Tracking Information Flow in Financial Text

Will Radford †‡ **Ben Hachey** ‡⋄ **James R. Curran** †‡ **Maria Milosavljevic** ‡

School of Information Technologies †
University of Sydney
NSW 2006, Australia

Capital Markets CRC ‡
55 Harrington Street
NSW 2000, Australia

Centre for Language Technology ⋄
Macquarie University
NSW 2109, Australia

{wradford,james}@it.usyd.edu.au {bhachey,maria}@cmcrc.com

Abstract

Information is fundamental to Finance, and understanding how it flows from official sources to news agencies is a central problem. Readers need to digest information rapidly from high volume news feeds, which often contain duplicate and irrelevant stories, to gain a competitive advantage. We propose a text categorisation task over pairs of official announcements and news stories to identify whether the story repeats announcement information and/or adds value. Using features based on the intersection of the texts and relative timing, our system identifies information flow at 89.5% F-score and three types of journalistic contribution at 73.4% to 85.7% F-score. Evaluation against majority annotator decision performs 13% better than a bag-of-words baseline.

1 Introduction

Financial news is an important resource for capital market participants and plays a central role in how they interact with the market. Companies must continuously disclose any information "a reasonable person would expect to have a material effect on the price or value of the entity's securities" (ASX, 2008). News agencies publish Finance stories that report on a broad range of events. Some stories report facts directly from announcements and may add value by presenting background knowledge, expert analysis or editorial commentary. The financial environment rewards participants that are alert and responsive to incoming information (Zaheer and Zaheer, 1997) and automated analysis of information flow is highly advantageous.

We define information flow between a pair of documents as when one document repeats information from the other. Textual similarity is central to this and has been addressed in a variety of research areas. Plagiarism detection concentrates on verbatim duplication of sections of text (Brin et al., 1995; Wise, 1996), while Information Retrieval techniques assess similarity at the broader topic level (Manning et al., 2008). Text Reuse examines a finer-grained notion of similarity (Metzler et al., 2005) between the verbatim copying and topic similarity. Topic detection and tracking (Allan et al., 1998a) focuses on tracking emerging events at a topic level over a news feed.

We examine two sources: the Australian Securities Exchange (ASX)[1] official announcements and the Reuters NewsScope Archive (RNA)[2], both of which release time-stamped documents tagged with one or more company stock *ticker* codes. In this research, ASX-RNA document pairs that share the same ticker and were published within a time window are extracted. The ASX-RNA pairs are annotated to indicate information flow (LINK) and, if so, whether the story is the first to report an announcement (FIRST), background (BACK) or analysis (ANLY) content. We formulate four tasks classifying whether each label applies to ASX-RNA pairs.

We design textual and temporal features to model information flow between the ASX-RNA pairs. The intersection of unigrams and bigrams from their texts and titles provides a baseline approach. Set-theoretic bags-of-words, similarity scores, sentence and number matches, and common sequence counts are used to capture textual similarity. Temporal features such as the pair publication times and lag represent the market news cycle. In the LINK classification experiments using Maximum Entropy models, we achieve 89.5% F-score and between 73.4% and 85.7% F-score in the BACK, FIRST and ANLY experiments. Evaluated against annotator majority decision, the system scores 13 points above a bag-of-words baseline F-score. With this new task and dataset, we demonstrate that it feasible to track information flow in financial markets.

[1] http://asx.com.au

[2] http://thomsonreuters.com/products_services/financial/financial_products/event_driven_trading/newsscope_archive

2 Background

Global news agencies operate on a 24 hour news-cycle in a highly competitive environment and are under pressure to report events as quickly as possible. Apart from timely reporting, they must isolate the salient facts from source material, simplifying them if necessary. Commonly available background information about people or events is provided to place the story in context. As well as reporting existing information, news sources generate novel information in the form of analysis, editorial content and commentary.

Identifying and measuring the value and timeliness of their contribution is a principal goal of our study. Textual similarity is the core of our approach to the information flow problem and has been explored by many research areas. In the information flow context, we propose that textual similarity will model ASX announcement facts and figures reported in RNA stories.

Information Retrieval provides many fundamental techniques for textual similarity. Perhaps the most simplistic of these is *bag-of-words*, which represents a document as an unordered collection of its words that acts as "a quantitative digest" (Manning et al., 2008). Stopword filtering and weighting functions such as TFIDF (Spärck Jones, 1973) attempt to emphasise information-bearing, or unusual, words. Having represented the text in these ways, vector space models treat them as vector parameters to a cosine function to quantify their similarity (Salton et al., 1975). Despite their simple model of language, these methods are robust and effective.

