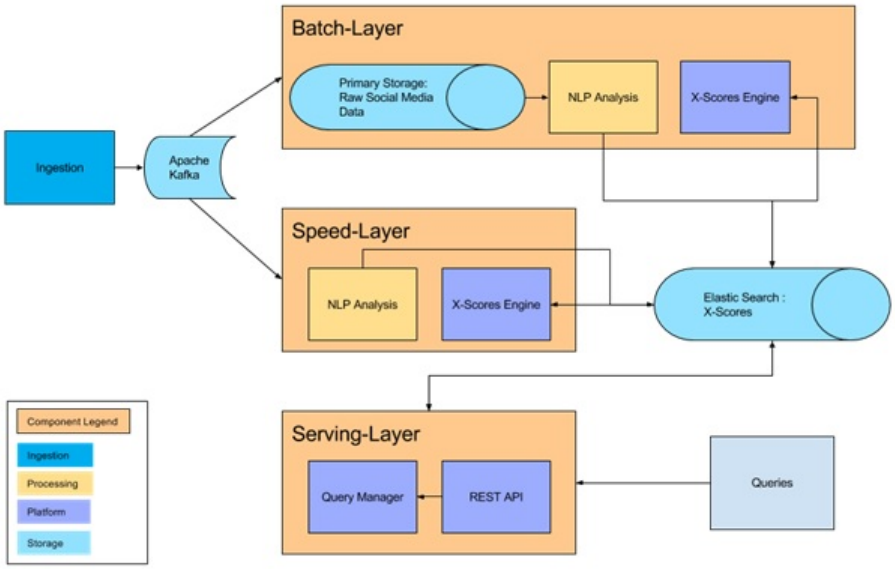

Figure 2. The SSIX Lambda System Architecture

[5]https://www.elastic.co/

3.4. Core Tiers

The core tiers of the SSIX platform are built around the functionality and work distribution of the SSIX project. They perform integral functions relating to different stages of data within the SSIX platform.

3.4.1. Ingestion Tier

The Ingestion tier is responsible for collecting, filtering, and enriching data. This data is then passed on to further tiers for processing.

3.4.1.1. Retriever The retriever component is responsible for collecting data from external sources, such as Twitter, Google News, Facebook, Stocktwits, etc. This component also maps different formats of external sources to internal data format.

3.4.1.2. Filter The filter component is a set of processes that carry out activities, such as removal of duplicates, and filtering on source or content based on specified criteria. This module is responsible for applying spam detection logics and for delivering cleansed contents according to the configuration of the request.

3.4.1.3. Enrichment The filtered data is subsequently processed by the Enrichment component that prepares the data structure before delivering the final content to the processing tier. This process includes operations of tagging with additional metadata (e.g., some contents are marked as spam or the identifier of the source is applied).

3.4.2. Processing Tier

The Processing Tier consumes the data stream produced by the Ingestion Tier to perform the analysis on each piece of data. Such analysis includes, normalisation, language identification and NLP analysis.

3.4.2.1. Normalisation Data from different sources is normalised under a common data model, so each analysis task further on just needs to implement support for a standardised API, simplifying integration of existing and additional analysis components.

3.4.2.2. Language Identification Given the quality provided by the Twitter built-in language identification[6], the first step which the content is subject to is a new language identification process implemented relying on Apache Tika[7] and other open-source libraries. The language identified is very important to then send the content to the right language specific NLP engine to provide concrete language support.

3.4.2.3. NLP Analysis The NLP analysis perform a parallel execution against all the NLP components. The first version currently provides sentiment analysis, but in future versions other NLP tasks (aspect oriented opinion mining, named-entity linking, summarisation, etc.) will be added to the pipeline.

[6] https://blog.twitter.com/2015/evaluating-language-identification-performance
[7] https://tika.apache.org/1.13/detection.html#Language_Detection

3.4.3. Platform Tier

The Platform tier performs two major tasks. It first performs statistical calculations on sentiment data to generate X-Scores and SSIX indexes within its X-Scores engine component. It also consists of a query manager to serve queries coming in through a REST API. These queries may be answered through simple lookups in storage or may trigger reprocessing of data.

3.4.3.1. X-Scores Engine The X-Scores engine calculates different X-Scores as configured throughout the systems. These are metrics, such as aggregated sentiment, polarity volumes, sentiment volatility, simple or exponential moving averages, but may also include custom algorithms and bespoke formulas.

3.4.3.2. Query Manager The query manager is responsible for handling query requests that come in through the REST API.

3.4.4. Storage Tier

The Storage tier provides persistent and temporary storage capabilities at different stages of the SSIX platform. Initially, the output of ingestion tier is stored within a non-relational database in Apache Cassandra[8] and the same is also passed to a Kafka instance for further collection by the processing tier. Output of the processing and platform tiers is stored in ElasticSearch to be used by the query manager.

[8]The Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance - https://cassandra.apache.org/

4. SSIX Templates and Scenarios

'SSIX templates' empower both the public and private sectors to develop innovative disruption-enabling mobile and cloud services and products, to leverage the massive amount of sentiment data that is constantly produced and published on various social media networks within multiple domains such as Finance, Economy, Government, Politics and Health.

