A Taxonomical NLP Blueprint to Support Financial Decision Making through Information-Centred Interactions

Siavash Kazemian† and **Cosmin Munteanu**‡ and **Gerald Penn**†
†Department of Computer Science
‡Institute of Communication, Culture, Information and Technology
University of Toronto
{kazemian,mcosmin,gpenn}@cs.toronto.edu

Abstract

Investment management professionals (IMPs) often make decisions after manual analysis of text transcripts of central banks' conferences or companies' earning calls. Their current software tools, while interactive, largely leave users unassisted in using these transcripts. A key component to designing speech and NLP techniques for this community is to qualitatively characterize their perceptions of AI as well as their legitimate needs so as to (1) better apply existing NLP methods, (2) direct future research and (3) correct IMPs' perceptions of what AI is capable of. This paper presents such a study, through a contextual inquiry with eleven IMPs, uncovering their information practices when using such transcripts. We then propose a taxonomy of user requirements and usability criteria to support IMP decision making, and validate the taxonomy through participatory design workshops with four IMPs. Our investigation suggests that: (1) IMPs view visualization methods and natural language processing algorithms primarily as time-saving tools that are incapable of enhancing either discovery or interpretation and (2) their existing software falls well short of the state of the art in both visualization and NLP.

1 Introduction

There are many stakeholders and agents that interact within the space of financial markets. *Investment management professionals* (IMPs) play the most prominent role here. On a macro scale, IMPs are responsible for the long-term strategies of institutions such as mutual funds, pension funds, sovereign wealth funds, etc. At the core of their activities lies information seeking - staying well informed by understanding market trends through reading external reports and developing their own predictive models based on thorough statistical analysis of large and varied sources of data.

Within the technological space of tools that support such information-seeking activities, natural language processing (NLP) research is already tackling tasks that IMPs perform, e.g., trading securities based on sources such as newswire, company quarterly reports, financial blog posts, and social media text (Bollen et al., 2011; Kazemian et al., 2016; Zhang and Skiena, 2010). Our study has revealed that textual and spoken documents are highly valued by experienced analysts, because they yield nuanced insights not available in aggregated, numerical data.[1]

Our critical survey of major financial-analysis software (e.g., Bloomberg Terminal, FactSet) reveals, however, that while this software is ubiquitous, its use of tools that could amplify understanding or enable discovery within natural-language sources is extremely conservative. This is particularly noticeable against the backdrop of a general trend in the financial sector towards automation of information processes, and an abstract awareness that ever-expanding NL datasets can facilitate more nuanced decision making (Flood et al., 2016).

Nevertheless, as we elaborate upon below, we have found that the "boots on the ground," the IMPs themselves, seem to assess the value of visualization and NLP techniques, as applied to their own use of unstructured natural language artifacts, exclusively in terms of *faster* analysis, with no prospect of *better* analysis — a world of "little data," mostly disconnected from the "big data" that they have read about in the popular press, in which computers can be relied upon to fetch and render natural language content but are largely superfluous to the analysis and interpretation of that content as IMPs require. This view persists because of a commodified view of NLP in which the literal under-

[1] As one of our participants bluntly explained: "*The thing about having a job in the market is at all times you're trying to not lose money and hopefully gain money. At any point when relevant information comes out, you need to know. For example, what Yellen said, everyone needs to know, if there is a loser who doesn't know, he is going to lose money at the expense of his ignorance*" (P1).

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 89 - 98

standing of speech and language is viewed as either trivial or at least a mostly solved problem, through the lens of commercial successes such as Siri, IBM Watson, and Google Now (Milanesi, 2016). In other words, zealous misrepresentations of what NLP research has already accomplished have tragically impaired IMPs' awareness of the goals and capabilities of contemporary NLP, and have been perhaps *the* major obstacle to a more pragmatic utilization of NLP within this community.

As will shortly become apparent, this is not an NLP research paper, nor have we attempted to reform the perceptions of IMPs. But because a central goal of the financial NLP community is to design intelligent interfaces and software that will better support the information practices of these IMPs, the ethnographic HCI research presented here is important to the financial NLP community, as it identifies critical aspects of the information practices of IMPs that are not being supported. The good news is that the problems being addressed by past offerings of this workshop series are well positioned to address many of these aspects.

Below, we first describe our investigation of the information practices and processes of IMPs from an Information-Seeking Process (ISP) perspective (Marchionini, 1995). We conducted a Contextual Inquiry (CI), from which we infer a taxonomy of information seeking tasks related to analyzing natural language documents (Study 1). We then conducted a series of Participatory Design (PD) workshops to validate the taxonomy and explore how revisions to the current software interfaces can better support the ISPs captured in our proposed taxonomy (Study 2). After presenting the insights from Study 2, anchored in the proposed taxonomy, we suggest design approaches for using visualization and NLP tools to support the ISPs of IMPs.