Plagiarism Detection uses concepts of textual similarity to identify wholesale copying of text or source code that indicates academic misconduct (Brin et al., 1995; Wise, 1996). Although the pathological case of verbatim copying is reasonably easy to detect, exact matching methods can be circumvented by simply reordering sections or changing a few words. Fingerprint techniques have separated documents into meaningful chunks, typically sentences, and sequences of these are hashed for later comparison against new documents. This reduces the complexity of the matching operation and allows the system to scale to large numbers of documents.

Fingerprinting techniques are also used in Coderivative Document Detection which identifies documents that share a common antecedent. Rather than direct copies, co-derivative documents are those where "long blocks of text are largely preserved, but possibly with intermittent modifications, and some original text is added" (Hoad and Zobel, 2003).

Text Reuse explores the reformulation and restatement of short phrases, part of a similarity spectrum between the specific matches of plagiarism detection and IR's topic similarity (Metzler et al., 2005). Motivated by the concentration of research at either end of this spectrum, the authors aim to track text and facts through corpora at sentence granularity using a variety of similarity measures. Clough et al. (2002) use similarity scores as features to classify newspaper stories as *wholly*, *partially* or *not* derived from UK Press Association newswires. They achieve their best *wholly/partially* F-score of 88.2% at the expense of 64.9% *not* F-score using Naïve Bayes classifiers.

Topic Detection and Tracking (TDT) was part of the TREC programme and focussed on events: "something that happens at a particular time and place" (Allan et al., 1998b). Subtasks, including Event Tracking and Link Detection, encouraged a wide range of approaches including relevance models (Lavrenko et al., 2002) and linguistic features such as noun phrase heads, synonyms and verb semantic classes (Hatzivassiloglou et al., 1999).

Novelty Detection was a later TREC task and models "an application where a user is skimming a set of documents, and the system highlights new, on-topic information" (Soboroff, 2004). Rather than TDT's document oriented approach, the input data is a sequence of sentences related to a topic and is a finer-grained task. Interestingly, the notion of novelty is often encoded as text *dissimilarity* with the already topic-related set of preceding sentences.

The Sentence Alignment task uses a loose notion of textual similarity to align sentences and their translations in parallel corpora and is typically a preprocessing step for Machine Translation training. Differing languages rule out word matching and so approaches tend to address structural features. Brown et al. (1991) report good results using sentence word length to align English and French sentences from Canadian Parliamentary Hansards, as do Gale and Church (1991), who use sentence character length.

News stories and announcements are inextricably linked to their release time and modelling temporal features is important. "Information streams" can be modelled as a 'bursty' time-series using a Hidden Markov Model over hidden states that specify an emission rate (Kleinberg, 2003). Highly-ranked bursts tend to reveal emerging technical terms and language change and the "landmark" documents these appear in is analogous to TDT. Gruhl et al. (2004) consider information flow as an epidemic, using hyperlinks and weighting TFIDF as word use changes.

Our approach appropriates some of these textual similarity techniques and, with temporal features, effectively models information flow.

3 Data

The information flow task requires that we collect documents from both primary and secondary sources cov-

Year	ASX	RNA Stories
2003	66,233	1,901,722
2004	80,570	1,954,259
2005	90,484	2,053,525
2006	102,235	2,298,462

Table 1: Document count per year.

Source	Count	Text (%)
ASX	10,404	83.9
RNA	8,277	99.6

Table 2: Document type distribution and text coverage.

ering the same time span and tickers. Sirca[3] provides ASX official announcements and RNA stories to subscribers. Table 1 shows the document count per year for our entire dataset. The overarching trend is that the volume of ASX and RNA data increases with time, though it is worth noting that the count for RNA data includes all Reuters stories released globally, which explains the disparity in size.

Table 2 describes our experimental dataset: a subset of the ASX and RNA datasets, all chosen from an 18 month period from the beginning of 2005. We show the counts of documents in each source and the proportion of those documents which yielded usable text. To filter the RNA stories specific to the ASX market, we select only those marked with ASX tickers and the English language tag. We choose 403 tickers from the ASX200 index[4] over the last day of each year from 2002 to 2008 to identify large and newsworthy companies.

The broad scope of the ASX's continuous disclosure rules means that almost any type of document can appear as an announcement. While these are all in PDF format, the dataset includes short letters, corporations law forms, long annual reports and presentation slides. In addition to these differences in length and form, companies' different industries mean that a wide variety of genres and topics can appear. For any content-level processing, the announcement text must be extracted from the PDF file, which may include scanned or faxed documents. Text for 83.9% of documents was extracted using the PDFBox[5] Java libraries. Plain-text metadata is also included, specifying the publishing timestamp, title and related tickers of each announcement.