The purpose of SSIX templates is to provide SSIX end users (Peracton, 3rdPLACE, Lionbridge) an easy way to personalise and customise several specific SSIX platform behaviours, such as time frame filter, keyword filter, spam filter, classification model, etc., without the need to change a line of code. Such approach is envisioned by generating and uploading (at load time[9]) –into the off-line SSIX engine– multiple SSIX templates, in order to initialise new/old variables that will alter in multiple ways the output of the platform at run-time[10]. The SSIX parters defined a SSIX Template as "SSIX Template is made of both configurable files (for example, but not limited to, XML) and software that implements the maximum number of variables possible to be declared in the data type file within SSIX software to allow personalisation for any targeted case study".

The SSIX templates are able to gauge the actual voiced sentiment from social media conversations, specifically emotional attributes, such as (but not restricted to) optimism and pessimism. These sentiment signals can be analysed to evaluate their influence on real world financial/economic/social/political outcomes and can act as valid indicators. An ideal paradigm that can benefit from the integration of SSIX templates is the field of investment decisions. Traditionally, research on securities, such as stocks, fixed income and foreign exchange relied on applying a Fundamental and/or Technical Analysis approach to determine the most efficient and lowest risk investment decision for a given amount of expected return. In this scenario, market sentiment is derived from the aggregation of a variety of these two disciplines (Fundamental and Technical analysis), including attributes, such as price action, price history, economic and financial reports/data, market valuation indicators, fund flows, sentiment surveys (e.g., ZEW Indicator of Economic Sentiment - A Leading Indicator for the German Economy), commitment of traders report analysis, analysis of open interest from the futures market, seasonal factors and national/world events. As a consequence, it is difficult to get a reliable and easy to interpret measure of a securities sentiment score without using a selection bias and almost impossible to measure a niche sector efficiently; this type of sentiment classification tends to be a lagging indicator to price movement but can act as confirmation.

The growth of social media APIs and the application of news analytics has provided a new method allowing sentiment analysis from a social media perspective to be carried out on financial securities, which has been proven to show a positive correlation to price performance ("Twitter is now a leading indicator of movement (up and down) of specific stocks - we can prove it.", Social Market Analytics). This data can be analysed to gain a greater understanding of sentiment behaviour and its correlation to price volatility for an individual security/sector or the entire market. By using this new sentiment data source,

[9]Load time is the period of time when a software program does not run. In this period of time, the software program is configured in order to be ready to run.

[10]Runtime is the period of time when a software program runs. It begins when a program is launched (executed) and ends with the program is quit or closed.

SSIX can deliver unique sentiment indices using X-Scores (a statistical framework of qualitative and quantitative parameters, such as regression, covariance and correlation analysis), such as the 'Social Sentiment Index for Healthcare' - SSIX_Health or the 'Social Sentiment Index for Technology' - SSIX_Tech, which will show the sentiment levels for their corresponding sectors, quantifying how market participants feel. X-Scores metrics can used in conjunction with industry standard technical parameters to analyse securities, such as Moving Average Convergence-Divergence (MACD), Relative Strength Index (RSI), Moving Averages (MA), Exponential Moving Average (MVA), Pivots Points, etc. SSIX X-Scores will provide real quantifiable data and tools to anticipate volatility and to analyse past performance, which will help develop alternative and more efficient approaches to reduce risk. SSIX can be used to identify trading signals, helping to make more informed investment decisions, resulting in a more efficient use of capital while reducing any associated risk. SMEs will be able to integrate the SSIX framework data into their own models for use in any area of application where sentiment analysis is used.

4.1. SSIX Scenario One

X-Scores metrics and SSIX Indices could be effortlessly incorporated into a quantitative trading strategy system. Security specific X-Scores sentiment indicators can be used to create buy and sell order strategies depending on real-world cross-lingual social conversations. For example, a currency trader can use an overall SSIX currency pair sentiment index to give a base for the current market sentiment of that currency pair but also integrate various X-Scores metrics data such as volume, short-term volatility, historical volatility and momentum of the index to get an indication of the change in market sentiment behaviour and find predictable patterns within the sentiment data, thus potentially forecasting its influence on future price dynamics. A short-term momentum trading strategy would benefit hugely from real-time sentiment data as it aims to capture gains by buying "hot" stocks and selling "cold" ones. A trader would take a long position in an asset, which has shown an upward trending sentiment, or short sell a security that has been in a down-trend. Using back testing and historical risk analysis, trading strategies can be improved to create more successful trades, generating increased profits while also lowering the risk of draw-downs. The cross-lingual analysis provided by SSIX would be able to identify regions where signals are strong and could be configured to give more priority to associated languages, enabling a niche sub-set of data to be analysed and correlated more precisely.

