Discovering News Events That Move Markets

Yuriy Gurin, Terrence Szymanski & Mark T. Keane

Insight Centre for Data Analytics &
School of Computer Science,
University College Dublin,
Belfield, Dublin 4, Ireland
Email: *firstname.lastname@insight-centre.org*

***Abstract*—Recently, there has been an explosion of interest in the use of textual sources (e.g., market reports, news articles, company reports) to predict changes in stock and commodity markets. Most of this research is on sentiment analysis, but some of it has tried to use the news itself to predict market movements. In this paper, we use 10-years of news articles – from a weekly, agricultural, trade newspaper – to predict price changes in a commodity market for beef. Two experiments explore the different ways in which news reports affect the market via (i) *major market-impacting events* (i.e., rare natural disasters or food scandals) or (ii) *minor market-impacting events* (e.g., mundane reports about inflation, oil prices, etc). We find that different techniques need to be used to uncover major events (e.g., LLRs) as opposed to minor events (e.g., classifiers) and show that no single unified predictive model appears to be able to do both.**

***Keywords*—Event Detection; Market Prediction, Text Analytics, News**

I. Introduction

On March 15th 2011, when the news broke that a radiation leak had been detected at the Fukushima power plant, following a 9.0 magnitude earthquake and a massive tsunami, the NIKKEI Index dropped 11%, heralding the beginning of a major dip in the Japanese economy. All markets – whether they be stock-markets or commodities markets – are impacted by events in the world. Some of these market-impacting events are major – such as, bankruptcies, natural disasters, wars – and lead to sudden, significant shifts in prices. Others are more mundane, minor events – changes in consumer sentiment, stockpiles of goods, inflation – which usually coalesce together to move the market in one direction or another. In this paper, we consider whether data-analytics can be applied to textual news sources to discover these two different classes of market-impacting events (the major and the minor), while reflecting on the fact no single technique seems to capture both types.

On the face of it, from a textual perspective, there appears to be a fundamental conflict between major and minor market-impacting, news events. *Major events* that impact the market, "stick out" from "normal" news coverage; these events tend to be rare, exceptional, and unusual and are described in rare, exceptional and unusual terms. For instance, the term "Fukushima" was probably seldom mentioned in market reports about the NIKKEI Index prior to March 2011, though after the earthquake it appeared in thousands of financial commentaries. In a suitable corpus of finance articles, this term would have a very distinctive signal, appearing out of nowhere and rapidly rising in its occurrence relative to more commonly-occurring terms in market commentaries. *Minor events* that impact the market, in contrast, are those that make up "normal" news coverage; these are the repeated, common and usual events that coalesce to shift the market (usually in small ways) and are described in repeated, common and usual terms. Indeed, most market reports are so repetitive and boring that reporters often jazz them up with metaphoric language (e.g., "Google assaulted the iPhone with Android"). Therefore if we want to find these two classes of events in textual sources, we face a conflict; on the one hand, we need to rely on repeated, often-used text-features to find minor events, but on the other we need to find rare text-features to capture major events. In this paper, we explore the hypothesis that the state-of-the-art, text-analytic techniques for discovering major and minor market-impacting events fundamentally conflict with one another, as they depend on very different regularities and exceptionalities in texts. We illustrate this conflict using 10 years (2005 - 2015) of news coverage and market commentary from a trade newspaper commenting on the beef-livestock market in Ireland (i.e., *The Irish Farmers Journal (IFJ)*).

The Beef Market in Ireland. The market we consider in this work is that for beef livestock in the Republic of Ireland, a €2.37 billion per-annum, export industry; the activities of which are captured in a country-wide database of mart prices for individual animals, with news articles on this market coming from a weekly-published, agricultural newspaper, *The Irish Farmers Journal (IFJ)*. We chose this market and news source because we wanted (i) a complete database of all the transactions in the market (i.e., we have the prices for every animal sold in a full year, which is akin to having every share transaction in a stock market), (ii) an authoritative news source that covered all aspects of this specific market (160,000+ articles over 10 years); so we have a comprehensive set of market reports on this specific trading ecosystem in a geographically-circumscribed location. Indeed, one of the contributions of this paper, lies in how this undifferentiated, large corpus of news articles was filtered down to a set that was specifically relevant to the beef market. Like any market the Irish Beef Market has, at times, suffered from a major market-impacting events such as the "horsemeat scandal" in 2013-14 or, more recently, the Russian ban on EU products following the conflict in the Ukraine.

A Text-Analytics Perspective. In recent years, a variety of text-analytics techniques have been used to predict and track changes in a variety of markets ([1], [2], [3]). Sentiment analyses are amongst the most commonly used methods, where positive/negative words and phrases are used to predict market movements or price changes in particular stocks (see e.g.,

[3]). However, the discovery of market-impacting events arises from a somewhat different literature. A variety of text-analytic techniques have been developed to identify and track events in news feeds (see TREC conferences; [4], [5], [6]) and, so-called, First Story Detection (FSD) has been well researched as a task in these challenges ([7], [8], [9]). In this paper, we draw on this literature exploring the use of log-likelihood ratio techniques to find major market-impacting events (see section IV) and a selection of classifiers to find minor, market-impacting events (see section V).

Overview. In section II, we briefly outline the most relevant literature to the current work on event detection and the techniques used to predict price changes in stock markets. In section III, we describe the data sets we used. In section IV, we describe our methodology for finding major market-impacting events using log-likelihood ratios. In section V, we show how the text-features of articles covering many minor events can be used to predict price changes, before considering the conflict between the two approaches (see section VI).

II. Recent Literature

The techniques examined here to discover market-impacting events rely on methods used for identifying structures and trends in textual time-series data. One widely-used method assumes an infinite-state Markov model governing the rate of appearance of each term in the corpus; in this model, "bursts" of appearances of a term correspond to periods when the Markov model is in a high-energy state, as opposed to periods when it is in a low-energy rest state [10]. Using this method, it is possible to enumerate all bursts over all terms and assign a probability-based weight to each one, in order to rank and identify the highest-scoring bursts [10]. This approach has been applied to track the rise and fall of textual features in various domains, for example tracking scientific research trends in the journal PNAS [11] and tracking "memes" (fixed, short phrases) in the news cycle [12]. However, the results produced by this method vary depending on parameter settings, and the weights have no natural cut-off for determining the significance of a given event.