2 Background

Central banks such as the US Federal Reserve (Fed) or the European Central Bank (ECB) play a prominent role in deciding the monetary policy of a jurisdiction (Bernanke and Kuttner, 2005). The leaders of central banks hold several press conferences a year to inform the public about their activities, and to give guidance on how they might act going forward. Similarly, publicly traded companies play a significant role in the capital markets by providing investment and risk mitigation opportunities to financial organizations. Public companies are required to hold regular earning calls to update the public on their activities. To IMPs, such events are critical to their risk mitigation efforts; the transcripts of these calls or conferences are thus valuable.

IMPs (often referred to as *financial analysts*) make investment decisions on behalf of their employers (*buy-side analysts*) or provide advice to large investment banks (*sell-side analysts*). The scale and complexity of their decision making sets them apart from retail analysts who advise individuals or small businesses on their investments.

In our first study, we examine how IMPs make use of spoken records of central bank news conferences and earning calls in their professional activities. In the second study, we will use their input from this first study to investigate better design approaches for software that supports the use of such records in their information seeking practices.

3 Related Work

Observational studies have investigated the workflow and information practices of IMPs, producing taxonomies of the information transfer process from sell-side to buy-side analysts (Ramnath et al., 2008) or details of accounting practices (Bouwman et al., 1995). However, these do not capture the IMPs' information-seeking needs themselves. They also do not describe how IMPs interact with information systems to satisfy their information needs.

Under the banner of Interaction Capture and Retrieval or ICR (Whittaker et al., 2008), however, there have been observational studies of somewhat related information practices that use spoken records of events. Whittaker et al. (1998), for example, investigated how recorded voicemail was used in a corporate setting, and incorporated their findings in the design of an improved voicemail indexing and retrieval system (Whittaker et al., 2002). Jaimes et al. (2004) studied why and how users review meeting records in order to guide their development of a cue-based meeting retrieval system. Whittaker et al. (2008) conducted another field study in which they observed how people were using records of meetings, and showed that although technology such as Speech Excision (Nenkova and Passonneau, 2004) is effective, it was not incorporated into state-of-the-art meeting browsers.

These prior studies, along with other research in the meeting domain (Bertini and Lalanne, 2007;

Jaimes et al., 2004; Lalanne and Popescu-Belis, 2012), have confirmed the relatively limited utility of more traditional meeting artifacts such as minutes, personal notes, and raw audio-visual recordings, and point to software-enabled tools as more effective. These include speech recognition and speech excision for voicemail and meetings (Whittaker et al., 2002, 2008), but there are many other promising candidate technologies: speech alignment (Goldman, 2011), disfluency detection (Liu et al., 2006), speaker segmentation (Budnik et al., 2016), information extraction (McCallum et al., 2000), answer selection, an important step in question answering (QA) systems (e.g., Jauhar et al., 2016; Rao et al., 2016), machine comprehension (Rajpurkar et al., 2016), and sentiment analysis (e.g., Rosenthal et al., 2017; Socher et al., 2013).

These technologies, furthermore, as well as the remarkable pace of their advancement, are known to software developers who support IMPs, despite the lack of a design investigation that explicitly connects these advancements to IMPs' needs and practices. The remarkable performance boost in answer selection between 2004 to 2016 on datasets containing financial news (from a mean reciprocal rank of 0.4939 (Wang et al., 2007) to 0.877 (Rao et al., 2016)), for example, was well publicized among these vendors, as were the significant improvements to machine comprehension, which extracts exact answer phrases to questions from raw text, in the space of a single year — from 50.5% (Rajpurkar et al., 2016) to 78.6% (Rajpurkar et al., 2018) (3.6% shy of human performance). Sentiment analysis of financial news was understood to have improved automatic trading from roughly 30% to 70.1% annualized returns (Kazemian et al., 2016), and the use of sentiment analysis in market analysis tools has been commonplace now for almost 10 years (Cambria et al., 2013).

With the exception of sentiment analysis, however, the absence of any serious, contemporary NLP functionality is notable. This paper takes a first step towards an explicit design investigation of the potential of this functionality by proposing (Study 1) and validating (Study 2) a design-minded taxonomy of information practices within the financial analysis domain.

The information-seeking process has been characterized as a highly variable process shaped by information seeking factors such as the task and information domain (Marchionini, 1995). For different tasks and information seeking factors, different types of support are needed by information seekers (Toms et al., 2003; Vakkari, 2003). Methods such as Contextual Inquiry (CI) are effective in uncovering such information seeking factors (Beyer and Holtzblatt, 1999), while approaches such as Participatory Design (PD) (Schuler and Namioka, 1993) are useful not only for engaging users in the design process but for refining the functional requirements of information support tools (Lalanne and Popescu-Belis, 2012).