The RNA data collects together stories taken from the global Reuters news feed and represents a unique multi-lingual resource. Each story is made up of a sequence of distinct *events* based on the Reuters work-flow. Figure 1, taken from the RNA documentation, shows an example of the evolution of a story. A newsworthy event might consist of alerts concisely stating the main information, followed by a newsbreak with a headline, two to four paragraphs and then any number of updates to Reuters' coverage. We use a unique 'story key' found in each event to aggregate them into a story, reconstituting the text to its final, canonical form. In addition to the text and title, each story is tagged with lists of relevant tickers, languages, topics and geographical areas. Only 55.7% of RNA stories are tagged with one ticker, in contrast to the 92.7% of ASX announcements. RNA stories, as such, are more likely to report about more than one company and possibly contain different threads of information.

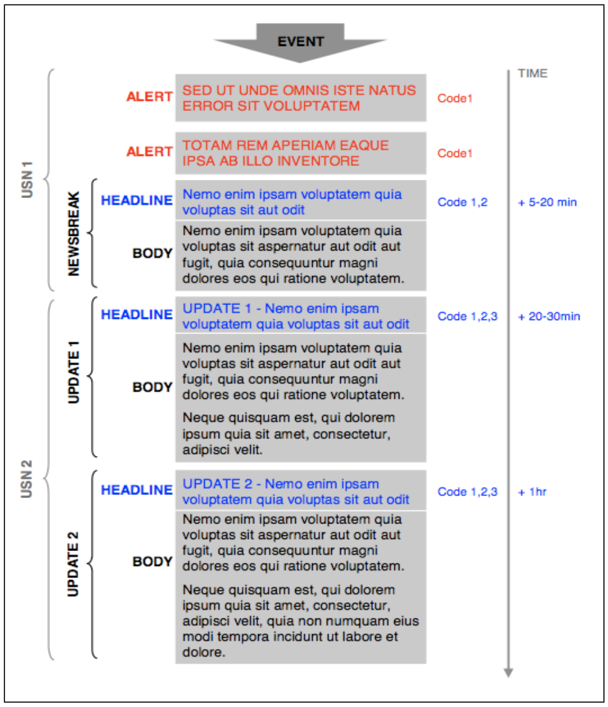

EVENT
TIME
USN 1
USN 2
NEWSBREAK
UPDATE 1
UPDATE 2
ALERT
SED UT UNDE OMNIS ISTE NATUS ERROR SIT VOLUPTATEM
Code1
ALERT
TOTAM REM APERIAM EAQUE IPSA AB ILLO INVENTORE
Code1
HEADLINE
Nemo enim ipsam voluptatem quia voluptas sit aut odit
Code 1,2
+ 5-20 min
BODY
Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, quia consequuntur magni dolores eos qui ratione voluptatem.
HEADLINE
UPDATE 1 - Nemo enim ipsam voluptatem quia voluptas sit aut odit
Code 1,2,3
+ 20-30min
BODY
Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, quia consequuntur magni dolores eos qui ratione voluptatem.
Neque quisquam est, qui dolorem ipsum quia sit amet, consectetur, adipisci velit.
HEADLINE
UPDATE 2 - Nemo enim ipsam voluptatem quia voluptas sit aut odit
Code 1,2,3
+ 1hr
BODY
Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, quia consequuntur magni dolores eos qui ratione voluptatem.
Neque quisquam est, qui dolorem ipsum quia sit amet, consectetur, adipisci velit, quia non numquam eius modi tempora incidunt ut labore et dolore.

Figure 1: The evolution of an RNA story.

Link	Text
FIRST	. . . Record BHP profit of $2.45 million. . .
BACK	. . . BHP has been moving into NSW. . .
ANLY	. . . The profit exceeds expectation, said. . .

Table 3: Examples of RNA story journalistic contribution given the ASX announcement information: *BHP posted record annual profits of $2.45 million.* .