A different method for finding exceptional events in text, uses chi-squared tests to identify terms that appear significantly more frequently in a given time window than outside that window, at a specified p-value [13]. In this work, related terms can be grouped together so that news stories can be identified as groups of terms (usually using pointwise mutual information) that appear together over a specific period of time. This approach is closely related to the method that we use to discover major market-impacting events; however we adopt the log-likelihood ratio static for significance testing, as it performs better than chi-squared when sample sizes are small [14].

Another relevant line of research has consistently shown that textual data sources can effectively predict upward and downward movements in markets. [1] used parallel time-series data (stock prices) and textual data (news articles) to train separate language models from periods of rising and falling stock prices and then used these models to predict future price trends from current news articles. Fung et al. [2] refined this approach in several ways. Basing their method on the Efficient Market Hypothesis, which states that stories reported in news have a instant impact on the market, they used chi-squared analyses to identify news articles relevant to the market, and then used an SVM classifier to label each news story as having either a positive, negative, or neutral impact on the market. More recently, [4] applied a more complex model, integrating multiple levels of feature extraction on semantic and sentiment levels, to predict foreign exchange markets; thus, demonstrating the effectiveness of this approach to financial domains other than the stock market. Our work on discovering minor market-impacting events is mainly influenced by [2] in that we use an SVM classifier to relate groups of news-articles to market movements.

III. Corpora, SubCorpora & Pre-processing

The analyses presented in this paper are based on a data set consisting of 10 years of articles on farming (2005-2015) from the archives of the weekly-published *Irish Farmers Journal (IFJ)*[1]. This data-set consisted of (i) over 140,000 articles from the IFJ Content Management System (CMS), their archive covering the years 2005-2015 and (ii) almost 20,000 articles gathered from their website using a web-crawler, mainly covering the years 2013 to 2015 (see Table I). These two corpora are very different. The CMS-corpus has every article published between 2006 and 2011[2]. In contrast, the website-corpus has meta-tags on the articles, some of which mention beef and beef markets. Hence, the website-corpus can be filtered using these tags, whereas the CMS-corpus needs to be filtered in some other way to extract articles relevant to the beef market. Both of these corpora present their own challenges; challenges that are typical ones for any project aiming to carry out text-analyses of any archive of news stories. As such, we systematically explored both corpora and pre-processed them in different ways to best prepare them for our tests of the different techniques used to find major and minor market-impacting events.

TABLE I. Number of articles per year in CSM & Website Corpora

Year	CMS	Website
2005	7,246	0
2006	16,884	0
2007	18,859	0
2008	22,538	0
2009	23,235	0
2010	21,127	1
2011	20,680	2
2012	7,136	188
2013	2,590	2,666
2014	711	9,807
2015	104	7,118
Total	141,110	19,782

Hence, from this data-set, we generated three distinct subcorpora to be used in our empirical tests; each represents a different way to filter out an appropriate subcorpus of beef-related articles:

- **Web-Subcorpus**: This subcorpus consisted of one year of beef-related articles from the website corpus (roughly 2000 articles from Nov. 2013 to Nov. 2014);

[1] http://www.farmersjournal.ie/

[2] In 2005, the coverage in the CMS is limited as it was started mid-year (so we exclude it in a lot of our analyses) and after 2011 the CMS numbers drop off as the company explored other archiving options.

it was a highly-filtered set of articles from the website corpus that had been explicitly tagged as being about beef (e.g., from the "Beef News" or "Beef Markets" sections of the website).

- **CMS-Subcorpus**: This subcorpus covered six years of all farming-related articles from the CMS, with racing-related articles excluded (roughly 31,000 articles from 2008-2014); this subcorpus is the CMS corpus roughly filtered by removing horse-racing articles (which account for almost half of the articles in these years).
- **CMS-Beef-Subcorpus** Six years of beef-related articles from the CMS, selected using topic-modelling (roughly 2,000 articles from 2008-2014); Greene [15] has shown that documents on a given topic can be selected from corpora using topic-modeling, so we applied non-negative matrix factorization to the CMS-Subcorpus to find articles that cluster around the beef topic.

These three subcorpora were pre-processed in a standard way, using the Stanford CoreNLP toolkit [16], with stop-word removal and lemmatization being carried out on all articles. Any term that appeared in less than 3 articles and more than 80% of the articles were also removed, reducing the feature space significantly. As the IFJ is published on a weekly basis, our basic time-step was the week. So, all the articles for a given week in the corpus were pre-processed in this way and time-stamped by week.

IV. Discovering Major Market-Impacting Events

In a similar way to the Fukushima example, the intuition we are pursuing to identify major market-impacting events is that, for example, in the Irish Beef Market, the term "horsemeat" was not a common term used in market reports before January 2013 but after the news story broke across Europe about supermarket beefburgers being contaminated with horsemeat, the term would have peaked in market commentaries. So, the term "horsemeat" stood out like it had never stood out before, as the beef market reeled from the negative publicity following the scandal[3].

As argued in the introduction, major market-impacting events tend to be unusual or exceptional and as such tend to involve terms that were not common in market commentaries up until the occurrence of the event (e.g., like "horsemeat", "Fukushima"). So, having created the three subcorpora of beef-related articles our initial aim was to explore whether the log-likelihood ratio (LLR) statistic could be used as a technique to identify such "stand-out" terms. LLR has been used to determine whether the frequency of a term some Input-Text is significantly different relative its frequency in a Background-Text (or collection of texts) and is typically used for first-story detection in the literature (see [13]).

