Affect Proxies and Ontological Change: A Finance Case Study

Xiubo ZHANG[1] *Khurshid AHMAD*[1]
(1) Trinity College, Dublin
xizhang@tcd.ie, kahmad@scss.tcd.ie

ABSTRACT
Traditional sentiment analysis has been focusing on inference of the sentiment polarity using sentiment-bearing words. In this paper, we propose a new way of studying sentiment and capturing ontological changes in a domain specific context in the perspective of computational linguistics using affect proxies. We used Nexis service to create a domain specific corpus focusing on banking sectors. We then created an affect dictionary from three kinds of lexica: sentiment lexica as in the General Inquirer dictionary; news flow represented by domain entities such as financial regulators and banks; and what we call *contested* term lexica, which consists of terms whose semantic implication is inconsistent over time. Univariate and multivariate analysis techniques such as factor analysis are used to explore the relationships and underlying patterns among the three types of lexica. Analysis results suggest that citations of regulatory entities show strong correlation with negative sentiments in the banking context. Also, a factor analysis was conducted, which reveals several groups of variables in which the *contested* terms correlate with positive and negative sentiments.

KEYWORDS: sentiment analysis, affect proxy, computational linguistics, factor analysis, contested terms, ontological change.

Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pages 99–114,
COLING 2012, Mumbai, December 2012.

1 An Introduction and the Case Study

In rapidly changing environments, for example the aftermath of the 2008 credit crunch, we saw the advent of US and EU economic and financial stabilization schemes, changes in the regulatory frameworks include major revisions of existing concepts (e.g. capital adequacy), introduction of new concepts (e.g. novel regulatory pathways), and constraints on existing concepts/practices (e.g. sub-prime loans). These changes are articulated in new or revised governmental legislation and voluntary codes of practice over a period of time – there are commentaries and interpretation of these changes. All these organisations produce prodigious quantities of documents on a daily or even hourly basis and broadcast the documents using data feeds and social media; there is a concomitant flow of new and revised keywords from the compliance and regulatory agencies.

The post credit squeeze language of the regulators and that of the regulated is suffused with negative affect – indeed the terms *credit squeeze*, *credit freeze*, *zombie loans/banks* are used to express the negative evaluation of the state of leading economies and their financial institutions. Times of change invariably involve the introduction of new terms, or more importantly old terms are retrofitted with new meanings or nuances. Indeed, the early pioneers of sentiment analysis, discussed the changing language of "American values" by an analysis of changes in the language of the two major political parties in the USA – the Democratic and the Republican parties (Namenwirth and Lasswell, 1970). The authors argue that the anti-slavery party (the Republicans) became less inclusive (compared to the Democrats). This claim was based on an analysis of "inclusivity" words in the election manifestos of the two parties between 1844-1864 and 1944-1964, the authors had used the General Inquirer system and the associated lexica (Stone, 1966). This text analytic approach suggests that major changes in the attitudes within a community can perhaps be discerned by examining the choice of words belonging to domain terms (political and economic) and the affect terms (negative/positive evaluation, strength and orientation). The question we ask in this paper is this: Are the changes in attitudes related to changes in the ontological commitments of the community (e.g. from pro-slavery to anti-slavery, from pro-federation to autonomous units in (Namenwirth and Lasswell, 1970)?.

Revert to the 2008 financial crisis: prior to the crisis, there was a vocal body of opinion that was in favour of *light-touch regulation*, and compliance and governance issues were expected to be dealt with within financial institutions. Things have changed considerably since 2008 what with the ever complex national and international compliance frameworks, direct governmental management of financial institutions, and a resurgence of regulators.

One iconic term which hallmarks the 2008 crisis is *light-touch regulation*: In the decade before the crisis, the banks, the regulators, and indeed the media and governments, wished for and implemented minimal (state) regulations, self-governance, and low-level of compliance, for the financial services industry. Things have changed after the decade and *light touch regulation* will be giving way to *abundant regulation*! A survey of associated sentiment with *light touch regulation* using Google search engine and selecting one of first 10 most relevant documents for the term sampled every two years from 2002 shows the contested nature of the headword – regulation. Furthermore, the affect terms associated with light touch regulation for sentiment evaluation changed polarity – from negative to positive (Table 1).