4.2. SSIX Scenario Two

SSIX indices and X-Scores metrics can provide national organisations and public sector bodies with actionable social sentiment data, which will enable them to gauge the general public's opinion on important issues that are effecting the public and on the services they provide, such as transportation and education. By understanding the sentiment of people living and working in a city, public sector bodies can make more informed decisions that will in turn lead to improved services for citizens and better use of valuable resources. Policy makers can monitor how policy proposals or budget proposals are being received as social media sentiment can act as real-time polling, social sentiment analysis offer governments and organisations new insights that can help them better understand and

respond to the public's concerns as they develop. Local authorities can use SSIX data to spot potential anti-social problems as they develop, helping them to react as soon as possible. Other areas can benefit from sentiment analysis, such as city planning, tracking health patterns like flu outbreaks, disaster response.

5. Big Social and News Data Management

Data retrieved from digital social networking and news sources provides significant data samples to the NLP component of SSIX. The entire process is developed through the following steps:

- Data download and gathering from different digital platforms (social networks, blogs, news sites, etc.) with different techniques (API usage, CSV download, Web scraping, etc.);
- Data cleaning and filtering to isolate significant information;
- Data processing to produce analysed and enriched data (smart data);
- Data sampling to extract pieces of smart data intended to be used by NLP component.

5.1. Big Data Challenges

In SSIX, multiple kinds of data are constantly collected, which process is continuous for the duration of the project. The following are types of data in question:

- Public available data from social networks
- Datasets part of the Linking Open Data (LOD) cloud
- LLD Cloud resources
- Public data available from domain-specific SMEs
- Survey data collected from independent events, such as technology summits, conferences, etc., or organised events, such as workshops, focus groups, etc.
- Financial and Economic trends outlined by the SSIX framework from analysis/mining of data
- Language Resources (LRs) either automatically acquired or reused from SentiWordNet (LR for opinion mining) and EuroSentiment (EU Project that provides a marketplace for LRs and Services dedicated to Sentiment Analysis).

Several challenges also arise due to the diverse nature of the gathered data. SSIX is able to deal with the three main challenges coming from the big data field namely, high volume, high velocity and high variety.

- High volume: constant growing of the data repository is managed through adoption of scalable technologies and architectures. The space required for the storage can be easily increased on request, while the technologies used are suitable to manage big quantities of data (e.g., Cassandra, Hadoop).
- High velocity: big stream of data is collected and managed with specific technologies and adequate processing capabilities. The project adopts high-performing servers with possibility to scale the computing power.
- High variety: the gathered data comes from multiple sources. In this case, each data source is treated separately. When required, an unstructured data model is implemented, in order to store information that can vary over time.

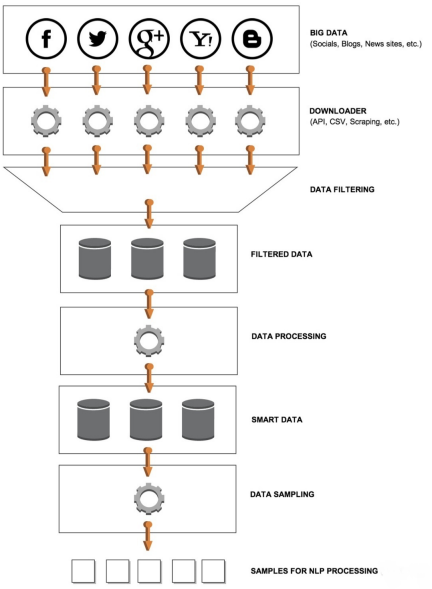

Figure 3. SSIX platform data-flow

5.2. From Big Data to Smart Data

Figure 3 illustrates the flow that all the data will follow before entering the SSIX platform for further NLP and analysis, which process transforms the data retrieved into smart data.

Each process is explained in more detail as follows:

- BIG DATA: indicates all the information available on different external platform in form of data sources (e.g., social networks, blogs, news sites, etc.)
- DOWNLOADER: the data are gathered from the different data sources using techniques, such as API usage, CSV download and parsing, web pages scraping, etc.
- DATA FILTERING → FILTERED DATA: a first process of noise removal and data processing that will produce a layer of filtered data.
- DATA PROCESSING → SMART DATA: in this phase of the process, all the data will be parsed and transformed into smart data.
- DATA SAMPLING → SAMPLES FOR NATURAL LANGUAGE PROCESSING: the last step will consist in the extraction of significant data samples destined for NLP.

All the smart data will be archived into a high performing repository. A cluster of servers will produce significant samples retrieved from the smart data repository that will be taken and streamed to the SSIX platform by and End Point component. The first prototype will use three physical servers to implement the architecture presented in Figure 4.