A. The Log-likelihood Ratio Statistic

To identify significant time-specific features (a.k.a. stand-out words in a specific time-period) we use the model comparison metric, the log-likelihood ratio test (also known as the G^2 test). In text-analytics, typically, this statistic is used to compare word frequencies in two text collections; an Input-Text *I* and a Background-Text(s) *B*. In this case, we need to assess whether a word stands out in Input-Texts for different time-periods; for example, we need to assess whether "horse-meat" stands out over a 1-week or a 2-week or 3-week period and so on. Our algorithm computes the LLR statistic for a given term over different sized time-windows (i.e., all windows from 1-10 weeks), each of which we slide across the full time period being analysed (e.g., 2008-2014; see Algorithm 1). For example, in Run#1 we compute the statistic for all words in a succession of Input-Text windows from 1-10 weeks in size, moving this window over the full year of the Web-Subcorpus (Nov. 2013 - Nov. 2014); with each Input-Text being compared against the Background-Text of the full Web-Subcorpus (less the Input-Text; see Table II).

Using these inputs G^2 assesses two hypotheses H_0 and H_1. H_0 asserts that the probability of the word frequency is distributed uniformly over the full time-span. H_1 asserts that the distribution is different; that the frequency of the term in the Input-Text window is significantly different to that of the term in the Background-Text.

$$H_0 : P(w|I) = P(w|B) \tag{1}$$

$$H_1 : P(w|I) \neq P(w|B) \tag{2}$$

$$\lambda = -2 \ln \frac{L(D; H_0)}{L(D; H_1)} \tag{3}$$

Formally, the log-likelihood test Eq. 3 calculates a metric λ based on the ratio of the likelihood *L* of observing the given data *D*, under the two alternative hypotheses H_0 and H_1. We calculate the likelihood of the data (the observed frequency at which word w appears over time) using the binomial distribution. So, words that appear constantly throughout the corpus will receive low scores using this statistic whereas ones that are rare or bursty within a given time period will receive high scores.

B. Algorithm: Using LLR in Time Windows

The method we use to find standout word-features that refer to major market-impacting events has two steps: (1) one that computes LLR scores for word-features in a corpus for different time-windows (see Algorithm 1), (2) one that groups word-features using Point-wise Mutual Information (PMI), to identify relevant events (see Algorithm 2):

1) Finding Standout Word-Features: To find standout word-features, we apply the LLR test to each word in a selected subcorpus corpus, for every time-window (between 1 and 10 weeks), sliding each of these windows across the full time-period covered by that subcorpus, and then ranking all results using the λ score Eq. 3. The highest scores correspond to the standout words in a given time-window, thus finding 'feature-words' that potentially reference major market-impacting news events (see Algorithm 1).

[3]See http://www.europeanmovement.ie/just-the-facts-horse-meat/ for details.

Algorithm 1 Finding standout word-features using LLR

1: $n \leftarrow$ Total rows of M
2: $m \leftarrow$ Total columns of M
3: $\alpha \leftarrow 0.05$
4: $e \leftarrow (\sum_{w=0}^{10}(n-w)) * m$
5: **k** $\leftarrow \left[\sum_{k=0}^{i} M_{k,1}, .., \sum_{k=0}^{i} M_{k,m}\right]$
6: $N \leftarrow \sum M_{i,j}$
7: $S \leftarrow 1-(1-\alpha)^{\frac{1}{e}}$
8: **windows** $\leftarrow Cut\ M_{n,m}$ at n by sliding window 1 .. 10
9: **for** W in **windows** **do**
10: **ki** $\leftarrow \left[\sum_{k=0}^{i} W_{k,1}, .., \sum_{k=0}^{i} W_{k,j}\right]$
11: **kb** $\leftarrow$ **k** - ki
12: $Ni \leftarrow \sum W_{i,j}$
13: $Nb \leftarrow N$ - Ni
14: $H_0 \leftarrow \left(\binom{Ni}{ki} p_0^{ki}(1-p_0)^{Ni-ki}\right) * \left(\binom{Nb}{kb} p_0^{kb}(1-p_0)^{nb-kb}\right)$
15: $H_1 \leftarrow \left(\binom{Ni}{ki} p^{ki}(1-p)^{ni-ki}\right) * \left(\binom{Nb}{kb} p^{kb}(1-p)^{nb-kb}\right)$
16: $\lambda \leftarrow -2\ln\frac{(H_0)}{(H_1)}$
17: $\lambda^{'} \leftarrow$ Apply S to λ
18: **T** $\leftarrow$ Append($\lambda^{'}$)
19: **end for**
20: **T** $\leftarrow$ Sort λ_terms by Score

2) Grouping Word-features Using PMI: Algorithm 1 is just the first step (finding standout words) towards identifying impactful news events; we also need to find feature-words that co-occur with these ones to have a longer word-feature list that allows the identification of the news event in question (see Algorithm 2). This step is completed by applying Pointwise Mutual Information (PMI) to the features in the document-set of the subcorpus; represented in Equation 4:

$$pmi(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

Algorithm 2 identifies word-features that are in the same temporal state (line 4) and computes co-occurrence scores for these word-features in the document-set for this state (line 7). PMI is calculated for terms that appear in more than 3 documents (see line 8). If PMI score is above a threshold, meaning if the co-occurrence of the feature is relatively high in relation to occurrence of all terms between documents then the features are grouped.

C. Procedure

These two algorithms were applied in sequence in five different runs involving different combinations of subcorpora, in which the Input-Texts, Background-Texts and window-sizes were systematically varied (see Table II). Across all five runs, we applied these techniques and examined the word-feature groups containing feature-words with the highest LLR scores. Taking a cut-off of the top-10 in this ranked list we manually identified the news-stories referred to by the feature-groups; these stories were then summarized in our results (see Table III and Figure 1).

Algorithm 2 Grouping word-features using PMI

Def. i represents beginning of window for f
2: **Def.** j represents end of window for f
for (f, i, j) in **T** **do**
4: $features \leftarrow$ Features_in_window(i, j)
for t in $features$ **do**
6: Note: All documents are preprocessed
$D \leftarrow$ Common_documents(t, f)
8: **if** $|D| > 3$ **then**
S $\leftarrow$ pmi(t, f) across docs in D
10: **end if**
if Threshold(**S**) **then**
12: Group(f, **S**)
end if
14: **end for**
end for

Unfortunately, there is no definitive ground-truth for what constitutes a major market-impacting event; one has to rely on the subjective assessments of commentators on the market as to which events were or were not important. So, to assess whether a given event was a major market-impacting event we relied on feedback from the journalists writing about this market. This method gives us some sense for how well the technique is doing in identifying market-impacting events.