A large number of US government agencies and professional bodies, around 12 at the last count, are involved in (a) monitoring financial institutions for compliance with existing laws and codes of practice; (b) producing regulations and regulatory frameworks; and (c) examining

Date	Headline and Source	KWIC
18 Nov 2002	average banking cost* (euro/year) [British Bankers' Association]	Historically *light touch regulation* ... has driven banks to be more efficient
8 Jul 2004	House of Commons - International Development - Written Evidence	We welcome the principles of *light touch regulation* ...
4 Dec 2006	SELLING THE CITY SHORT? [Open Europe Think Tank]	Bermuda ... enjoy *light-touch regulation* ...
22 Jun 2006	Gordon Brown's Mansion House speech | Business [Guardian.co.uk]	... the future, advance with *light touch regulation*, a competitive tax environment ...
17 Oct 2008	The days of light-touch regulation in the City are over,' warns head ... [Daily Mail]	The City watchdog ... warned the days of *light-touch regulation* ... are over.
12 Mar 2010	(UK FSA) calls time on FSA's "light touch" regulation – [Telegraph]	(UK) will drop its long-held commitment to ... *"light-touch" regulation*
3 May 2012	Switzerland says goodbye to light touch regulation [Reuters. Blog]	Switzerland says goodbye to *light touch regulation* ...

Table 1: Changes in the polarity associated with a contested term light touch regulation between 2002-2012

the governance of financial institutions. In itself, the involvement of agencies in (a)-(c), appears a normal, routine matter in that business-critical institutions should by default comply with laws, have good regulatory framework, and demonstrate exemplary governance. The fact that concepts related to *compliance*, *governance* and *regulation* are still being contested in the media is an interesting manifestation of regulatory change from *light-touch regulation* and/or *self-regulation* to something else, e.g. *smart regulation*. The evidence of this continuing debate can be perhaps seen in news reports relating to the key financial institutions – the *banks* and its regulators.

It appears that a major shift in (inter-)national policies regarding an area of human enterprise, that is a major change in the ontological basis of the enterprise, is accompanied by changes in the use of domain specific terms including named entities in the domain, changes in evaluation of the domain specific terms through a change in associated affect terms, and changes in what we call *contested* terms. Contested terms generally include terms related to the basic operation of an enterprise. For instance, banks to have to comply with existing law, banks should have transparent governance structures, and banks have to be regulated well. But the question is to what extent and by whom: *lightly* by the banks themselves or *strictly* by the regulators.

It is important to note that affect can be expressed at three different levels pragmatic description: First, the number of news stories in a fixed interval of time can be used as a measure of affect evaluation – the so-called news flow is an important affect proxy. Second, the changes in the distribution of the contested terms can also be used as a proxy for changes in affect or sentiment. And, third, the distribution of the domain independent affect terms, if computed accurately and with appropriate degree of disambiguation, can be used as a more direct measure of sentiment.

All three measures of affect or sentiment, news flow and the distribution of the contested and evaluation (positive/negative affect) terms closely follow the boom and bust within the world economic system.

Sentiment analysis is an interdisciplinary enterprise involving computer scientist, linguists, literature experts, cognitive psychologist and domain experts. One can argue that sentiment

Banco Santander	BNP Paribas	Deutsche Bank	Mitsubishi UFJ
Standard Chartered	Bank of America	Citibank	Goldman Sachs
Mizuho Financial	State Street	Bank of China	Commerzbank
HSBC	Morgan Stanley	Sumitomo Mitsui	Bank of New York Mel.
Credit Agricole	ING	Nordea	UBS
Banque Populaire	Credit Suisse	J.P. Morgan Chase	RBS
UniCredit	Barclays	Deixa	Lloyds
Societe Generale	Wells Fargo		

Table 2: 30 named entities used in the corpus design

analysis encompasses computational linguistics and has psychologists and domain experts additionally. In this paper, we look at the analysis of sentiment by looking at dictionaries compiled by psychologists and linguists. We begin by describing the design and implementation of our corpus (c. 12 million words) and a specially designed lexica for dealing with affect and affect proxies in Section 2. This is followed by a description of the method we used. The results section comprises the results of univariate and multivariate analysis reported in Section 4 and then we conclude.