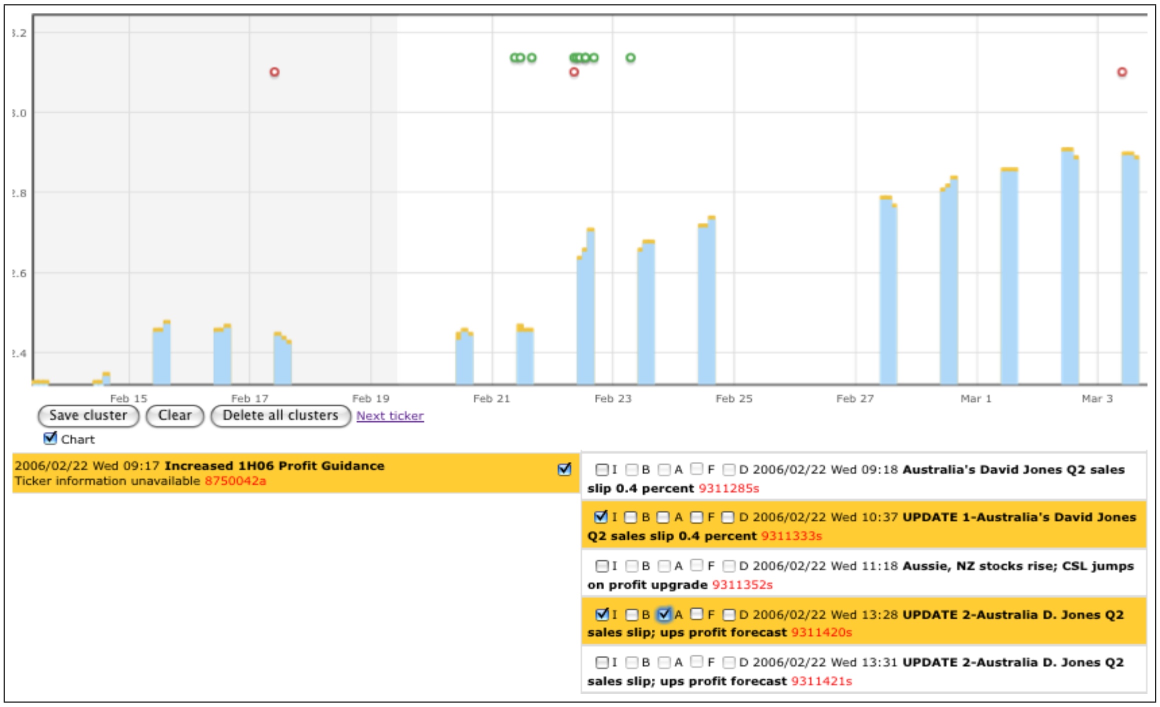
4 Annotation Scheme

We developed a scheme that codifies information flow in Finance text. The scheme describes two phenomena: whether an RNA story contains information from an ASX announcement (LINK) and, if so, the journal-

[3] http://www.sirca.org.au

[4] http://www.standardandpoors.com

[5] http://incubator.apache.org/pdfbox



2006/02/22 Wed 09:17 **Increased 1H06 Profit Guidance** ☑
Ticker information unavailable 8750042a

□ I □ B □ A □ F □ D 2006/02/22 Wed 09:18 **Australia's David Jones Q2 sales slip 0.4 percent** 9311285s

☑ I □ B □ A □ F □ D 2006/02/22 Wed 10:37 **UPDATE 1-Australia's David Jones Q2 sales slip 0.4 percent** 9311333s

□ I □ B □ A □ F □ D 2006/02/22 Wed 11:18 **Aussie, NZ stocks rise; CSL jumps on profit upgrade** 9311352s

☑ I □ B ☑ A □ F □ D 2006/02/22 Wed 13:28 **UPDATE 2-Australia D. Jones Q2 sales slip; ups profit forecast** 9311420s

□ I □ B □ A □ F □ D 2006/02/22 Wed 13:31 **UPDATE 2-Australia D. Jones Q2 sales slip; ups profit forecast** 9311421s

Figure 2: A *screen* from the annotation interface showing links between an announcement and two stories

istic contribution that the RNA story makes. The LINK label applies if the story in an ASX-RNA pair contains information from the announcement. In this case, information is defined as company details that are legally required to be disclosed *first* through an announcement to the ASX (i.e., continuous disclosure). Journalistic contribution mainly concerns the RNA story and can be indicated by any of FIRST, BACK or ANLY. FIRST indicates that the story is the first to cover the information in the announcement. BACK refers to background information regarded as common knowledge and publicly available before the release of the announcement. ANLY describes new information added by the news source, such as analysis, editorial content and new quotes from industry commentators. Table 3 shows examples of the link types. The distinction between BACK and ANLY is subjective since it the annotator must decide whether information is already known (i.e., BACK) or whether it is novel analysis.

The scheme also defines an annotation unrelated to information flow: DIGEST. The Reuters dataset includes stories that contain snippets of news that relate to multiple events and companies, often a daily market report of reviews of ‘Hot stocks’. These RNA stories would be likely to report information from many sources and the DIGEST annotation allows their exclusion (although we do not do so in our experiments).

5 Annotation Task

The annotation task and interface were designed to allow annotators to read ASX-RNA pairs and identify any information flow. A pilot annotation team of Finance PhD students assisted in the development of the scheme and interface by participating in a shared annotation task. The initial scheme specified that the links were mutually exclusive but resulted in low Cohen’s Kappa agreement scores (Cohen, 1960). After several iterations of relaxing and refining the scheme, Kappa inter-annotator agreement was sufficient to begin the main annotation phase.