Across the five runs we attempted to vary the complexity of the problem by moving from smaller, highly-filtered data-sets (Runs #1 and #2) to much larger, noisier data-sets (Runs#3 and #4) and finally, to a large data-set that was significantly pre-processed (Run#5; see Table II).

TABLE II. FIVE RUNS REPRESENTING DIFFERENT SUBCORPUS COMBINATIONS FOR INPUT- AND BACKGROUND-TEXTS BY DIFFERENT TIME-WINDOW SIZES

Run	*Input*	*Background*	*Window Size*
#1	Web	Web	1 - 10 weeks
#2	Web	Web+CMS	1 - 10 weeks
#3	Web+CMS	Web+CMS	1 - 10 weeks
#4	Web+CMS	Web+CMS	2 - 10 weeks
#5	CMS-Beef	CMS-Beef	1 - 10 weeks

Run#1 uses the smaller 2000-article subcorpus, as Input-Text and Background-Text, based on the articles that were tagged as beef-related on the IFJ's website (i.e., the Web-Subcorpus). So, Run#1 has the benefit of applying the LLR technique to a highly-filtered data-set but one that is limited by the Background-Text being only one year of articles. Run#2 provides a contrast to Run#1 in that it expands the Background-Text to six years of articles (adding in the CMS-Subcorpus).

Runs#3-5 examine larger Input-Texts/Background-Text combinations. Runs #3 and #4 combine the large CMS-subcorpus and the Web-Subcorpus, as both Input- and Background-Texts, to present the technique with a much more varied and noisier data-set (recall, the CMS-subcorpus has articles about all aspects of farming, not just beef-related ones). The only difference between #3 and #4, is that the latter only considers windows between 2 and 10 weeks, thus it excludes any word-features that peak in a 1 week period (this change was designed to exclude transient events that were unlikely to

be important to the market). Finally, Run#5 uses the CMS-Beef subcorpus based on an automated, pre-processing step to reduce the large CMS corpus to beef-relevant articles and then determine the events found from it.

D. Results & Discussion

For each run the top-10 word-features ranked by LLR scores were noted and the word-feature lists found from these were used to identify key news articles and, hence, the referring events. For the most part, Run#1 and Run#2 are a natural group as they cover a recent single-year and Runs#3-5 are a separate group covering a six-year period. Though there were some minor differences in the events identified in different runs we report in detail on the runs that provide the "best" results; they are Run#2 for a 1-year Web subcorpus and Run#4 for the 6-year Web+CMS subcorpus (see Tables III and IV, respectively).

Over the 1-year period, Run#2 compares the highly-filtered Web-Subcorpus against the combined Web+CMS subcorporus and identified several events, 60% of which were identified as potentially major market-impacting events (see Table III and Fig. 1): including (#1) a legal case by HSBC against the TLT livestock export company, (#2) the Horsemeat scandal, (#3) changes in the Ecological Focus Area (which has knock-on effects for EU farm payments), (#4) a major dispute between producers and processors (aka the Factory Dispute 2014), (#7) EU actively pursuing a trade deal with the US in TTIP that could strongly affect the Irish and EU Beef sector and (#9) a new beef trading agreements between Ireland and US surrounding a visit from the US Secretary of Agriculture, Tom Vilsack. However, it should also be noted that this technique also identifies some events that were not market-impacting (e.g., a National Sheep-Shearing Event, Beef Genomics Scheme and articles about lawnmowers). Figure 1 shows the temporal extent of these events based on the weeks over which the LLR score peaks for the word-feature in question. So, for instance, it is interesting to note that Figure 3 there is a large drop in prices between late-2013 and early-2014 that corresponds to the TLT and horsemeat-scandal events for the same period (see events #1 and #2 in Figure 1).

Over the 6-year period, Run#4 compares the larger Web+CMS subcorpus against itself, with the added constraint that the windows are limited to 2-10 weeks (as such, this foreshortening of the window excludes any event that is mentioned only within a single week). Again, this run identifies several events in the top-10 highest scoring terms, of which, around 60% could be considered to be major market-impacting events (see Table IV and Fig.2): including (#1) a legal case by HSBC against the TLT livestock export company, (#5) the Horsemeat scandal, (#6) Irish investors looking to invest into West Missouri which may have an impact on the market in some subtle way. (#7) talks about protests occurring on cuts in farm income, which may effect the market, (#9) represents the roundtable talks, referring to the factory dispute between producers and processors and (#10) the Irish pork crisis from 2008 which involved international recall of pork Irish products due to high dioxin levels. It should also be noted, that again several events were found that were unlikely to impact the market: (#2) human trafficking and women right in Nepal, (#4) festivals with Irish VIPs, and (#8) the Haiti earthquake (though note this could be considered a major catastrophic event in some contexts). Figure 2 shows the temporal extent of these events based on the weeks over which the LLR score peaks for the word-feature in question.

In contrast to the runs on the smaller subcorpora, the three runs on the larger subcorpora found much fewer salient events (i.e., Runs #3 and #5). The best set of outputs arose in Run#4 which excluded events that lasted for a short period of time (i.e., 1 week). Run#5 with the CMS-Beef-subcorpus, has been used to select beef-related articles only, also produced similar, though slightly different results (not reported here).

Overall, it is clear that the LLR technique can find market-impacting events from among the 1000's of articles published about the beef market. But, the results show some false positives (e.g., events about the Haiti Earthquake or drug trafficking in Nepal). Arguably, one could probably live with such false positives, as it is better to find a potentially impactful event than to miss such an event.

Our runs have looked at several filtered variants of the original large corpus (e.g., the CMS-Beef subcorpus). None of these filtering manipulations successfully removed these irrelevant articles. So, we believe that the removal of such false postives is probably best achieved by a post-processing, content check on the article found, to determine if it mentions beef or the beef market, at all.

Finally, it is interesting to note that different sets of market-impacting events are found depending on the number of years covered by the corpus. We when compare the results of the 6-year period versus the 1-year period, different market-impacting events come to the fore. So, some calibaration of the time-window adopted needs to be also considered.