The schema defined in Figure 4 illustrates the ideal architecture delegated to retrieve data from the identified data sources, in order to process it and to create data samples for the NLP phase. The business case studies that will be executed in the duration of

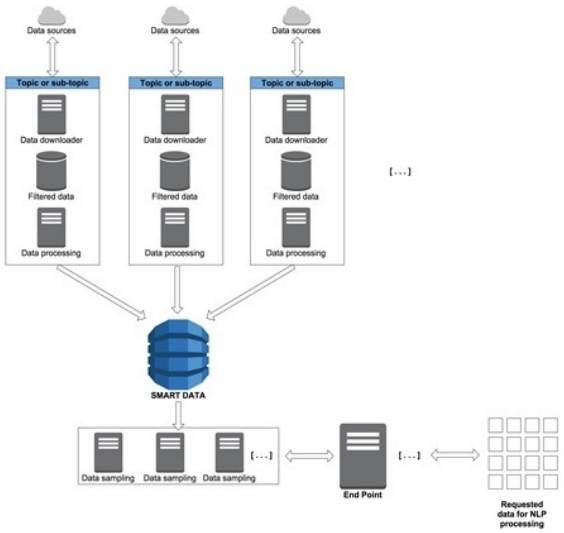

Figure 4. SSIX platform data architecture

the project (i.e. the ones discussed in Section 7) will be managed by a cluster of machines that will include: i) a software component that will interface with the different data sources, which will retrieve the data from them; ii) a repository of filtered data; and iii) a software component for data processing.

6. Natural Language Processing Services and Analysis

Analysing trends in social media content results in the process of a very large number of comparably short texts in near real-time. Therefore, the major challenge for the implementation of the NLP pipeline is in the orchestration of the different analysis components in a way that is potentially scalable in a cluster of servers that is able to handle hundreds of messages per second. Special care has to be taken to provide the NLP process as a distributed near real-time computation system that can reliably process unbounded streams of data. SSIX implements this process based on Apache Spark (as already mentioned in Section 3.2). Moreover, SSIX addresses the following major objectives:

- Automatic execution planning of NLP analysis processes: based on the descriptions of existing analysis components, available input and infrastructure, and desired output, SSIX automatically computes an appropriate execution plan;
- Standardised API for analysis components: a common problem in NLP processing is that there are many components for different, but related tasks, but they all implement completely different APIs, making it hard to combine them efficiently in a process. SSIX provides a standardised API and a standardised component description format to simplify integration of existing and additional analysis components.
- Sufficient collection of initial components: a big challenge in building this pipeline is to provide a sufficient collection of initial components so that we can (1) validate our execution model and API, and also provide examples for developers, (2) provide a process for real-time analytics, and (3) integrate with queuing and database technologies provided by SSIX. Figure 5 provides an overview architecture of the NLP pipeline.

6.1. Multilingual Language Resource Acquisition and Management

The multilingual language resource acquisition and management occurs in two phases:

1. Identification and resource of existing language resources for adaption for SSIX business cases (three business use cases are discussed in more detail in Section 7), i.e. exploitation of multilingual sentiment and domain specific lexica from European projects, such as EuroSentiment [5] –which provides a shared language resource pool for fostering sentiment analysis from a multilingual, quality and domain coverage perspective– or the adaptation of LLD resources and carry out any necessary localisation of monolingual resources where target language equivalents are scarce, such as Asian languages.
2. Exploration of unsupervised and/or semi-automatic corpus based methods for acquisition of multilingual lexica to support entity and sentiment analysis tasks.

6.2. Sentiment Analysis

Sentiment Analysis –the most important NLP task in SSIX– is all about "the computational study of peopl's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes" [17]. This is one of the most complex computational problems; ambiguities in language and issues, such as sarcasm, poor

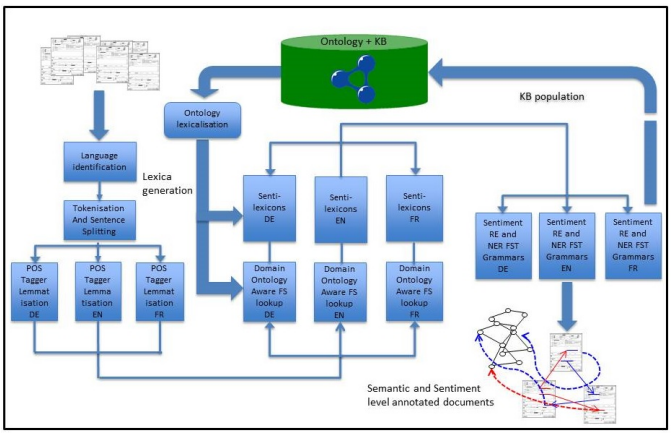

Figure 5. SSIX platform Knowledge-based NLP pipeline

spelling, sub-language, lack of context and the subtleties of sentiment make it difficult for an algorithm to understand the real sentiment behind a text expression. Such a problem is even more complex when the text to analyse does not follow formal grammatical rules, but it is colloquial text used in social networks, including abbreviations and new symbols. To understand the complexity behind this issue, it is enough to look at the result of three people tagging, for example 1,000, Tweets with their respective sentiment; often it is even difficult for humans to agree [15]. In order to avoid increasing the complexity, we restrict the scope of the sentiment analysis to the actual message content, not taking into account for now related content, such as images or external links.