2 Design of the Corpus and *Affect* Lexica

2.1 Corpus Design

Our analysis is targeted on a corpus comprising news articles related to 30 major banks around the world as shown in Table 2. We have used the Nexis database of news and related documents to collect the bank-related news over an 11 year period (2001-2011); our choice of this news source was motivated by the availability of rich meta-level information that is used to annotate, and subsequently retrieve each news document in Nexis. Our data set contains 22 sub-corpora each comprising six months of news. For each of the six month period, a query is issued to search the articles using the bank names as keywords over a pre-defined set of sources called "Major World Newspapers (English)" within Nexis: the top 1000 most relevant articles returned by the search are retained. We did not restrict our search to a particular news paper because we believe the overall prospect of the banking sector might be better captured in a global perspective. Our use of the relevance metric provided by Nexis was motivated by the thought that the sampling process should remain consistent and largely free of any biases or framing during manual selection of media sources [1].

The meta-level information was extracted automatically from raw text downloaded from Nexis data base [2]. The information can be used to extract the date of publication and news source. The publication dates come with the documents allow us to aggregate the daily news stories into lower frequency data – weekly, monthly or yearly. The news source information help us to use all news from all sources or to dis-aggregate the news according to sources. The time period aggregation and news source dis-aggregation can help capture the effect of time scale or the news source.

[1] The duplication-removal option in Nexis was used, the actual amount of articles obtained per search is usually less than 1000, but as the occurrences of duplication can be regarded as random events, we believe the corpus created this way is consistent and representative.

[2] The raw documents downloaded are unstructured and a Java program was written to extract the meta-data annotation from the text, which contain the date on which the news was published as well as the source of the news.

Title	Articles	Tokens	Average Article Length
Year: 2001	1890	1157837	612.61
Year: 2002	1857	877891	472.75
Year: 2003	1794	846660	471.94
Year: 2004	1886	939377	498.08
Year: 2005	1593	1002570	629.36
Year: 2006	1953	1271229	650.91
Year: 2007	1918	1231522	642.09
Year: 2008	1879	1416622	753.92
Year: 2009	1771	1285902	726.09
Year: 2010	1791	1247322	696.44
Year: 2011	1797	1254569	698.15
Total	20129	12531501	
Mean	1830	1139227	622.94

Table 3: Yearly breakdown of the corpus

For the 30 banks, Nexis yielded 20129 relevant articles over the 10 year period, which enabled us to build a specialist corpus of 12.5 million words with a mean number of 1830 documents per year and an average length of 623 tokens (Table 3).

2.2 Lexica Design

Three lexica were used in our analysis:

2.2.1 Domain Lexica: The Financial Regulator / Banking Dictionary

The motivation behind the creation of this dictionary is the assumption that frequent mentions of financial regulators might imply the existence of inadequacy in regulatory enforcement, making the announcement of such agencies a proxy to negative sentiments. The dictionary contains 4 categories: US Regulators, UK Regulators, and Eurozone Regulators, with the fourth category containing a list of prominent banks, as nominated in (Forbes, 2011).

2.2.2 Affect Lexica: Harvard Dictionary of Affect

Harold Lasswell (Lasswell, 1948) has used sentiment to convey the idea of an attitude permeated by feeling rather than the undirected feeling itself. Such analyses of documents in the political and economic domain were boosted by the use large digitized dictionaries, notably the *GI Dictionary* also known as the *Harvard Dictionary of Affect* which formed the backbone for the *General Inquirer* system (Stone, 1966). The *GI Dictionary* currently comprises over 11,000 words. Each word in the Dictionary has one or more "tags". Some of these tags refer to the connotative meaning of the word, whilst others to its cognitive orientation, and some to the belongingness of the word to a specific domain. The words in the Dictionary have between one and 12 of the 128 "tags". These tags are divided into 28 or so categories.

The original, and linguistically rather dated *Harvard Dictionary of Affect*, has been used in our analysis purely for the evaluation affect words – negative and positive. Note that the *Harvard Dictionary* has affect tags associated with domain specific terms which can be misleading when an affect count is carried out. For example, *Harvard* has the word *competition* tagged as negative evaluation word, and the words *share* and *company* as positive evaluation words. This may

Keyword	Identity	Opposite	Remark
regulation	*control*		direct synonym of regulation
	supervision		synonym of synonym of control
		relinquishment	antonym of synonym of control
		disorganization	direct antonym of regulation
	coherence		antonym of antonym of disorganization
		dissolution	synonym of antonym of disorganization

Table 4: Semantic identity and opposition of the contested term *regulation*

have been true in everyday language of the 1940's and 50's (the times when the Dictionary was compiled), but today these words are used as keywords in the domains of economic and finance.