A second team of Finance students was hired and after completing a shared task, the seven annotators with consistently high average Kappa score were chosen to continue. An 18 month period from 2005 to mid-2006 was used to create screens. A *screen* consists of the ASX announcements and RNA stories released for a ticker over a fortnight and allows annotators to view the pairs in context and apply information flow links. We randomly sampled three subtasks for each annotator consisting of 215 screens to be completed individually, 50 shared screens and a final 215 individual screens. The shared task midway through the project allowed re-checking of agreement figures and is also used as held-out evaluation data. Due to low agreement, only five annotators’ data was used for training and evaluation and not all targeted screens were annotated. Consequently, 1779 individual screens were an-

Task	Mean Kappa
LINK	0.75
FIRST	0.71
BACK	† 0.66
ANLY	† 0.55

Table 4: Mean Kappa inter-annotator agreement scores (N=5). † indicates lower than acceptable Kappa.

Label	Count	%	ASX	RNA
LINK	6,380	4.71	17.1	33.6
FIRST	2,638	1.95	16.7	17.7
BACK	4,707	3.47	15.1	27.2
ANLY	1,759	1.30	7.2	9.9

Table 5: Count and percentages of links in both training and evaluation datasets (135,537 pairs, $lag_{max} = 7\ days$). Percentage of linked documents for each source.

notated for use as training data and 42 shared screens for evaluation.

Figure 2 shows a *screen* from the annotation tool. A screen consists of three main sections: a time-aligned navigation panel spanning the top, then two vertical document lists below. The top timeline panel shows the fortnight of interest and a context week either side (though the right context week is not shown in this figure), showing stories that might be related to the announcements at the edges of the target fortnight.

The panel displays the stock price and rows of dots indicating an ASX announcement or RNA story. In this example, the ASX announcements on the bottom row are followed by a burst of RNA stories on the top row and this visualisation helps annotators to quickly navigate large complex screens with many announcements and stories.

The two time-ordered lists of documents on the bottom of the panel show the ASX announcements on the left and RNA stories on the right. The titles and timestamp are always visible and annotators can click to reveal the RNA story text or open the ASX PDF file. Checkboxes corresponding to the information flow link types are arranged towards the middle of the lists.

Although all evaluation here applies to ASX-RNA pairs, the list presentation format allows annotators to cluster related documents. We define a cluster as an announcement and one or more stories related to that same announcement. The highlighted cluster in Figure 2 consists of the ASX announcement on the left and second and fourth RNA stories on the right hand side. The RNA I checkboxes show cluster membership and the A checkbox shows the ANLY link for the second story. The main benefit of this strategy is efficiency since the top-down view allows annotators to easily isolate clusters without re-reading documents and see which clusters they had previously created.

It is also possible to add multiple ASX announcements to the same cluster, for example when a meeting announcement is followed by a set of presentation slides. However, the main constraint is that clusters be as minimal as possible and any announcements containing new information should form new clusters. For example, a company takeover might span several months of offer and counter-offer but our annotation should pick out the individual stages of the overarching process. Moreover, the minimality constraint encourages conceptual clarity and mirrors the way information is released piecemeal while still allowing later aggregation of clusters if required.

Inter-annotator agreement for our five annotators is assessed using Cohen's Kappa over the shared task of 42 screens. Table 4 shows that acceptable Kappa scores are achieved for LINK and FIRST with borderline Kappa scores for BACK– relative to the threshold at 0.67 (Carletta, 1996). ANLY was annotated with lower agreement and is consistent with annotator feedback during scheme development, where ANLY links were the most difficult to disambiguate from BACK since the distinction between existing and new information proved subjective.

We placed an upper bound of a week on the time lag between ASX and RNA publishing time. Though primarily an optimisation step to reduce the number of pairs for comparison, it is also consistent with the cluster minimality constraint; annotators were encouraged to split clusters that spanned too long a time.

Table 5 shows the count of links (with a maximum lag of a week) across the individual and shared screens and, in the second column, the distribution of the link types. BACK is the most frequent type beyond the prerequisite LINK link, followed by FIRST and then ANLY. While the low ANLY proportion might be due to stories that emphasise topic and background, low Kappa scores for BACK and ANLY makes it difficult to rule out annotator confusion. Table 5 also indicates that while roughly even proportions of documents are linked by FIRST and ANLY, far more RNA stories are linked by LINK and BACK. Indeed, no more than a third of each journalistic contribution link type appear without another (they all co-occur with LINK) , indicating that the annotators applied them with a high degree of overlap.