In SSIX, sentiment represents the sentiment value in the range of -1.00 to 1.00. A continuous representation of the sentiment has been selected to better support later aggregation and development of more complex metrics. However, for the purpose of supporting class-based classifiers in the pipeline, a mapping based on the distribution of labels from manual gold standard annotation is also considered. Although only one analyser is used at runtime, this current SSIX pipeline ships four different implementations of sentiment analysers: two native in Java (StanfordNLP and GATE), one accessing an external commercial API (Redlink) and another one implemented in Python (NLTK). NLTK is used by default, given that it obtained the best results. All four sentiment analysers were evaluated in a qualitative nature on a Twitter data sample, that contained randomly selected tweets via a keyword search on Twitter for specific company cashtags. A high percentage of the sample suffered from poor quality or irrelevant content as only basic filtering was performed.

6.2.1. StanfordNLP

The first analyser is based on Stanford Core NLP[11]. Two different part-of-speech configurations are available: one using the default Part-of-Speech (PoS) tagger and another using the GATE Twitter[12] one.

[11] http://nlp.stanford.edu/software/corenlp.shtml
[12] https://gate.ac.uk/wiki/twitter-postagger.html

Stanford outputs sentiment as a vector (class membership probability vector) of five elements. Each element of the vector represents the probability of membership in one of five classes, where position 0 represents very negative sentiment, 1 - negative, 2 - neutral, 3 - positive, and 4 very positive. Thus [0.51, 0.38, 0.11, 0.0 , 0.0] would represent a negative sentence; because that is the class with the highest probability, but a tendency towards low negative and neutral sentiment is also probable. Based on that probability vector, a calculation that weights all classes, produces a more meaningful continuous sentiment polarity.

6.2.2. GATE

The second analyser is built using the General Architecture for Text Engineering[13] (GATE). GATE is a natural language processing framework developed in Java. GATE Embedded[14] has been used to integrate the application with the pipeline runtime. Although Embedded could have also been used to develop the pipeline, GATE Developer was used instead owing to ease of use. The finished pipeline is saved as a GATE application and loaded into the pipeline using GATE Embedded.

The analyser loads the GATE application and processes Tweets that are passed through it. Tweets are passed in plain text, one at a time, but GATE documents can also be created from JSON or XML format Tweets for later stages. These are preprocessed to annotate text as a Tweet annotation. This annotated text is then fed to the application for sentiment annotations. The application currently returns sentiment classes as positive, negative, or neutral that are mapped to 1.0, -1.0 and 0.0 respectively.

6.2.3. Redlink API

The Redlink API[15] consists of a number of services for content analysis, Linked Data publishing and Semantic Search. It provides Sentiment Analysis support based on Apache Stanbol[16]. Redlink grants SSIX unlimited free access for research purposes during the project execution.

The implementation of this analyser is simply a wrapper to the Redlink Java SDK[17] requesting the sentiment analysis there, so the pipeline does not handle the actual computing.

6.2.4. NLTK

NLTK, the Natural Language Toolkit, is the fourth analyser. It is a Python-based API for developing NLP solutions, providing interfaces to NLP resources (e.g. WordNet, corpora) and tools (e.g. TreeTagger). The NLP Sentiment Analyser uses the Vader rule-based model for sentiment analysis on social media (micro-blog) data [Hutto2014].

The implementation of this analyser consists of a RESTful Flask wrapper[18] to a NLTK microservice implemented in Python.

[13] https://gate.ac.uk/
[14] https://gate.ac.uk/family/embedded.html
[15] http://dev.redlink.io/api
[16] http://stanbol.apache.org/
[17] http://dev.redlink.io/sdk#java
[18] http://flask-restful.readthedocs.org/

7. Business Case Studies

Three business use cases were defined by three of the SSIX industrial partners: Peracton, 3rdPLACE and Lionbridge, in accordance with their business needs and market objectives as further discussed in the sub-sections below. The goal behind these business use cases is to exploit the SSIX technologies commercially.

7.1. Augmented Investment and Trading by Social Sentiment Indexes

In order to sharpen and improve the decision making capability of Peracton's financial analytics MAARS platform (http://peracton.com/maars), with the help of SSIX project, Peracton will add market sentiment data to enrich their financial analysis. Market sentiment data will allow improvements, such as more complex type of searches, advanced what-if scenarios as well as multiple financial back-testing scenarios. In this respect, Peracton require a highly customisable social sentiment data input to be provided to MAARS. Such input (what is called a 'SSIX sentiment index template') would need to be personalised and integrated (in a light manner) with the MAARS analytics algorithm. The SSIX template will be used to calculate social sentiments for securities (i.e., stocks in Peracton's case study). The results (unique sentiment numbers attached to stocks) will be added then as independent parameters within the MAARS platform. By this means, Peracton will be able to do more complex investment/trading type evaluations. Performance analysis will be made over historic data and real-time, in order to determine the impact of SSIX indices and then, recommendations will be provided back to the consortium.

The following five phases are required in order to integrate the SSIX templates with MAARS:

Phase 1: Establish data sources and targets in order to generate unique SSIX Indices

In this phase Peracton together with SSIX partners, will identify the suitable data sources available on various social networks (Twitter, Facebook, LinkedIn, Google+). Peracton will provide the list of stocks to be tracked on social networks.