4 Study 1: Observing Spoken Document Use

Spoken documents contain unique and critical information for IMPs. They are rich with both factual and affective data, and yet this medium is not adequately supported by existing financial analysis software such as Bloomberg or FactSet. Moreover, these spoken documents contain both content authored by the institutions holding the events (e.g. Federal Reserve), as well as Q&A from journalists and analysts that, as will be discussed, give transcript readers clues about their future publications, and thus about the markets' reaction to the events. Hence, the focus of our taxonomy is on IMPs' use of spoken documents such as transcripts from the Federal Reserve. In particular, we focus on overall information and decision-making practices, instead of users' interaction, or the use of specific elements of the text, a topic extensively studied in linguistics.

We conducted a contextual inquiry, observing how IMPs utilized spoken records. Eleven analysts (4 female, 7 male) who actively use transcripts responded to our participation call, which had been distributed through our professional network and word of mouth. All participants had more than 5 years of experience in their field, and were currently working at Wall Street (New York, USA) hedge funds, asset management firms, central banks, and large multinational investment banks. The study was conducted at participant offices. Participants were instructed to choose spoken records they would be interested in reading as part of their professional activities. The documents that they chose were transcripts of earning calls of publicly traded companies or of news conferences given by leaders of central banks.

A researcher observed them during reading; after they read, he later conducted a semi-structured interview with the participants to gain insights into

the 6 information-seeking factors defined by Marchionini (1995) that characterize their ISP, the lens through which we view their interactions with information systems. These are: *setting, information seeker, domain, task, search systems,* and *outcomes.* The interviews were recorded and transcribed. The study's data consist of these transcripts as well as the observation notes. An Inductive Thematic Analysis was used to extract the major themes in the dataset (Braun and Clarke, 2006), conducted under an essentialist epistemological approach, in which language is seen as a reflection of intended meaning and individual experience (versus a constructional perspective, in which meaning and experience are viewed as socially constructed). In this paradigm, one can theorise about individual motivations. No theoretical framework was used, however, as our goal was not to measure fit with a particular theory.

The participants/readers all work for organizations that are market participants, entities that buy and/or sell assets in the investment markets. When describing their professional duties, the participants noted that they exclusively made decisions in a group, highlighting the collaborative nature of their ISPs. 8 of the 11 participants noted that they usually updated their team about what they learned after reading transcripts. Their task can therefore be formulated as extracting from the content of these spoken records key takeaway points that could be referenced later or shared with colleagues.

They furthermore noted that in this industry, time is of the essence. Even minor delays in investment decisions could be very costly. This is the major reason that IMPs tolerate working with error-laden transcripts, so long as they are available sooner. It also explains their expert proficiency in skimming and skipping over information they already know or find irrelevant.

The meta-goal of participants is to increase institutional returns. For this, participants need to develop insights about future actions (e.g., whether the Fed would raise rates) and outcomes (e.g., whether a company's total revenue would appreciate over the next year) of the organizations they study (e.g. a company or a central bank). Just as important, the users also need to develop a good understanding of the markets' expectations of those actions and outcomes. The success of an IMP hinges not just upon a more accurate grasp of the organizations' futures actions and outcomes, but of a differentially better understanding than the general market consensus.

Table 1 summarizes the information our participants tried to extract: the *Essential Predictive Knowledge (EPK)*. In order to assess the future actions and outcomes of an institution and the markets' reactions to them, readers mined information related to the organization, the speakers in the recordings, and external factors (T1 in Figure 1).

4.1 Taxonomy of ISP Subtasks

Our interviews reveal that none of what is communicated by the speakers is viewed by our participants as ground truth. Instead, the content is interpreted by comparing it to previous communications from the same organization and speakers, and in the context of their activities and market perceptions. The speakers representing the institution know about this complexity. Their aim is to send carefully drafted messages to their audience (T2a), which may or may not be supported by all of the facts available. From these messages, and by considering contextual information, our participants aim to extract "the truth" about the organizations. To do so, they performed several sub-tasks, which we summarize as a taxonomy in Figure 1. First, all participants interpreted facts about EPK from transcript content (the what). This starts with forming a solid understanding of the company's past actions and outcomes, as well as the "dialogue" about the company. Next, readers take notes on disclosed information as well as referencing related information not shared in the transcript. For instance, in his analysis of a company's unusually large reported loss in revenues, P7 had to consider market rumors that the company was losing its largest institutional client, concluding that the rumors could be true, and that they would negatively impact the company's long-term profitability.