6 Features

We model the information flow problems using a variety of text similarity and temporal features extracted from the ASX-RNA pairs. Each feature value is binary and real-valued features are placed into equally sized bins (with the exception of lag as mentioned below). The text and title of the announcement and story were both tokenised using the NLTK's word tokeniser (Bird et al., 2009) and implementation of the Punkt sentence tokeniser (Kiss and Strunk, 2006). Unigram and bigram features are extracted, ignoring punctuation and any n-grams that include a token in NLTK's list of 127 English stopwords.

To model fine-grained textual similarity, we define three set-theoretic classes of bag-of-words features depending on where content is found: intersection (ASX $\cap$ RNA), *only* in the announcement (ASX $\setminus$ RNA) and *only* in the story (RNA $\setminus$ ASX). These methods are applied to unigrams and bigrams in the title and body text of the ASX-RNA pair. The intersection text/title features are used for a baseline approach. The set-theoretic features are mainly designed to model information flow's similarity, and the story's contribution (RNA $\setminus$ ASX).

The information flow problem requires tracking of short units of text such as distinctive terms and figures. We encode this using Text Reuse similarity scores over text unigrams, title unigrams and tokens containing one or more digits (Metzler et al., 2005). The scores calculated are symmetric overlap, asymmetric overlap favouring the RNA story with and without inverse document frequency weighting, a TFIDF overlap score and two cosine similarity scores, one unweighted and one TFIDF weighted.

To capture longer units of reused text, we take tokens *including* stopwords from each sentence and count the number of exact sentence matches between the ASX-RNA pair. Common token sub-sequences are found using Python2.6's difflib library[6]. The sequences include stopwords and have a minimum length of three since we already calculate bigrams. Both the lengths and counts of these sub-sequences are rounded into bins to produce features such as: *seq-len* and *seq-len-count* indicating that there were matches of *len* and that there were *count* of them respectively.

We also extract features to represent the precision of financial figures mentioned in both texts: the more precision used, the more important the figure. For each number string appearing in both texts, if it consists of a non-zero digit followed by any number of zeros or periods, the characters are replaced with `0`. Otherwise, the characters were replaced with `#`. For example, a round number like `5000` would be replaced by `0000` and a more interesting number like `45.3` would be replaced with `####`. The set of these precision-hashed numbers are used as features.

Time is an important factor in news and we propose that the placement of announcements and stories in the news cycle is significant. The temporal features consist of the *time lag* between the release time of the story and the announcement, mapped to bins that increase in size either side of zero (to account for stories that occur before announcements). For example, the bins around zero are: `[-15...-5)`, `[-5...0)`, `[0...5)`, `[5...15)` and are left-closed and right-open so that a pair released at the same time will have a feature value of `[0...5)`. In addition to this, the time of each document release is rounded to the half-hour, generating a feature to represent the ASX's news cycle.

Label	Training	%	Evaluation	%
Total pairs	30,249	100.0	1,621	100.0
LINK	5,596	18.5	231	14.3
FIRST	2,394	7.9	81	5.0
BACK	4,118	13.6	166	10.2
ANLY	1,472	4.9	72	4.4

Table 6: Distribution of links in the training (30,249 pairs) and evaluation (1,621 pairs) datasets - both use a lag of less than 1 day.

7 Experimental Methodology

The information flow problem is framed as four binary text categorisation tasks over the ASX-RNA pairs – one task for each link type. The development experiments use 10-fold cross validation and we report precision, recall and F-score for classifying pairs as labelled. We do not report scores for classifying *unlabelled* pairs since these are far more common than the labelled pairs. The experiments use the MegaM Maximum Entropy classifier (Daumé III, 2004) with the `binomial` options to represent the binary features.

To compare to human performance, a model is trained using the development data and used to classify the pairs from the shared annotation task. Gold standard *majority* data is created by positing a *true* link where it is marked by a majority of the five annotators – a more difficult task. Each annotator is compared against the majority and the mean result is used as the upper bound on system performance.

Table 5 showed a highly skewed annotation label distribution in the ASX-RNA pairs released within a week of one another. However, approximately 92% of the links occur within 24 hours of one another and Table 6 shows that applying the *short time lag* improves the class distribution. We still consider all pairs in the evaluation dataset, though our system only classifies pairs within the 24 hour lag and thus classifies the 30 LINK labels that lie outside as not linked. Given the lower prior probabilities of true links in the evaluation dataset, we expect the performance to be worse than in development experiments.