The system used in our analysis has been so designed that when a token from a given document is analyzed for its belongingness to affect categories, and if the token is found in a domain specific dictionary then the system ignores the affect category.

2.2.3 Contested Term Lexica

The contested terms lexica are a hybrid of domain specific terms and words in an affect lexicon. We use three ontological primitives – *compliance*, *governance* and *regulation* and populate the hybrid lexicon with synonyms and antonyms of each of the three primitives. The hypothesis we wish to test is the identity terms, especially synonyms of a given ontological primitive will reinforce messages related to the unit whilst the opposition terms, especially antonyms, will create a negative empr? of the primitive.

This population process can be accomplished by traversing a general thesaurus or a thesaurus similar to *WordNet* "intelligently" and to scrape data from synonymous and antonymous relationships between synsets as demonstrated in a variant of *WordNet – SentiWordNet* (Baccianella et al., 2010). For our study, we use a general language thesaurus that is freely available on-line at this time[3].

The dictionary is populated using an expansion algorithm, which starts with the three keywords, *governance*, *regulation* and *compliance*. The algorithm then iteratively populates the dictionary by assigning direct as well as indirect synonyms and antonyms of the three seed words to appropriate. An example expansion from the seed word "regulation" is shown in Table 4. The table demonstrates how the affect category *regulation identity* and *regulation opposition* are populated using synonyms and antonyms of the seed word regulation. A synonym of a word is considered to have the same affect evaluation as the word while an antonym of a word has the opposite affect evaluation. This rule is also applied iteratively to synonyms and antonyms of the seed word as well.

2.2.4 Merging Strategy

The above three dictionaries are merged together to form a single affect dictionary to be used in the analysis. After the merge, the set of categories to which a word belongs is the union of

[3]The thesaurus used in our study is an on-line thesaurus at *http://thesaurus.com*. Synonyms and antonyms that are shorter than four characters were excluded from the lexica to avoid common close-class words.

	Relationship	
	Identity	Opposites
Governance	337	34
Regulation	370	110
Compliance	185	291
Affect Evaluation	4923	6870

Table 5: Lexica statistics

the original three sets of categories the word is associated with. A summary of the statistics of the lexica is shown in Table 5.

3 Methodology

We employ a methodology similar to vector space model, where each document in the corpus is represented by a vector of N dimensions. The difference lies in the semantics of the space – the vectors measure affect strength rather than word frequency.

The merged dictionary created as described in Section 2.2 is used to transform documents to vectors. The dictionary is essentially a many-to-many mapping between words and dictionary categories, where each word in the dictionary is associated with one or more categories. The documents in the corpus are then converted into vectors where each element in a vector corresponds to the relative frequency of a specific affect category in that document.

The relative frequency of a category is computed as the sum of the absolute frequencies of words belonging to the category over the total number of words in the document. Formally, the strength of the category C in document D is given as Equation 1.

$$\text{AffectStrength}(C, D) = \frac{\Sigma_{d \in D} |\{w | w \in d \wedge w \in C\}|}{\Sigma_{d \in D} |\{w | w \in d\}|} \tag{1}$$

The next phase of the method is to aggregate the document vectors based on the time of publication of the document. Vectors that associate with documents from the same time period of interest are added together to form a single vector representing the affect characteristics of the specific period. The result of this aggregation is a multivariate time series. For our analysis, the documents are grouped into a monthly scale.

4 Analysis and Results

4.1 Univariate Analysis

4.1.1 News Flow

In text analytics in general, and in sentiment analysis in particular, news flow, typically number of relevant articles published in a given time interval, is used as a sentiment or affect proxy – see for instance Kim and Barnett's work in international marketing (Kim and Barnett, 1996), Cain's in political science (Cain, 2012), and Hafez and Xie's in finance (Hafez and Xie, 2012). A study of the aggregated monthly news flow in our corpus shows that the coverage of banks in news media during the three periods (c. 2001, 2002-2005, 2006-2011) is different: below the mean news flow in the boom period and above the mean during crises.