In addition, the users assessed the communication acts themselves in the transcript (the how). Special attention was paid to tone or sentiment of communication (P1, P3-5, P7-10), which was described as "bullish" (or "bearish"), "unabashed" (or "reserved"), "positive" (or "less positive"), "gung ho" (or "defensive"), and "hawkish" (or "dovish"). According to these participants, the expressed sentiments were not only a good clue about the organizations' future actions, but also have an effect on the short-term market reaction to the communicated content. Communications tactics used by the speakers were also discussed (P1, P4-5, P7,

	Past	Present	Future
External Factors	Market / A&O of other institutions	Analysts' Q&A	*Market reactions*
Speakers	Professional history, previous remarks	Cognitive and affective state	*Leading actions*
Studied Institutions	A&O, Guidance	A&O, Guidance	*A&O*

Table 1: The sought-after knowledge for predicting future actions and outcomes (A&O).

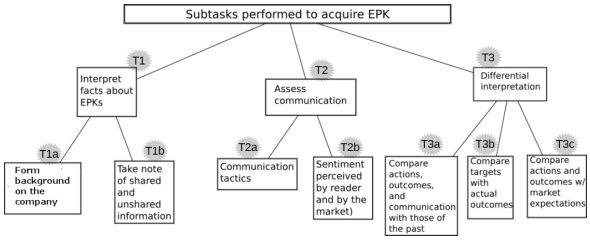

Figure 1: Tasks performed by IMPs to predict future action and outcomes of the studied organization and the market's reaction.

P10-11), such as side-stepping questions, providing "evasive" responses, repeating important content to signal salience, or providing more details about key subjects during Q&A.

Finally, as our participants integrated their newly gained EPK and made higher-level assessments, they compared it with past actions, outcomes, past communications, market expectation, and (differentially) the organization's past guidance.

5 Study 2: Participatory Design

The competitive nature of the financial markets has made this user group largely unavailable to participation in user studies. We were fortunate to be able to recruit four professionals to participate in Participatory Design (PD) workshops (3 males, 1 female, identified as D1 to D4). D1 and D2 had also participated in Study 1. All had more than six years of experience. D3 and D4 were sell-side analysts for investment banks and equity firms, while D1 and D2 were buy-side analysts for large global asset management firms and hedge funds.

5.1 Methodology

The four PD workshops were attended by a facilitator, a participant, and a visual artist. The visual artist's role was to assist participants with sketching their proposed ideas and to help facilitate the visual conversation, mitigating participants' potential lack of sketching or drawing expertise. In the design workshops, the participants started with a warm-up by reading a transcript as they normally would in their work routine. They were subsequently asked questions about how and why they read transcripts. The participants were then introduced to a scenario similar to the first study. The scenario involved examining the content of a transcript relevant to an investment decision that the participant's hypothetical employer was considering. The participant then wrote five takeaways from the transcript, potentially including an investment recommendation, for the purpose of sharing it with their team members.

After presenting the scenario, the participants were provided with drawing tools, large sheets of paper, and a new transcript of their choosing. The participants were told that 'the sky is the limit' for technologies they can incorporate into their designs: visualization, navigation, and artificial intelligence. We deliberately described the available technologies vaguely, to avoid priming participants toward specific technologies. They were also asked to think about tools that would provide appropriate assistance for their ISPs given the ecosystem of platforms they regularly use (e.g. Bloomberg Terminal or FactSet).

5.2 Data Collected

Each workshop (1.5 to 2.5 hours), was video recorded, and produced one design sketch. The components of the designed systems (UI features, labelled as Fi in our analysis) were identified by examining the produced sketches alongside the sessions' video recordings. Affinity diagramming was used to categorize the elements into themes that were present in all of the designs.

5.3 Analysis

5.3.1 Content Themes Presented in Design Components

Functionally speaking, the components could be categorized into three groups, with many components providing functionality from multiple categories. In support of our hypothesis, each of the functional categories in fact did assist in the performance of one sub-task depicted in Study 1's task taxonomy (Figure 1): (1) Elements showing useful shared and unshared information about EPK ("the what"), (2) Elements assessing communication acts ("the how"), and (3) Tools for differential

interpretation.

5.3.2 Showing Useful Shared & Unshared Information about EPK

Some components in this category presented important qualitative data such as management outlook, past and current guidance, and the organization's performed strategic and corporate actions, in a bullet list to make them easier to access (D1, D3).

Other features augmented disclosed information with additional data to enhance their interpretability (D1, D3). A Cashflow Overview feature visualized components of key performance figures such as cashflow (D3). The visualization showed a graph of the historic and forecasted values of key figures and their components, allowing the user to rapidly uncover the causes of change.[2]

This feature also facilitated the rapid comparison of quantifiable outcomes across companies, which are often calculated differently within or between industries. Although much of the information presented in D3's tools exists in products such as Bloomberg Terminal, they could not all be accessed simultaneously, forcing IMPs to often collect the information into a spreadsheet for analysis.