A proprietary methodology of targeting the right audience to infer the sentiment indices is being developed.

Phase 2a: Set-up MAARS server to connect to SSIX engine

While the SSIX engine will be under development, the MAARS server will be set-up and tested, in order to be ready to connect and take in SSIX sentiment indices data.

Phase 2b: SSIX indices generation and storage Once the data sources are established, the SSIX engine will be instantiated to generate the very first sentiment indices to trace U.S. stocks on AMEX, NASDAQ and NYSE. This first instantiation will be for testing purposes in order to test SSIX output, indices storage, and prepare integration with MAARS. A first period of testing versus benchmark will follow, prior to integration with MAARS platform, in order to have SSIX indices output calibrated.

Phase 3: SSIX indices integration within MAARS analytics software

From a technical point of view, SSIX should produce numerical values between (-100;100) that will represent the sentiment value for each stock tracked. The values will be delivered in a database (e.g., MySQL). MAARS will retrieve this data automatically from there.

The generated index sentiment values, will be integrated (from an IT technical perspective only) within MAARS analytics. This will be a very first attempt, as this integration may take two to three iterations. A second calibration of SSIX will follow, based upon the feedback provided to the consortium by Peracton.

Phase 4: Trading and Investment with SSIX indices

As sentiment data starts to be updated within MAARS analytics, the MAARS team will perform simulations tests of investing and trading. There will be trading and investment tests with no SSIX sentiment data (control tests) and then in parallel, same investment and trading tests involving sentiment data.

Phase 5: Feedback to be provided to SSIX consortium in addition to the feedback already provided during the iteration process

Based upon Phase 4 tests, feedback will be provided to the consortium with regard to the performance, and changes in results of investment/trading exercise due to using sentiment data.

Figure 6 presents how these phases integrates for the investment and trading case study.

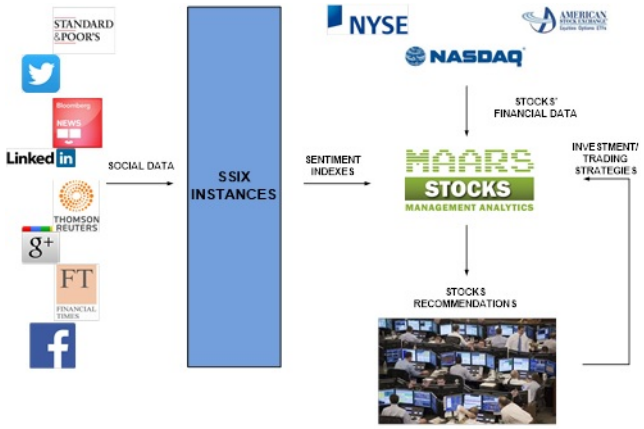

Figure 6. Investment and Trading Case Study

7.2. Boosting the editorial production of financial publishers with 3rdEYE

The needs of publishers operating in the finance sector are not utterly different from the expectations of a more general online publisher. They both need to find interesting topics on which to create effective articles that can generate revenues instantly. However, the difficulty when writing content related to the finance world is connected to the lack of authoritative sources that can provide fresh news timely. The high changing frequency of the markets is strongly influencing the news production; it is therefore important to acquire the ability to discern between irrelevant rumours and significant trending topics.

This specific demand resulted in the definition of a software by 3rdPLACE that can help the publishers to easily identify fresh and significant informations regarding the companies about which they want to write articles. Another key feature is related to the possibility to discover the most important influencers that can be contacted and

engaged for the activities of dissemination and distribution of the published contents. Hence, the final aim of the integration between 3rdEYE (http://3rdplace.com/en/3rdeye/) –a proprietary software for Data Management that supports businesses, allowing them to utilise data driven decision making– and the SSIX platform's output is to provide the finance publishers with an enhanced tool which could assist them during their editorial activities.

The final product will involve a set of consecutive phases. Figure 7 represents the implementation workflow, from the initial data source to the final automation on the publisher's website.

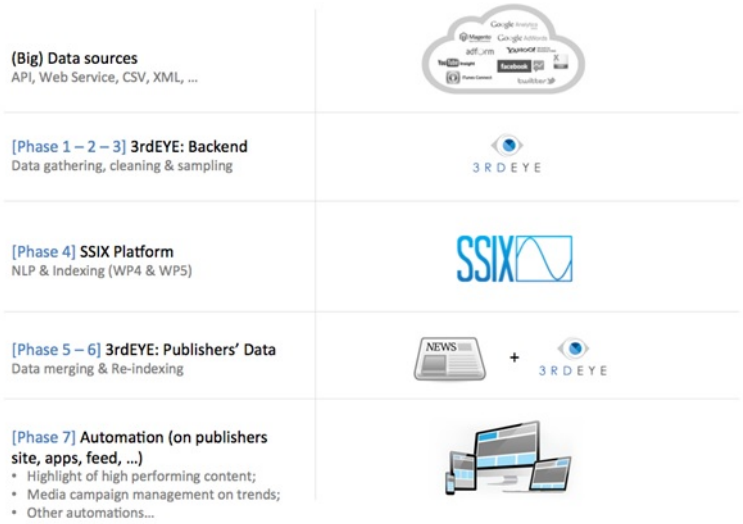

Figure 7. Publishers and Finance Case Study

Phase 1: Data Collection The available procedures collect data from distributed external sources, using different techniques. During this phase, accurate studies take place in order to define the keywords used to gather the data for a specific topic.