V. Discovering Mundane, Minor Market-Impacting Events

The previous tests using the LLR technique have shown that it can be used to identify major market-impacting events, but this technique tells us nothing about the minor events that routinely change the markets in more subtle ways (e.g., changes in consumer sentiment, stockpiles of goods, inflation rates). To capture such regular, mundane events we need very different techniques that use the broad swathe of word-features used in everyday market commentaries. Specifically, we constructed predictive models following [1], [2], using a classification framework and selected word-features.

This classifier was designed to predict, at a given point in time, whether market prices are currently rising or falling, using the text of news articles at that point in time, as input to the classifier. We give particular attention to feature selection, partly to improve the classifier's accuracy by avoiding overfitting, but also because it is important to know which features from the text data are most predictive of movements in the market price. Whereas the major market-impacting events discussed in the previous section are rare and unusual events, these minor events that shape the market are likely to be reflected in word-features that regularly recur throughout the corpus and have a strong correlation with rising or falling trends in the market price data.

TABLE III. TOP 10 FEATURES FROM RUN#2 USING LLR ON SLIDING WINDOWS OF 1-10 WEEKS ACROSS 2013-2014, COMPARED AGAINST THE 10-YEAR WEB+CMS SUBCORPORUS; WITH ASSOCIATED FEATURE-LISTS, TOPIC DESCRIPTIONS AND LLR SCORES

	Feature	Feature List	Topic Description	LLR Score
1	tlt	tlt receivership garavelli costelloe davide paolo	TLT went into receivership owing 3 million to marts and HSBC bank	602.63
2	horsemeat	horsemeat mcadam	Lessons learned from the horsemeat scandal	346.22
3	efa	efa fallow glas efas equivalence	EFA greening requirement, Crop diversification requirements obligation clarifications on greening	312.98
4	roundtable	roundtable activation roadblock	Tension between Downey and Coveney on Beef price crisis and roundtable talks. Factories abuse farmers on the QPS. Roadblocks impeding trade to Northern Ireland	191.11
5	lawnmower	lawnmower stihl hustler	Advices on Lawnmower and Machinery purchases	149.21
6	bdp	bdp bgs	Applications for the Beef Genomics Scheme (BGS) and Beef Data Programme (BDP)	148.07
7	ttip	ttip	Potential beneficiary of TTIP trade of goods and services between the EU and US	145.71
8	overage	overage underage	Commentaries of cattle prices	127.38
9	vilsack	vilsack	Irish -US beef trading . Visit to Ireland by US Secretary of Agriculture,Tom Vilsack	125.54
10	nematodirus	nematodirus	Warning by Nematodirus Advisory Group, Periods of Nematodirus Larvae prediction. Disease affecting lambs	112.62

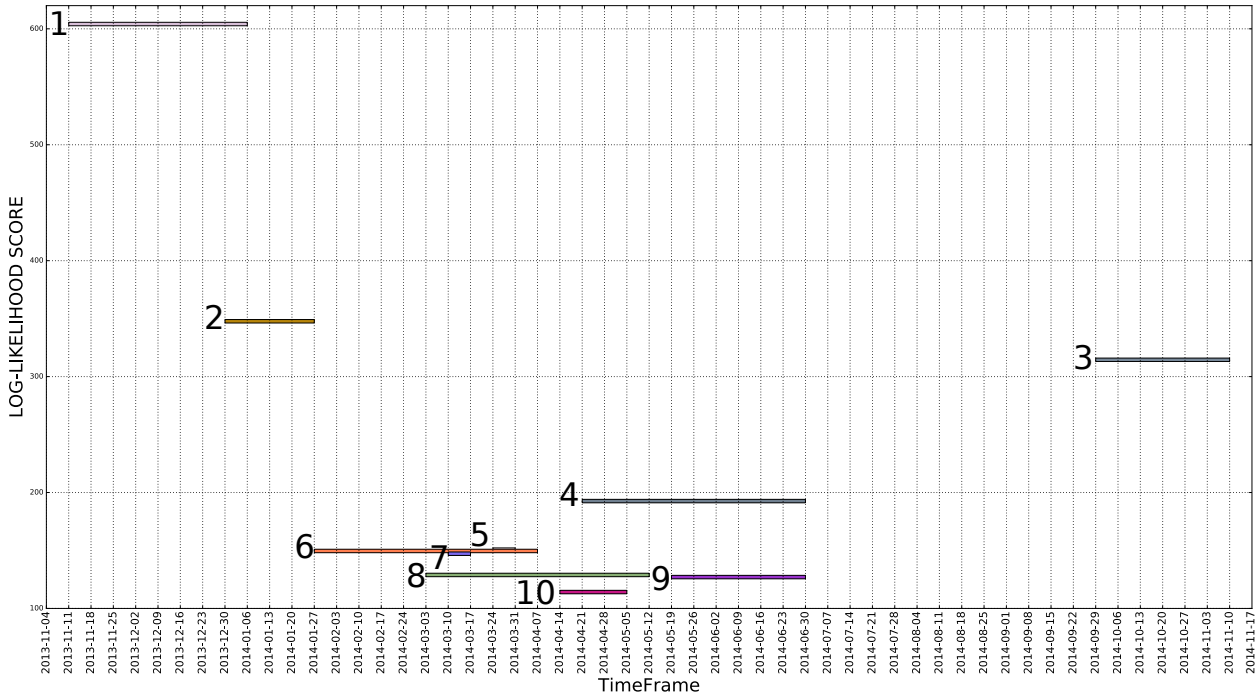

Fig. 1. Time Periods of the Top-10 Events for 2013-2014; Numbers Refer to Word-features and Events listed in Table III

TABLE IV. TOP 10 FEATURES FROM RUN#4 USING LLR ON SLIDING WINDOWS OF 2-10 WEEKS ACROSS 2008-2014, COMPARED AGAINST THE 10-YEAR WEB+CMS SUBCORPORUS; WITH ASSOCIATED FEATURE-LISTS, TOPIC DESCRIPTIONS AND LLR SCORES