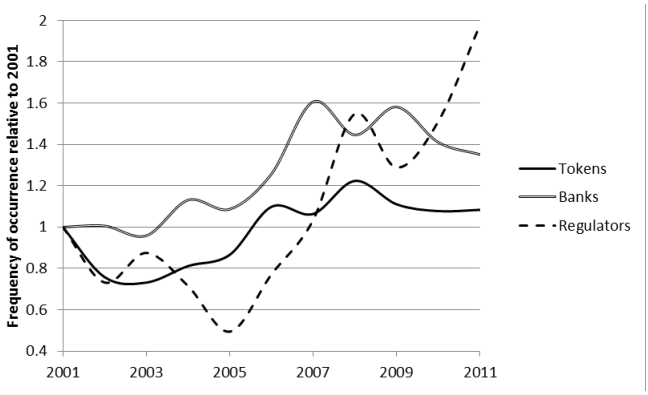

Figure 1: Annual frequency of total number of tokens and two named entities, banks and regulators, relative to 2001. (2001 frequencies – $N_{tokens} = 1157837$, $N_{banks} = 11326$, $N_{regulators} = 593$)

Three points to note here: (1) that following the dotcom boom (c. 2001) and until the first signs of the credit crunch (c. 2007), the yearly average word count, 500 tokens/news story, was much lower when compared with the pre-dotcom period, c. 600 tokens/news story, and the post boom period (c. 700 tokens/news story 2008 – to date); 2) there was a significant increase on the average length of articles talking about banks starting from 2005 and again another boost around 2009. The increase in the average length of the article pertinent to banks might be a result of the shift of public attentions towards banking sector during the financial crisis. A further Augmented Dickey-Fuller test for unit root shows that the series is non-stationary, which implies such shift must be structural rather than by chance.

The average frequency of the domain primitives, banks and regulators, i.e. the use of the names of banks and the regulators, in our 12.5 million word corpus, is 1.25% and 0.06% respectively. The annual distribution of the total number of tokens in our corpus is similar to that of the frequency of use of bank related tokens – higher in bust periods and lower in the boom periods (Figure 1); this is not surprising in that the corpus was created using the names and abbreviations of banks listed in Table 2. However the asymmetry in the distribution of bank related tokens and regulator related tokens is interesting in the sense that regulator related terms showed a drop in pre-2005 period but then there is an almost linear increase in the citations of regulators. Overall there is a 2.61% per annum increase in the regulator-related tokens whereas that of banks is 1%; these increment figures were computed using the historical return of the frequencies (logarithm of the ratio of this year's frequency of usage over last year's).

4.1.2 Contested Term Flow

The average annual frequency of the tokens related to the contested terms, *compliance*, *governance* and *regulation*, is 0.12%, 0.71% and 0.23% respectively in our 12.5m token banking corpora. The peak usage of three terms was in 2004 (*compliance*), 2006 (*governance*) and 2008

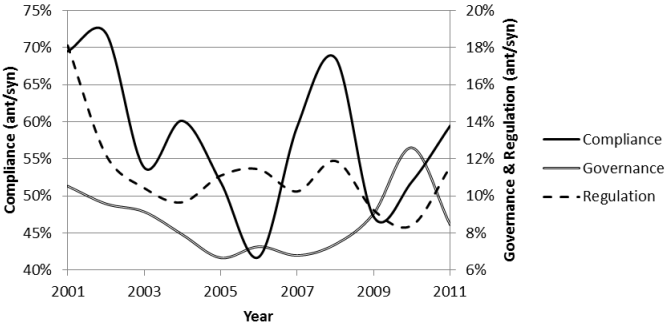

Figure 2: Annual frequency of total number of tokens and two named entities, banks and regulators, relative to 2001. (2001 frequencies – $N_{tokens} = 1157837$, $N_{banks} = 11326$, $N_{regulators} = 593$)

(*regulation*). The maximum usage of the three terms is within two standard deviation of the mean for each of three contested terms over the 10 years (2001-2011), showing a degree of stability of usage and perhaps our choice of the term and their synonyms and antonyms.

However, the distribution of the tokens related to synonyms and antonyms of the each of the contested terms is asymmetric, with synonyms being more widely used than the antonyms in each year of our observation. One can see the same effect in the language of general purposes: we have looked at the very broad coverage Google search engine and the more restricted American National Corpus (comprising 450 million words used in newspapers, fiction and other texts published during 1990-2012) and found a similar asymmetry in the distribution of a token and its antonyms.

What is interesting is the change in the asymmetry ratio over time: The average asymmetry for the compliance-related synonyms and antonyms is 58%, however, the maximum is around 70% (in 2002 and 2008) with a minimum of 40% in 2006. The ratio for the other two contested terms, *governance* and *regulation* is around 10% for every synonym used 10 times the antonym is used only once. The ratio again changes over our observation period (2001-2011) with a peak (18%) in 2002 (and minimum of 8% in 2012) for *regulation*. The asymmetry ratio for *governance* has a peak (13%) in 2010 (and a minimum of around 6% in 2005). The term compliance appears to be more contested than the other two (Figure 2).