Phase 2: Smart Data Identification Not all the collected data is useful for the analysis. A comprehensive analysis approach would be deleterious or at best not very effective and efficient. It is therefore useful to apply a first level of data filtering, which aims to reduce the amount of data and determine its quality. The resulting data set will be ordered on the basis of a predefined number of logical and business criteria and will configure the group of data fundamental for the analysis (also known as Smart Data).

Phase 3: Data Sampling The group of Smart Data, even though reduced compared to the initial Big Data, may still be prohibitive to perform analysis. It is therefore necessary in many cases to provide a mechanism of statistical data sampling that is functional in order to obtain rapidity and ensure representativeness.

Phase 4: SSIX Integration NLP services developed for the SSIX platform will be providing a fundamental contribution to the contents' evaluation process, since they will be used to increase or decrease the importance attributed to tweets and news. The SSIX

indices will enable a level of contextualisation of the contents and will allow to enhance the sorting algorithms implemented inside the 3rdEYE platform.

Phase 5: Integration with Publishers' Data The indications that work well for a publisher may not work as well for another one. It is therefore essential to take into account the specificities of each project and to consider, in addition to external data - the ones not owned by the publisher - also the publisher's internal data, which will influence the insights (parameters: traffic, monetisation, engagement and sharing).

Phase 6: Relevant Suggestions 3rdEYE will be finally able to provide enriched information to the publisher, suggesting to produce effective content on the basis of reliable financial trend topics.

Phase 7: Automation 3rdEYE can be queried via an API from the publisher's systems in order to provide the classification of the more performing elements.

The publisher will be able to use this information to automate some processes: highlighting of the most performing contents, automatic changing of the bids on the SEM campaigns based on the trends, etc.

7.3. Improving sales performance and reducing risk by using natural language processing

This business case study will focus on how sentiment analysis and opinion mining, can be used to create business value for companies. Here, SSIX will be used as a tool for Business-to-business (B2B) sales representatives that monitor company activity, including sentiment about a company's activities. More specifically, Lionbridge will aim to improve the identification of new opportunities or threats, and enhance existing customer relations. The underlying idea is that social media produces signals which can be turned into valuable business intelligence. Usually these signals remain unnoticed, but sentiment analysis can be used to detect them, and it is the aim of Lionbridge to tap into these signals. Other NLP techniques will be used to monitor sources, such as online news aggregates and financial information providers. Some tools for this already exist in the market, but the aim is to widen the scope of sentiment analysis by including multiple languages in the NLP pipeline.

The amount of published financial commentary is massive, and overall, the greatest challenge seems not to be finding suitable companies or information about them, but to find a way to get into discussions with the interesting companies. Social media sentiment in general and the SSIX platform in particular can help with this by identifying trending topics and products, forecasting the performance of tracked companies, and by filtering and distilling the information flow, among other things. Lionbridge will approach the case study as an opportunity to build a completely new product for their sales teams. The end users will access the sentiment data and indices created by the NLP processes directly via the SSIX platform. This strategy will enable Lionbridge to focus more on fine-tuning the NLP processes, and will also give them a good opportunity to test the platform in actual use.

This business case study is conducted in four phases:

Phase 1: Identify and collect relevant user data The targeted user group was Lionbridge B2B sales representatives.

The data collection phase will adopt the concept of user stories as they are used in agile software development [22]. User stories are concise descriptions of what the end

users do in order to reach their goals and they are good for expressing the business value of features being developed. To collect the user stories, Lionbridge identified pilot end users within their company, and approached the group with a questionnaire (Q1) about their current means of gathering the information they need in their sales activities.

Phase 2: Refine collected user data The user stories extracted from the Q1 represented the sales representatives' current situation: the tools and sources they used to gather certain facts to reach their objective of finding the business goals and directions of other companies and match them with Lionbridge's services. The SSIX project aims to provide a different tool to reach this objective in a more efficient manner, including information and sources that our user group currently does not make use of. The user stories were therefore modified and extended (called Selected User Stories) to better support development of the SSIX platform.

The selected user stories were given a priority between 1-3 (1 being high priority, 2 medium priority and 3 low priority) taking into account their importance for Lionbridge's user group and feasibility within the SSIX project. The data sources mentioned in Q1 were also classified with respect to their importance for Lionbridge's user group, per their mentions in Q1, and the feasibility of using the sources within the SSIX project.

In a second round of user feedback (Q2), the selected user stories with priority 1 and 2 were presented to a larger group of end users, including the original small group of users, who were asked to estimate the importance of each user story based on their own needs. This method allowed Lionbridge to confirm and refine the importance of the user stories. The users were asked questions about the SSIX functionality, and about how often they visited the sources collected in the first round of user feedback, e.g. Twitter or Google News, with the purpose of gathering information about other companies.