	Feature	Feature List	Topic Description	LLRScore
1	tlt	tlt receivership garavelli costelloe davide paolo gore	TLT in receivership owing 3 million to marts and HSBC bank	602.63
2	nepal	nepal shah nepalese widow trafficked dishonour	Inequality for women Nepal Trafficking of women	450.00
3	chlorine	chlorine	Detergent trichloromethane detection in some milk supplies	403.87
4	farmfest	farmfest	FarmFestVisited by VIPs Ministers of Agriculture	393.41
5	horsemeat	horsemeat mcadam	Lessons learned from the horsemeat scandal	346.21
6	missouri	missouri	Irish and New Zealand investors looking to establish farming in South West Missouri	267.54
7	activities	activities ifap	IFA's protests against further Government cuts in Farm Income (2009)	240.98
8	haiti	haiti earthquake	Haiti earthquake Donations	219.60
9	roundtable	roundtable activation roadblock	Tension between Downeyand Coveney, beef roundtable talks. factories abuse farmers on the QPS. Roadblocks on trade to Northern Ireland	191.11
10	pcb	pcb millstream rogan	2008 Irish pork crisis. Dioxin contamination incident	175.52

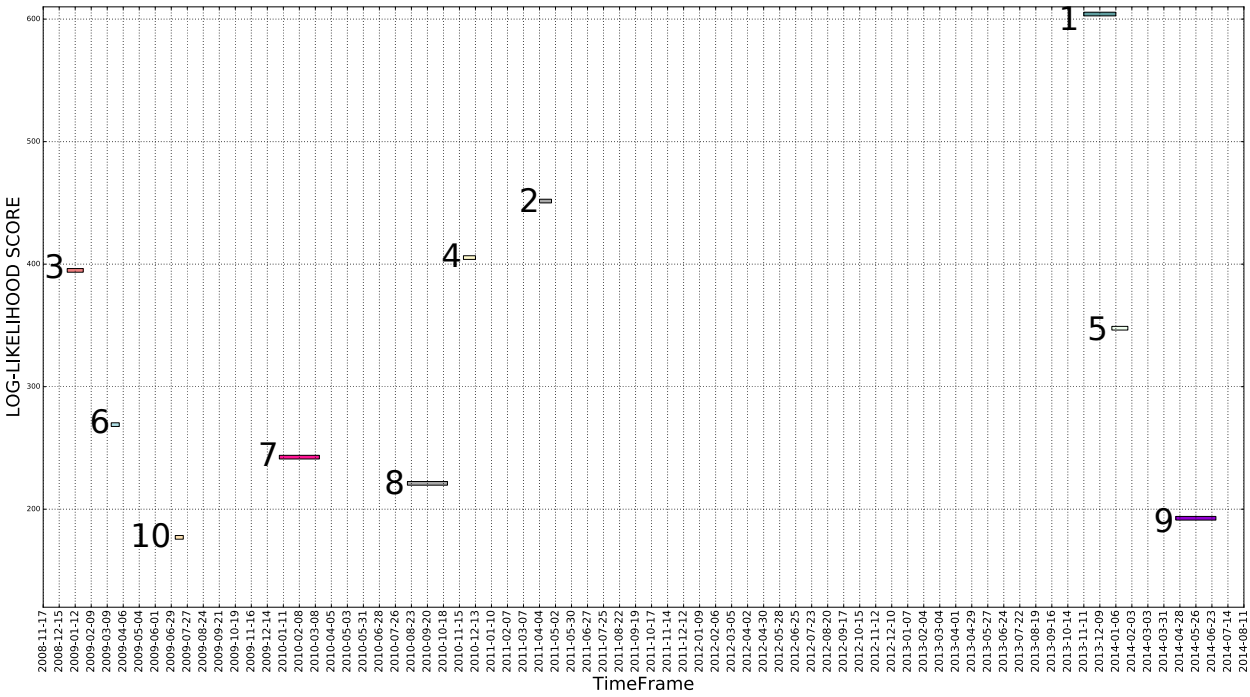

Fig. 2. Time Periods of the Top-10 Events for 2008-2014; Numbers refer to Word-Features and Events listed in Table IV

A. Data Sets & Corpora

To carry out this analysis, we supplement our news corpora with numeric time-series market price data. Our initial data source is a listing of over 2 million individual transactions of livestock at 37 marts across Ireland, covering all sales of cattle between farmers in the country. While our data set contains many details of each transaction (such as the age, weight, and breed of the animal, among others), we are primarily interested in the price and time of the sale. From this list of transactions, we extracted records of all single-animal sales with a nonzero price, and we limited our results to only those from the five busiest marts, in order to limit the variance in prices. From this filtered list, we aggregated and averaged the prices by day, in order to construct a daily-sampled time series. This gives us a single time series representing the mean price per head of cattle sold at market on a given day in Ireland.

The daily sampling gives a time series with 2,019 points spanning the years 2008–2015. We further aggregated this series to give weekly-sampled and monthly-sampled series, which effectively smooth over the daily fluctuations in price. Past work [1], [2] used piece-wise linear segmentation to further smooth the data and categorize each point in time as belonging to either a rising, falling, or neutral trend. However, our experiments using linear segmentation on both the monthly and weekly time series yielded unsatisfying results. Instead we used a straightforward weekly sampling, that matches the weekly publication timeframe of the *Irish Farmers Journal*. The weekly price data is shown in Figure 3. Each week is assigned a binary class label of either "rising" or "falling", based on the change of that price in that week, compared to the price in the prior week.

Finally, we directly align each week of price data to the corresponding week of news articles, as simultaneous alignment has been shown to be effective in prior work [1]. Hence, we used our three constructed subcorpora: (i) *Web-Subcorpus* containing around 2000 articles and 5,700 unique feature-words, (ii) *CMS-Subcorpus* 27,000 articles and 36,000 unique feature-words and (iii) *CMS-Beef-Subcorpus* roughly 2000 articles and 7,500 unique feature-words, the news articles range from 2008–2014 (see section III for subcorpus construction) and the corresponding prices for the same period (i.e., the prices for 2008-2014 from Figure 3). For each subcorpus we constructed a matrix of word frequencies where each row represents a week; the word frequencies for a week are aggregated from all documents published in that week. We then apply a TF-IDF transformation and unit-row normalization to this matrix. This generates a data set with one data point per week, consisting of both the term frequencies extracted from the text corpus and the rising/falling label extracted from the time series, which we use as training input for the classifier. By training a classifier on this data, we are investigating the extent to which textual features extracted from the *Irish Farmers Journal* are able to predict contemporary rising and falling trends in the cattle mart price data.