Another component in this category provided additional detail about the company's productions facilities, enabling the user to better interpret the consequences of production stoppages on the company's future profitability (D1). The component visualized different production facilities on a zoomable map, annotating each facility with its production capacity as well as production costs per unit, and highlighting the facilities that were affected by a production stoppage (F1a in Figure 4). D1's design allows users to rapidly assess the extent to which the company's profits would be affected. Although information about production stoppages also exists in products such as Bloomberg Terminal, it is typically dispersed amongst multiple text documents. Extraction techniques are required to populate such visualizations, which are now becoming possible given machine comprehension's success under similar scenarios (Wang et al., 2017a,b) (F1b).

Yet another designed component, named "Sensitivity Analysis" (see Appendix, Figure 2), augmented the organization's guidance about future outcomes (D1). Each predicted outcome (e.g., revenue), is based on assumptions made by the company (e.g. oil prices, exchange rates), which may not be reasonable from the reader's perspective. To alleviate this, the "Sensitivity Analysis" tool extrapolates the provided guidance to a range of alternative values, enabling readers to inspect the sensitivity of guidance to the company's key assumptions (F2a).

Sensitivity analysis is currently performed manually by junior analysts on Wall Street (D1). To automate this, one needs to build a model of the company's outcomes (e.g., revenues) as a function of one or more assumed variables (e.g. oil prices, exchange rates). Such models are currently built using spreadsheets. As participants in both studies have indicated, the information needed to build these models can be found verbatim in earnings call transcripts as well as the company's filings. This is also true of the map widget discussed above. What is missing in current tools is the effective visualization designed by D1, which requires the use of machine comprehension techniques (F2b) to be fully automated. All four of the participants stressed time pressure as the motivation for automation, but not accuracy of the resulting computations, nor recall rates of important information from source material.

5.3.3 Elements Assessing Communication

Visualizing sentiment in transcripts was envisioned as one of the tools for assessing communication (D1, D5). D1 designed widgets to visualize the sentiment score of the transcripts along with the distribution of sentiment scores from the company's previous communications using a box chart (F3a). The widgets also facilitated the comparison of sentiment information across companies in the same industry (see Appendix, Figure 3). Moreover, the widget allowed users to track the evolution of sentiment over time using a popup line chart (F3a), allowing them to account for "sentiment inflation" in companies exhibiting the common habit of finding the "silver lining in the cloud" (D1). The IMP can use the widget to determine whether the studied company beats its competitors or peers in positive sentiment. Interestingly, D1's sentiment visualization included two copies of the described widget, one representing sentiment in the prepared remarks and another for the Q&A content.

The central feature of D4's prototype compared expressed sentiment across time on a hawkish-

[2] "*this quarter EBITDA went down, why was it? was it because your revenue went down... was management taking out some money um what was it...*" (D3).

dovish scale (F3a). Depending on the value of sentiment, the system would also recommend a set of trades to the end user. D4's prototype contained four tables: a table containing hawkish terms in the transcript along with their frequencies, a table containing dovish terms along with their frequencies, as well as two similar tables representing frequencies in only the most recent transcript (F3a). The component showing the change in sentiment on a scale would be used most often, with the term tables being used only when an explanation was needed about how the system arrived at the computed sentiment. This would work naturally for current sentiment analysis algorithms, which assign sentiment based on frequencies of sentiment-bearing words (e.g., Pennebaker et al., 2007). Since IMPs often access sentiment to appraise short-term investment opportunities, successful sentiment analysis technology used in automatic trading algorithms (e.g., Bollen et al., 2011; Kazemian et al., 2016; Zhang and Skiena, 2010) is an excellent candidate to provide the accurate sentiment scores needed to populate these widgets (F3b).

Participants also designed tools to support mining information from the Q&A sections. D3's design allowed for more rapid access to Q&A content by initially hiding the answers in order to quickly scan all the questions at once before choosing which answer(s) to view. D1's designs aimed to characterize how speakers responded to questions, by measuring: the amount of time taken by speakers before formulating a response,[3] the average length of answers relative to with the company's peers,[4] and the percentage of responses that resulted in the disclosure of specific facts or quantifiable information.[5] Companies that have direct and quantifiable responses are viewed by the market as more certain investment opportunities (D1). The goal of these widgets, using similar visualizations to D1's sentiment widgets (F4a), is to convert qualitatively expressed metadata about a speaker's communication tactics into a quantitative score depicting the investment attractiveness of the company.