4.1.3 Sentiment Flow

Typically, in financial studies, the negative sentiment has been found to be the causal variable that impacts the return on investment: Tetlock and colleagues have looked at a restricted set of tokens associated with negative affect and found a correlation between the variance in the frequency of such tokens and risk on the return. The author has argued that “high values of media pessimism induce downward pressure on market prices” (Tetlock, 2007): by media he means a financial gossip column in the Wall Street Journal and market “prices” refers to the logarithmic return of the daily values Dow-Jones Industrial Average Index. Elsewhere, we have

noted that the historical volatility (proxied as standard deviation) of a negative affect time series (Devitt and Ahmad, 2008).

We have looked at the annual frequency distribution of the negative and positive affect tokens in our corpus, together with the logarithmic value of the ratio of the frequency of the current year and the previous year – usually called return. The asymmetry of the average value of the relative frequency, over the 10 years of our coverage, for negative and positive affect is 2:3, the values over the 10 year period for both affect series is within two standard deviation of the mean. However, the average value of return is 0.1% for negative affect but -0.02% for the positive affect: the volatility for negative affect is 5% whereas for positive affect 2% only. The differences are even starker when we divide the series of affect values in "boom" years (2002-2006) and "bust" period (2007-2011). The negative affect decreases overall in the boom period and vice-versa for the positive affect; contrariwise is the case for the bust period. The volatility of negative sentiment is much higher in the bust period.

4.2 Multivariate Analysis

The variables we have discussed thus far in the context of changing nature of(world-wide) financial systems dealt with three inter-related categories of tokens: domain specific tokens, contested tokens, and affect tokens. We have chosen to study not only the tokens but have constructed a polar space where we have (a) banks and their regulators; and (b) not only we have looked at contested issues, compliance, governance and regulations, but also at the identities and opposites of these tokens. In this section we will look briefly at the correlation between the distribution of the terms and attempt to identify combinations of these categories account for the variance of frequency distribution of tokens within the categories.

4.2.1 Correlation Analysis

We have looked at the correlations between the three categories of tokens and correlations across the categories. Correlations at 99% significance level appear between (a) negative affect tokens and (synonyms of) regulators, the correlation is positive, and (synonyms) of compliance anti-correlate with negative affect; positive affect tokens correlate with (synonyms of) governance and 90% significance level with the identities of compliance and regulation; (b) the frequency distribution of regulators is correlated with compliance; (c) the identity and opposites of compliance are positively correlated as are those of governance; the latter is correlated with the synonyms of regulation. (See Table 6 for details).

4.2.2 Factor Analysis

Pair-wise correlations in some cases help to identify relationships between two variables. However, the method makes it somehow difficult for human to gain insight into data, especially in terms of relationships between groups of variables. To obtain a better understanding of the overall picture between the variables, we performed a factor analysis on the data to explore latent patterns that may dictate the observed behaviors of the affect categories[4]. Factor analysis was initially developed in the discipline of psychology as a statistical approach to explain correlated variables using reduced number of "factors". In our study, we are mainly interested in its capability of grouping variables so that they can be better understood.

[4]The principal component analysis and factor analysis was done using Minitab 16.

			Affect		Domain		Contested					
			Negativ	Positiv	Banks	Regulators	Compliance		Governance		Regulation	
							Iden.[a]	Oppos.[b]	Iden.	Oppos.	Iden.	Oppos.
Affect	Negative		-									
	Positive		-0.039	-								
Domain	Regulators		0.543***	0.172**	-							
	Banks		0.033	-0.047**	-0.026	-						
Contested	Compliance	Iden.	-0.325***	0.175**	0.062	-0.266***	-					
		Oppos.	-0.089	0.036	0.061	-0.210**	0.322***	-				
	Domain	Iden.	-0.069	0.290***	0.165***	-0.017	0.188**	0.210**	-			
		Oppos.	-0.032	-0.092	-0.065	0.112	-0.106	-0.207**	-0.275***	-		
	Regulation	Iden.	0.062	0.212**	0.158*	0.011	-0.032	0.186**	0.288***	-0.013	-	
		Oppos.	0.190**	-0.139*	-0.019	0.060	-0.127*	0.115	0.198**	-0.081	0.142*	-