[6] `http://docs.python.org/library/difflib.html`

Task	Features	P (%)	R (%)	F (%)
LINK	Baseline	85.0	73.1	78.6
LINK	Best	90.9	88.1	89.5
FIRST	Baseline	66.0	43.9	52.7
FIRST	Best	77.0	70.1	73.4
BACK	Baseline	83.4	67.1	74.4
BACK	Best	88.4	83.2	85.7
ANLY	Baseline	78.9	56.0	65.5
ANLY	Best	86.7	75.0	80.4

Table 7: Precision, recall and F-score for cross-fold validation experiments.

8 Results

Table 7 summarises the experimental results, showing baseline and best precision, recall and F-score for *true* link classification. Text intersection unigrams and bigrams, title intersection unigrams and bigrams were used as baseline features and those scores were exceeded for all link types. While higher F-scores were achieved, for the most part, in the tasks with higher prior link probabilities, scores in ANLY were surprisingly high given its low prior of 4.9.

Table 8 shows the best performing (by F-score) feature combinations for each link type. To test the contribution of each feature, subtractive analysis was performed on the best performing feature set for each link type. An experiment is conducted that uses all but one feature and the results compared to best using approximate randomisation (Chinchor, 1995) to assess whether adding the omitted feature results in a statistically significant improvement.[7] Features used are marked with ·, while features are marked with ⋆ or ⋆⋆ if their removal results in significantly worse F-score (at $p<0.05$ and $p<0.01$ respectively).

The first observation to make from the table is that the tasks can be separated into two groups on the set of features that was most successful: LINK/BACK and FIRST/ANLY, though this may also be related to the different prior link probabilities, higher and lower for each group in this case.

Features based on the text play perhaps the broadest role, both modelling information flow and journalistic contribution. Although intersection unigrams and bigrams appeared in all feature sets, text intersection bigrams were only significant in LINK and BACK. One reason might be that they more effectively model topic-level textual similarity while being less susceptible to single words appearing by chance in both texts. The textual similarity measures were significant for the LINK, FIRST and BACK experiments, perhaps because they are able to weight terms more effectively.

Features	LINK	FIRST	BACK	ANLY
ASX∩RNA TEXT-1G	·	·	·	·
ASX∩RNA TEXT-2G	⋆⋆	·	⋆⋆	·
ASX\RNA TEXT-1G				
ASX\RNA TEXT-2G				
RNA\ASX TEXT-1G		·		·
RNA\ASX TEXT-2G		·		⋆⋆
ASX∩RNA TITLE-1G	·	·	·	·
ASX∩RNA TITLE-2G	·	·	·	·
ASX\RNA TITLE-1G	⋆⋆		⋆⋆	
ASX\RNA TITLE-2G	⋆⋆		⋆⋆	
RNA\ASX TITLE-1G	·	·	⋆⋆	·
RNA\ASX TITLE-2G	⋆⋆	·	⋆	·
TEXT SIMILARITY	⋆⋆	⋆⋆	⋆	·
TITLE SIMILARITY	·	⋆⋆	·	·
SENTENCES	·	·	·	·
SEQUENCES	·	⋆⋆	·	·
NUM SIMILARITY		·		·
NUMBER PRECISION		⋆		·
TIME LAG	⋆⋆	⋆⋆	⋆⋆	·
TIME OF DAY	·	·	·	⋆

Table 8: Feature combinations for the best performing development experiments. Features significant from subtractive analysis are annotated ⋆ ($p<0.05$) and ⋆⋆ ($p<0.01$)

Common sequence matching proved significant in detecting FIRST and no other experiments. Reported figures and information are more likely to be reported verbatim, rather than be subject to editing, and this may play a role in the features' success. Of the text set difference features, only bigrams that appeared in the RNA story and not the ASX announcement were significant and only then for ANLY, suggesting that the feature effectively represents commentary. Interestingly, text present only in the ASX announcement was not used in any well-performing experiment. One potential explanation is that the wide variety of text sizes is simply too noisy a feature for the model to generalise.

Titles play an important role in announcements and stories, summarising the event that they report on. Rather than intersection, the ASX and RNA set differences proved to be more significant features. The ASX unigram and bigram varieties of this feature were significant for both LINK and FIRST experiments and the RNA unigrams and bigrams less significant for the same classes. This may indicate that cues of ASX announcement newsworthiness may appear in ASX titles, yet not be repeated in the titles of stories that report on them. Conversely, title terms that indicate that a story reports directly on an announcement may not be found in that announcement's title. In addition to this, titles are often constrained by space and the need for concise communication and are less likely to contain

[7] We adapt a parsing evaluation script `http://www.cis.upenn.edu/~dbikel/software.html`

Task	Baseline	Best	Upper
LINK	62.5	⋆⋆76.7	86.4
FIRST	38.8	53.0	83.1
BACK	56.7	⋆⋆68.0	80.8
ANLY	38.2	45.7	72.5

Table 9: Model F-score agreement with *majority*. Upper is the mean of the F-scores for each annotator and *majority*. ⋆⋆ indicates significance (p<0.01)

non-indicative terms. Title similarity was important in LINK and FIRST, the only significant title-based feature for the latter task. Further exploration would be required to measure how much information is transferred in the titles alone.