Although D3's design does not require the use of NLP, D1's three widgets do. For these widgets, using established tools such as speaker segmentation (Budnik et al., 2016) and speech alignment (Goldman, 2011), each transcript portion can be aligned with its underlying audio signal, and to also calculate average duration of responses. Disfluency detection (Liu et al., 2006) can help find the time taken by disfluencies before a coherent response is produced. Thus, current NLP technology can be used to populate D1's first two widgets (F4b). However, to the best of our knowledge, state-of-the-art NLP tools such as answer selection (Rao et al., 2016) or fact extraction (Pasca et al., 2006) have not yet been evaluated in scenarios similar to the third widget.

An important observation can be made about the tools designed so far. With the aid of visualization and NLP techniques, these tools extract information from examined transcripts, augment it with information from other sources, and present it to users visually. All users noted the time savings accrued in comparison to manually reading and producing similar visualizations with current software. No user noted that such tools may also help them because they are more accurate or methodical in their detection of implicit information such as sentiment or communication tactics into the analysis. Participants, when questioned, saw no particular advantage to cognitively offloading to a computer the interpretative or analytic activity that followed upon the information gathering sub-tasks because: 1) they themselves were highly effective at doing it, whereas 2) a computer might make mistakes.

Although all users enlisted the aid of NLP to populate their visualizations, in one case, the use of NLP even here was doubted. D2 reluctantly considered automatic highlighting to mark a transcript's salient parts, but indicated that this amounts to the system thinking on her behalf. There are areas of NLP such as summarization and information extraction that could indeed be used to highlight text, but this falls within the purview of interpretation, whereas parsing complex syntactic constructions in free-flowing text to identify objective quantities was considered more reliable. D2 remarked that she was only willing to use automated highlighting when extreme time pressure prevented her from reading the entire transcript. Observations such as this suggest that IMPs do not embrace NLP when it removes their own decision-making agency.

[3] "*did the candidate... dilly-dally a lot or was he very forthcoming . . . [with] answers*"

[4] "*what was the average length . . . usually they give longer answers when they don't have an answer*"

[5] "*what percentage of the time was he BS-ing and what percentage of the time was he giving a clear direct answer*"

This, together with the prior important observation, highlights a key theme running through all the features Fi mentioned: our participants view such designs and the possible underlying NLP technology simply as time-saving tools, and not tools that may enhance discovery or interpretation. This suggests the need to preserve decision making agency when using software that provides assistance during information-seeking tasks - software that must be transparent in the use of the NLP tools.

5.3.4 Tools for Differential Interpretation

Most tools designed by participants compared actions and outcomes to those of the past, to those of the institutions' peers, and to published projections. Although the described comparisons are not supported for natural language data in current analysis software, comparing curated, quantitative data to historic values or projections is well-supported in current products such as Bloomberg Terminal or FactSet.

Similarly, many components in the sketched prototypes included easy-to-access links to related research reports that complement users' analysis of market perception and anticipate market reaction to transcript content. Again, links to research reports are available in software such as Bloomberg Terminal and FactSet, but are not integrated with tools for the qualitative analysis of transcripts.

6 Conclusion: HCI-NLP Co-Design

Our studies have revealed many information practices of IMPs. Several are not well supported by existing software marketed to IMPs, partly due to the complexity of the processes that IMPs typically carry out (Figure 3). Our studies suggest that IMPs need more and more detailed visualizations than what currently exists in their software. They also suggest that NLP technology will be most enthusiastically received when it is bundled with visualization techniques as an extraction mechanism that populates visualizations in such a way that preserves the IMPs' sense of agency over decision making proper.

An extensive taxonomy (see Appendix, Figure 4) synthesizes our findings, capturing the requisite high-level information practices, the software functionality that would serve the typical cases envisioned by analysts, the available NLP tools to support this, and the common UI elements in which these tools can be encapsulated (as drawn or described by the PD workshop participants). Note that the "desired" functionality here consists mainly of very close variants of problems that have already received considerable attention from the NLP community, such as aspect-based sentiment analysis, but recast as more vertical tasks that IMPs will assign value to. Without that domain-specific context, the more abstract tasks that NLP researchers generally ascribe to their own work are more likely to be construed by IMPs as a combination of trite and insufficiently nuanced, because their own vocational expertise is more highly prized by them than the general cognitive mechanisms that the AI community focus on in popular representations of their accomplishments.