[a] Identity
[b] Opposite
*** $p \leq 0.01$
** $p \leq 0.05$
* $p \leq 0.15$

Table 6: Pair-wise correlation matrix and associated correlation significance

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Communality
Negativ	-0.83	0.00	0.14	-0.12	-0.06	0.72
Regulators	-0.82	-0.08	-0.07	0.02	0.09	0.69
compliance+[a]	0.58	-0.21	-0.18	-0.34	0.05	0.53
Positiv	0.02	-0.75	-0.42	-0.10	-0.06	0.75
regulation+	-0.01	-0.68	0.40	0.27	-0.16	0.73
governance+	0.06	-0.65	0.21	-0.37	-0.04	0.61
regulation-[b]	-0.17	-0.03	0.81	-0.08	0.05	0.69
compliance-	0.40	-0.18	0.42	-0.38	-0.04	0.51
governance-	0.01	0.04	-0.07	0.86	0.06	0.75
Banks	0.01	-0.12	-0.04	-0.06	-0.98	0.98
Variance	1.87	1.55	1.27	1.25	1.01	6.94
Var	0.19	0.16	0.13	0.13	0.10	0.69

[a] "+" denotes "identity"
[b] "-" denotes "opposition"

Table 7: Factor loadings from factor analysis

Firstly, a principal component analysis was carried in an attempt to determine the number of factors that would appear in the factor analysis. The result indicates that the first five factors combined explain 69 % of the variances, while the contribution of including the sixth factor is negligible. The factor analysis was then carried out using 5 factors on 10 variables: two variables each for both affect and the domain categories and two for each of the three contested token categories. The resulting factor loadings are rotated using Varimax Rotation for better interpretability. A total of 69% of the variances are explained by a combination of five factors as expected from the previous principal component anlaysis. The variables are explained fairly well, with seven of them having more than 65% of their variances explained by the factors (Table 7).

We then tried to interpret the factors by labelling them with semantic descriptions.

Compliance Factor *compliance+*[5], *compliance-*, *Negativ* and *Regulators* all have strong loadings on Factor 1, where compliance topics load to the opposite of Negative sentiment and regulator references. This conforms to what we observed in the correlation matrix in the previous section, where *Negativ* positively correlates with regulators and the compliance terms negatively correlates with *Negativ* as well as references to regulators. We suggest that this factor to be labeled as "Compliance Factor".

Positive Factor *regulation+*, *governance+* and *Positiv*, as we can see from the factor loading table, load heavily on Factor 2. Considering the supporting nature of the *regulation+* and *governance+* variables, we believe it makes sense to label Factor 2 as "Positive Factor".

Regulation Factor Factor 3 loads heavily on regulation, *regulation-* together with *compliance-*, and to the opposite of *Positiv* category. This could suggest that Factor 3 is related to the concept of regulation and compliance, while the concept generally occurs in a non-positive context. Therefore we suggest that Factor 3 be labeled as "Regulation Factor".

[5] *compliance+* denotes the identity concepts of *compliance* while *compliance-* denotes the opposition concepts of *compliance*. The same notion is applied to *governance* and *governance* to keep things concise.

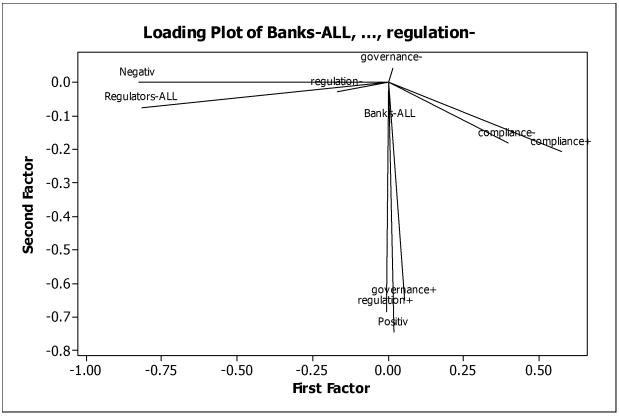

Figure 3: Factor loading plot

Cluster No.	Variables
1	*Banks-ALL*
2	*Negativ*, *Regulators-ALL*
3	*Positiv*, *compliance+*, *compliance-*, *governance+*, *regulation+*
4	*governance-*
5	*regulation-*

Table 8: Variable clusters

Governance Factor The category of *governance+* predominates Factor 4. The *governance-* category, however, loads to the opposite of *governance+*. This suggests that the discussions of governance are polarized – the contexts where governance are mentioned are either supportive or non-supportive of the governance concept. Therefore, we suggest that the Factor 4 be labeled as "Governance Factor".