Phase 3: Define custom SSIX indices Based on the Q2 user ratings of the importance of the user stories Lionbridge finalised the prioritisation of user stories, taking into account importance for the users, feasibility within SSIX and overlap with the business use cases of Peracton (Section 7.1) and 3rdPLACE (Section 7.2). This allows SSIX platform development to focus first on the user stories with highest priority, and to add new features incrementally. For the priority 1 and 2 user stories, Lionbridge then defined custom SSIX indices.

Phase 4: Evaluate the SSIX platform

Evaluation of the SSIX platform with respect to the defined business use case i.e. as a B2B tool used by sales representatives to monitor company activities. Several factors, such as high accuracy of the sentiment signal, data sources and multilingual support, will be evaluated.

8. Conclusions and Future Directions

SSIX seeks to extract and measure meaningful financial sentiment signals in a cross-lingual fashion, from a vast multitude of social network sources, such as Twitter, Facebook, StockTwits, LinkedIn, and public media outlets, such as Bloomberg, Financial Times and CNBC. It will generate custom X-Scores powered index for a given sentiment target or aspect, i.e. company or financial product. The primary domain is Finance although SSIX has scope for Environment, Health, Technology, Geopolitics and beyond. The X-scores will be used by the industrial partners and bundled with their financial analytics, in order to increase the accuracy of their output combined with either end of day financial data or, real time data feeds. SSIX will adapt existing mature, proven and scalable open source text mining tools in order and circumvent language barriers with respect to unexploited multilingual financial sentiment content by harvesting cross lingual Big Social Media and News Data. Semantic Analytics will be employed to generate SSIX indices.

In terms of future directions, several considerations towards improving the overall SSIX system will be taken. These include using massively parallel computation for NLP or statistical calculations, researching faster text indexing techniques for categorization and searching, considering terms other than cashtags to categorise information and determining a better weight vector and improvement to NLP processing or machine learning components. As for the Sentiment Analysis task within the project, several Deep Learning-based techniques are being looked at to improve the overall quality of the proposed technology.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme ICT 2014 - Information and Communications Technologies under grant agreement No. 645425.

References

[1] Applying the Big Data Lambda Architecture. *http://www.drdobbs.com/database/applying-the-big-data-lambda-architectur/240162604*, 2013.

[2] FIRST - large scale inFormation extraction and Integration infrastructure for SupporTing financial decision-making. *http://project-first.eu/*, 2013.

[3] MONNET - Multilingual Ontologies for Networked Knowledge. *http://cordis.europa.eu/fp7/ict/language-technologies/projectmonnet_en.html*, 2013.

[4] AnnoMarket. *https://annomarket.eu/*, 2014.

[5] EuroSentiment. *http://eurosentiment.eu/*, 2014.

[6] OpeNER. *http://www.opener-project.org/*, 2014.

[7] TrendMiner. *http://www.trendminer-project.eu/*, 2014.

[8] LIDER: "Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe". *http://www.liderproject.eu/*, 2015.

[9] Freme. *http://www.freme-project.eu/*, 2017.

[10] MixedEmotions. *http://mixedemotions-project.eu/*, 2017.

[11] Sven-Kristjan Bormann. Sentiment indices on financial markets: What do they measure? Economics Discussion Paper 2013-58, Kiel Institute for the World Economy, 2013.

[12] Brian Davis, Keith Cortis, Laurentiu Vasiliu, Adamantios Koumpis, Ross McDermott, and Siegfried Handschuh. Social sentiment indices powered by x-scores. 2016.

[13] P. A. Gloor, J. Krauss, S. Nann, K. Fischbach, and D. Schoder. Web Science 2.0: Identifying Trends through Semantic Social Network Analysis. In *International Conference on Computational Science and Engineering. CSE '09.*, pages 215–222.

[14] David Greenfield. Social media in financial markets: The coming of age... Gnip whitepaper, GNIP, 2014.

[15] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pages 27–35. Association for Computational Linguistics, 2009.

[16] Jay Kreps. Questioning the lambda architecture. *O'Reilly Online article, July*, 2014.

[17] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[18] Nathan Marz and James Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2015.

[19] Alex Micu, Laurens Mast, Viorel Milea, Flavius Frasincar, and Uzay Kaymak. Financial news analysis using a semantic web approach. In *Semantic Knowledge Management: an Ontology-based Framework, Paolo Ceravolo, Ernesto Damiani, Gianluca Elia, Antonio Zilli (Eds.)*, pages 311–328, November 2008.

[20] Piotr Mirowski, Marc'Aurelio Ranzato, and Yann LeCun. Dynamic auto-encoders for semantic indexing. In *NIPS 2010 Workshop on Deep Learning, Proceedings*.

[21] Ontology2. FSI: Financial Semantic Index. *http://financialsentimentindex.com/fsi/*, 2012.

[22] Kenneth S Rubin. *Essential Scrum: A practical guide to the most popular Agile process*. Addison-Wesley, 2012.