Numbers are central to information flow in Finance and the two features based on numbers, similarity measures and precision were present in the FIRST and ANLY experiments. The number precision feature was significant for the FIRST experiments, while similarity and precision were just under significance for ANLY. Though these initial results are encouraging, the importance of number to information flow means that more work is required.

Finally, news has a strong temporal dimension and we expected the lag feature to be significant for all link types. While it was for LINK, FIRST and BACK, the time-of-day feature was more significant for ANLY. That the analysis and commentary are the only link types sensitive to their placement in the news cycle points, potentially, to less time critical stories released at regular times.

Table 9 shows performance between the baseline and *majority* in the evaluation task. While the lack of significant results for FIRST and ANLY is somewhat discouraging, reasonable results for LINK and BACK indicate the feasibility of our approach to the information flow problem.

9 Conclusion

Our paper presents a formalisation of the information flow problem in the Finance domain. We present an annotation scheme that codifies flow of facts from primary to secondary sources and apply it to ASX announcements and RNA stories. Moreover, the scheme models three types of journalistic contribution and, despite its difficulty, can be applied with high agreement.

We explore a range of features from diverse fields and combine them to classify the different information flow types. Textual features based on the intersection and differences of document texts and titles prove useful, while number features show promise at identifying financial figures. Temporal features allow modelling of the news cycle and news source responsiveness to identify linked documents.

This paper presents a new approach to the information flow problem in the Finance domain, essentially text categorisation over the *pair* of documents. While bag-of-words performs predictably well in this task, we are able to take advantage of temporal and textual features to classify information flow at 89.5% F-score and journalistic contribution from 73.4% to 85.7% F-score. In evaluation against human performance of 86% F-score, our system scores 77% for flow classification; demonstrating we can feasibly track information flow in Finance text.

10 Acknowledgements

We would like to thank all reviewers, both anonymous and from the Sydney Language Technology Research Group for their feedback. We also thank Dominick Ng, Silvio Tannert, Amy Kwan and Ingo Wiegand at the Capital Markets CRC for their assistance preparing the corpora and managing the annotation project. This research was conducted with the support of a Capital Markets CRC High Achiever's Scholarship.

References

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998a. Topic detection and tracking pilot study final report. In *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

James Allan, Ron Papka, and Victor Lavrenko. 1998b. Online new event detection and tracking. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45, New York, NY, USA. ACM.

ASX. 2008. Continuous disclosure. *ASX Listing Rules*, Chapter 3.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python.

Sergey Brin, James Davis, and Héctor García-Molina. 1995. Copy detection mechanisms for digital documents. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, pages 398–409, New York, NY, USA. ACM.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Nancy Chinchor. 1995. Statistical significance of muc-6 results. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 39–43, Morristown, NJ, USA. Association for Computational Linguistics.

Paul Clough, Robert Gaizauskas, Scott S. L. Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *ACL '02: Proceedings of the 40th Annual Meeting on*

Association for Computational Linguistics, pages 152–159, Morristown, NJ, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Hal Daumé III. 2004. Notes on cg and lm-bfgs optimization of logistic regression. Aug.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, Morristown, NJ, USA. Association for Computational Linguistics.

Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. 2004. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, New York, NY, USA. ACM.

Vasileios Hatzivassiloglou, Judith L. Klavans, and Eleazar Eskin. 1999. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 203–212, College Park, Maryland, June. SIGDAT.

Timothy C. Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology*, 54(3):203–215.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Jon Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Victor Lavrenko, James Allan, Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas. 2002. Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research*, pages 115–121, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Chris Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. *Cambridge University Press New York, NY, USA*, Jan.

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524, New York, NY, USA. ACM.

Gerard Salton, A Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Ian Soboroff. 2004. Overview of the trec 2004 novelty track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*.

Karen Spärck Jones. 1973. Index term weighting. *Information Storage and Retrieval*, 9(11):619–633.

Michael J Wise. 1996. Yap3: improved detection of similarities in computer program and other texts. *SIGCSE Bull.*, 28(1):130–134.

Akbar Zaheer and Srilata Zaheer. 1997. Catching the wave: alertness, responsiveness, and market influence in global electronic networks. *Manage. Sci.*, 43(11):1493–1509.