Finally, this investigation has shown the importance of conducting user studies to assess the usefulness of technology (in this case, for supporting ISPs) alongside the development of the technology. Blindly pursuing a "deep-learning" crusade for general intelligence is unlikely to result in widespread adoption of black boxes, even at the level of speech recognition and sentiment analysis, by IMPs. To some extent, this is a Catch-22. Their current software does not incorporate advanced NLP, and so IMPs are unaware of its potential specific to their needs, and thus they are resigned to reserving agency over even the minutest of their decision-making tasks, which software vendors capitulate to in the design of their products. For their real potential to be embraced by IMPs, NLP tools need to be embedded in designs and visualizations in a manner that emphasizes superior extractive accuracy and generative quality over the time value of using the tools, while maintaining a sense of ISP agency.

References

B.S. Bernanke and K.N. Kuttner. 2005. What explains the stock market's reaction to Federal Reserve policy? *Journal of Finance*, 60(3):1221–1257.

E. Bertini and D. Lalanne. 2007. Total recall survey report. Technical report, University of Fribourg, Department of Computer Science.

H. Beyer and K. Holtzblatt. 1999. Contextual design. *Interactions*, 6(1):32–42.

J. Bollen, H. Mao, and X. Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

M.J. Bouwman, P. Frishkoff, and P.A. Frishkoff. 1995. The relevance of gaap-based information: a case

study exploring some uses and limitations. *Accounting Horizons*, 9(4):22.

V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

M. Budnik, L. Besacier, A. Khodabakhsh, and C. Demiroglu. 2016. Deep complementary features for speaker identification in tv broadcast data. In *Proceedings of ODYSSEY*, pages 146–151.

E. Cambria, B. Schuller, Y. Xia, and C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

M. Flood, H.V. Jagadish, and L. Raschid. 2016. Big data challenges and opportunities in financial stability monitoring. *Financial Stability Review*, 20:129–142.

J.-P. Goldman. 2011. Easyalign: an automatic phonetic alignment tool under praat. In *Proceedings of INTERSPEECH*, pages 3233–3236.

A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. 2004. Memory cues for meeting video retrieval. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 74–85.

S.K. Jauhar, P.D. Turney, and E. Hovy. 2016. Tables as semi-structured knowledge for question answering. In *Proceedings of ACL*, pages 474–483.

S. Kazemian, S. Zhao, and G. Penn. 2016. Evaluating sentiment analysis in the context of securities trading. In *Proceedings of ACL*, pages 2094–2103.

D. Lalanne and A. Popescu-Belis. 2012. *User requirements for meeting support technology*. Cambridge University Press.

Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540.

G. Marchionini. 1995. *Information Seeking In Electronic Environments*. Cambridge University Press.

A. McCallum, D. Freitag, and F.C.N. Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of ICML*, pages 591–598.

C. Milanesi. 2016. Voice assistant anyone? yes please, but not in public! Technical report, Creative Strategies, Inc.

A. Nenkova and R.J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152.

M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006. Organizing and searching the World Wide Web of facts-step one: the one-million fact extraction challenge. In *Proceedings of AAAI*, pages 1400–1405.

J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, and R.J. Booth. 2007. *The development and psychometric properties of LIWC2007*. UT Austin.

P. Rajpurkar, R. Jia, and P. Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of ACL*, pages 784–789.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

S. Ramnath, S. Rock, and P. Shane. 2008. The imp forecasting literature: A taxonomy with suggestions for further research. *International Journal of Forecasting*, 24(1):34–75.

J. Rao, H. He, and J. Lin. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *Proceedings of ACM CIKM*, pages 1913–1916.

S. Rosenthal, N. Farra, and P. Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of SemEval*, pages 502–518.

D. Schuler and A. Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

R. Socher, A. Perelygin, J.Y. Wu, J. Chuang, C.D. Manning, A.Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

E.G. Toms, L. Freund, R. Kopak, and J.C. Bartlett. 2003. The effect of task domain on search. In *Proceedings of CASCON*, pages 303–312.

P. Vakkari. 2003. Task-based information searching. *Annual Review of Information Science and Technology*, 37(1):413–464.

M. Wang, N.A. Smith, and T. Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of EMNLP/CoNLL*, pages 22–32.

T. Wang, X. Yuan, and A. Trischler. 2017a. A joint model for question answering and question generation. In *ICML Workshop on Learning to Generate Natural Language*, page 7 pages.

W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. 2017b. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of ACL*, pages 189–198.

S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg. 2002. Scanmail: a voicemail interface that makes speech browsable, readable and searchable. In *Proceedings of CHI*, pages 275–282.

S. Whittaker, J. Hirschberg, and C.H. Nakatani. 1998. All talk and all action: strategies for managing voicemail messages. In *CHI-98 Conference Summary*, pages 249–250.

S. Whittaker, S. Tucker, K. Swampillai, and R. Laban. 2008. Design and evaluation of systems to support interaction capture and retrieval. *Personal Ubiquitous Computing*, 12(3):197–221.