B. Procedure & Classifiers Used

We use the Python package *sklearn* to evaluate the performance of three machine learning methods against two baselines. The three classifiers selected are known to perform well with textual classification, involving large, word-feature sets: Support Vector Machines (Linear SVM) are known to perform significantly better than other textual classifiers [17], and we included Multinomial Nave Bayes (MNB) and Stochastic Gradient Descent (SGD) for comparison. Our two baselines are dummy classifiers which make predictions based on the output label distribution either stochastically or deterministically (i.e., always choosing the most-frequent label).

For each classifier we trained a series of models with k input features, with k ranging from 2 to N (where N is the total unique word-features in a given subcorpus). For each k value, the most-informative features were selected using *sklearn*'s chi-squared feature selection module. Each model was evaluated using non-shuffled 10-fold cross validation. Note that each fold of data consists of several months of contiguous data points, which minimizes the possibility that temporally-local features could bias the classification of nearby points. It should also be pointed out that while the feature-selection method, using the chi-squared test, is a variant of LLR, the way it is being used here to do feature selection over the classifier data is quite different to how it was used earlier.

The key goal of this experiment was to find the word-features appearing in a given temporal state that accurately predict market movements, where these words reflect the repetitious, minor events that impact the market. We assume that in a corpus with tens of thousands of word types, the vast majority of these will have no relation to market prices, while a small number will be strongly correlated with the rises and falls of cattle prices. For this reason, we experimented extensively with feature reduction methods to identify the most informative lexical features. A common method for feature selection in classification contexts is the chi-squared method, which measures the correlation of each feature with the output class.

C. Results & Discussion

The classification accuracy results are shown for the CMS subcorpus in Figure 4. In the scatterplot, each point represents a single trained classifier; the y-axis denotes the classification accuracy and the x-axis denotes k, the number of features used to train the classifier, which ranges from 2 to 36,000. The blue line represents the Linear SVM Classifier. Around it we can see a thicker green cluster of points that represents performance from SGD Classifier. The Multinomial Naive Bayes Classifier is shown as the yellow line. The plot shows that all three classifiers outperform the two baselines, and all three perform best at around 5,000 features, for this CMS subcorpus. For the Web subcorpus the classifiers perform best at 1000 features and for CMS-Beef-Subcorpus at 918 (Web subcorpus and CMS-Beef subcorpus data are not shown here). Adding more features beyond this point causes progressive overfitting of the data. In addition, the thickness of the different bands illustrates the variance of each classification method: while SGD seems to achieve the highest overall accuracy, it also has a much higher variance profile than the Linear SVM.

Overall the classification performance is good, as the SVM achieves an 88% accuracy for CMS subcorpus (Webs subcorpus - 82% accuracy, CMS-Beef subcorpus - 91%) at predicting rising vs falling trends, showing that textual data

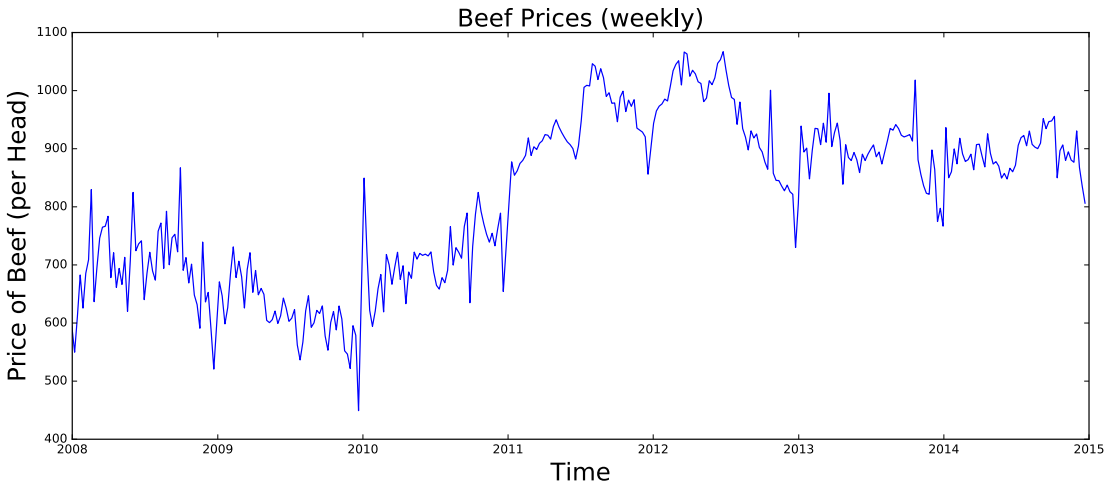

Fig. 3. Price Trends of Mart Beef Prices from 2008-2015: Weekly Aggregated, Average Price Per Head of Cattle in Euros.