Bank Factor Factor 5 is almost entirely dedicated to the citations to banks, hence we named it "Bank Factor".

Figure 3 shows the plot of the variables against the top two factors that explained the variances most, giving an intuitive representation of the distribution of the loadings. It can be seen fairly easily that the variables form three clusters. Following this intuition, we conducted a further analysis in which the variables are clustered according to their correlations[6]. Five clusters are identified and reported in Table 8.

It is worth noting that factor analysis only reveals correlations rather than casual relationships between the variables. In our case, the factors could be interpreted in two different ways. First, it could be argued that the sentiment variables are the "consequences" while the domain ones

[6]The analysis is performed using Minitab 16's "Cluster Variables" function.

are the "causes". For instance, in Factor 1, it might be reasonable to say that the contexts in which the regulators were cited are mostly negative in sentiment. This interpretation conforms with the conventional expectation from sentiment analyses, where we learn about the polarity of opinions with regard to certain topics. The second perspective of seeing the factors are to think the domain and contested variables as "proxies" or "indicators" of sentiment. Again, for Factor 1, it may be inferred that excessive citations of financial regulators indicates there is something "wrong" with the banking sector (thus negative).

Conclusion

In this paper, we proposed a hypothesis that the usages of domain entities (*financial regulators* and *banks*) and contested terms (terms relating to concepts that had bear much debate) could serve as proxies of ontological shifts in the general sentiment of the news in financial sectors.

We use a bag-of-words method for analyzing texts for computing the affect content. A univariate analysis of the distribution of three different types of terms in a large corpus of news about banks shows that the general level of negativity in the news about banks has increased. A multivariate analysis, based on correlation and factor decomposition, shows references to *regulatory bodies* strongly associated with *negative affect*, forming a heavily loaded factor in the analysis. We believe this might be strong evidence supporting our argument that those terms other than pure sentiment bearing words, for example, news flow and contested terms could possibly serve as proxies to sentiments in domain context. This, perhaps, is due to the fact that frequent discussions about a domain concept such as regulators or fierce debate over a contested term might imply the absence of such concept, which, in our case, is the regulation of the financial institutions. We have identified several other factors which could provide further insight to the relationships between contested terms and sentiments: a "positive" factor which also loads with pro-governance and pro-regulation terms; an anti-compliance and anti-regulation factor that has opposite loadings on positivity; an anti-governance factor, and a bank factor. Interpretation of the factors were attempted.

Our future work would focus on the refinement of contested term lexicon as well as exploring techniques from time series analysis to model the changes of news flow, contested terms and sentiments, which would help capturing the dynamics of the system better. We also plan to leverage lexical information more in the future to enhance the accuracy of analysis.

Acknowledgments

Thanks to Yorick Wilks for comments and advice on this research. This work was supported by Enterprise Ireland grant #CC-2011-2601-B for the GRCTC project and a Trinity College research studentship to the first author.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Chair, N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, volume 25, page 2010. European Language Resources Association (ELRA).

Cain, G. (2012). How Can an Information Campaign Win Support for Peacekeeping? *Journal of International Peacekeeping*, 16(1-2):1–2.

Devitt, A. and Ahmad, K. (2008). Sentiment Analysis and the Use of Extrinsic Datasets in Evaluation. In *Proc. of the 6th Intl. Conf on Language Resources and Evaluation*.

Forbes (2011). World's Most Important Banks.

Hafez, P. and Xie, J. (2012). Factoring Sentiment Risk into Quant Models. *Available at SSRN*.

Kim, K. and Barnett, G. A. (1996). The Determinants of International News Flow A Network Analysis. *Communication Research*, 23(3):323–352.

Lasswell, H. D. (1948). The Structure and Function of Communication in Society. *The communication of ideas*, 37.

Namenwirth, J. Z. and Lasswell, H. D. (1970). *The Changing Language of American Values: a Computer Study of Selected Party Platforms*. Sage Publications.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. (1967). *The Measurement of Meaning*, volume 47. University of Illinois Press.

Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. M.I.T. Press; First Edition edition (January 1, 1966).

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.