W. Zhang and S. Skiena. 2010. Trading strategies to exploit blog and news sentiment. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 4(1):375–378.

Appendix: Design Artefacts

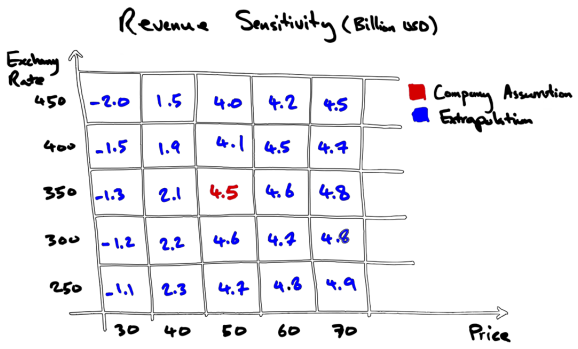

Figure 2: Sensitivity Analysis extrapolating the value of a company's important outcomes (e.g., revenue) under different assumptions about key performance factors.

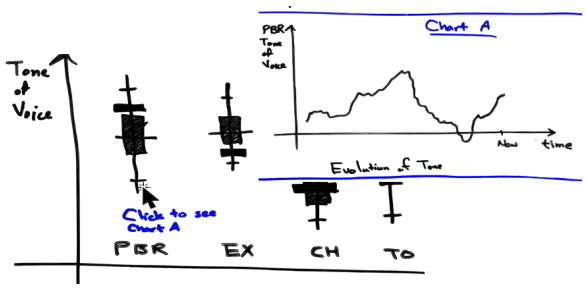

Figure 3: Components comparing the document's sentiment with the company's previous communication (using box chart and line graph), and with competitors' communications.

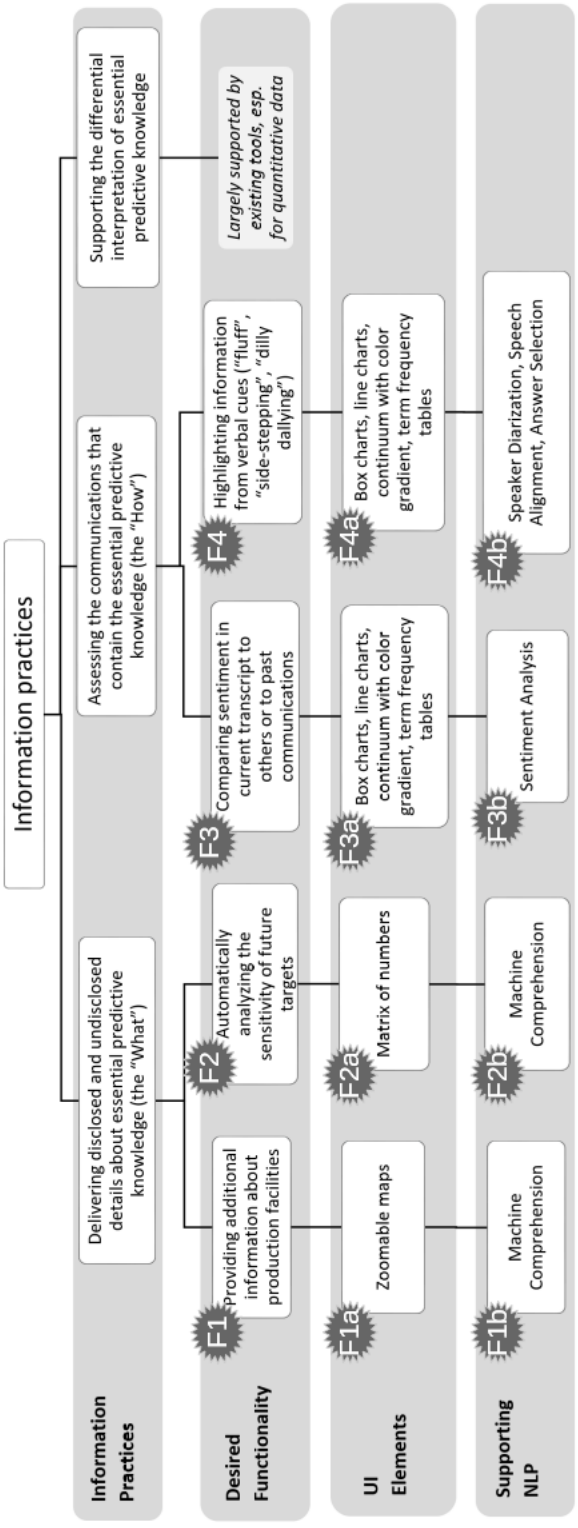

Figure 4: The proposed taxonomy, capturing high-level information practices and related software functionality, along with available NLP technology and common UI elements that can implement the functionality to support information practices.