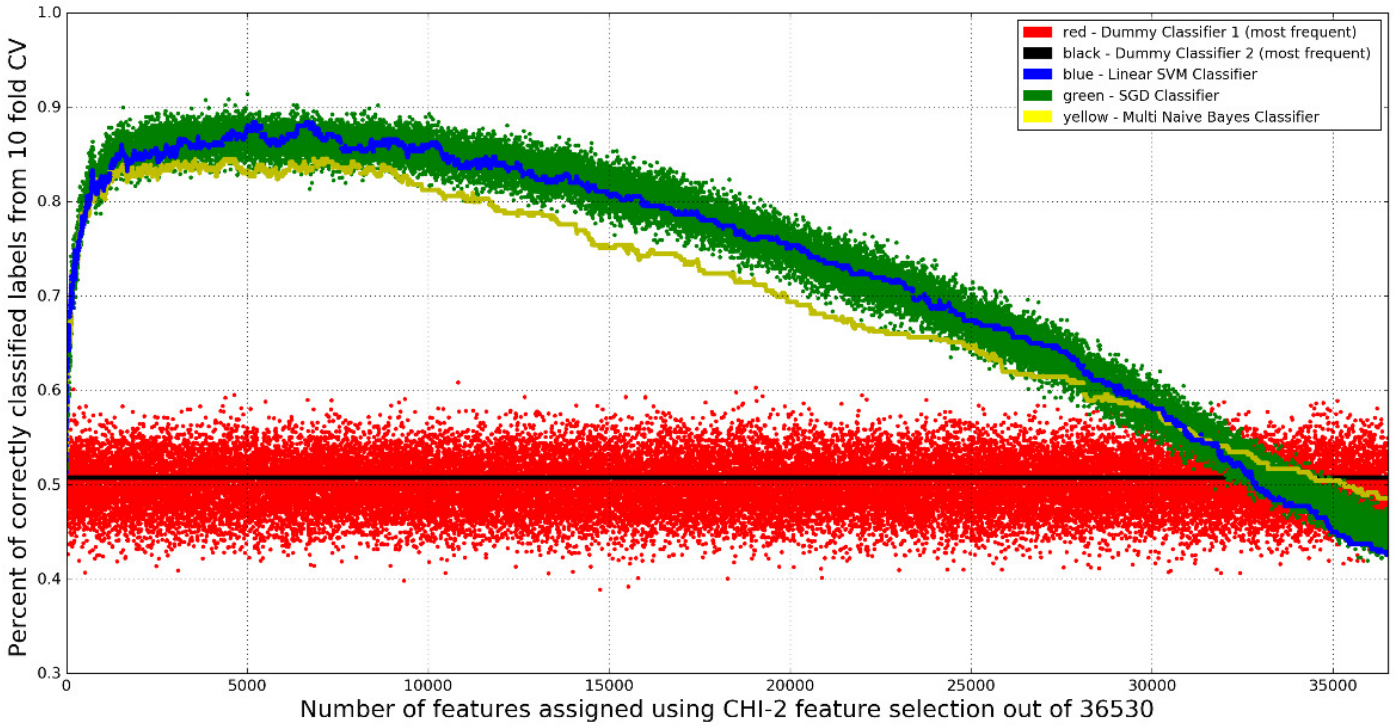

Fig. 4. Results for Different feature selections (from 2–36,530 Word Features) Where Features Were Selected using the X^2 statistic. Each Point Represents the Average Accuracy Score for Each 10-Fold CV Plotted by Each of the Classifiers. Data based on 27.7K Documents Containing 36K Feature-words Covering 2008-2014.

can be successfully used to predict price fluctuations in cattle markets, confirming similar findings of [1] and [2] for stock markets. The feature selection results confirm the hypothesis that a relatively small set of word features (10% of total feature set) are highly correlated with market price movements, and these features are most useful for price trend prediction.

Comparing Word Features from LLR & SVM. Notably, when we compare this list of highly-predictive word-features (in each subcorporus) against those discovered by the time-window LLR method in section IV, for the same time period (2008-2014), we find that there is a low correlation (Web Sub-corpus r=0.34, CMS subcorpus r=0.40, CMS-Beef subcorpus r=0.38) between the two, suggesting that these two methods effectively identify quite distinct sets of market-impacting word-features. It is interesting to note that LLR does find some similar attributes, however these attributes are not the ones we found to be associated with major once-off events. We applied an alterative observation restricting feature-words to

the Top-30 Major events with the highest score from the LLR method. This will allow us to identify whether the high major events relate to the minor event model. This resulted in the following correlations for the Web subcorpus - r=-0.12, CMS subcorpus r=-0.06 and CMS-beef subcorpus r=0.43. Intuitively, this is what we expected to find: in order for a word-feature to be an effective predictor in the classification setting, it must appear multiple times in the data corresponding to the rising/falling market trend (for example, in our cross-validation experiments a feature must at minimum appear in two cross-validation folds in order to have any predictive power). This contrasts with the major-event word-features identified by the LLR method, which by definition appear highly frequently in one time window but infrequently outside of that time window.

VI. Conclusions

In this paper, we have considered two different methods that uses the text of news articles to find events that impact a market (i.e., the Irish Beef Market). We distinguished between major, once-off events that radically impact markets and the minor, more everyday events that have gentler impacts on prices. Both types of events are represented in text sources, and both have the potential to impact markets, but the two types of events have different properties and thus lend themselves to different methods.

For identification of mundane, minor market-impacting events we have shown, using three differently filtered sub-corpora of news articles, that an SVM classifier can predict changes in market price data with very high accuracy. We also showed that the LLR method can identify rare, standout and exceptional events (like the "horsemeat" event in beef markets) that also affect markets.

However, our comparative studies have shown these two classes of events are distinct from one another. There is a degree of conflict between these methods as neither one can adequately meet the task done by the other. No single method appears to be able to find both types of market-impacting events. So, our conclusion would be that any complete future system that aims to identify both type of events from text sources would need to use these two methods, albeit separately.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289. We would also like to thank *The Irish Farmers Journal* and Padraig Foley for their support in this project.

References

[1] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, "Mining of Concurrent Text and Time Series," in *Proceedings of the 6th ACM SIGKDD Knowledge Discovery and Data Mining*, 2000, pp. 37—-44.

[2] G. Fung, J. Yu, and H. Lu, "The Predicting Power of Textual Information on Financial Markets." *IEEE Intelligent Informatics Bulletin*, vol. 5, no. 1, pp. 1–10, 2005.

[3] A. Gerow and M. T. Keane, "Mining the web for the "voice of the herd" to track stock market bubbles," in *In Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, ser. IJCAI'11. AAAI Press, 2011, pp. 2244–2249.

[4] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306–324, 2015.

[5] G. Kumaran and J. Allan, "Using names and topics for new event detection," *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 121–128, 2005.

[6] X. Wang and A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 424–433.

[7] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Association for Computational Linguistics, 2010, pp. 181–189.

[8] M. Osborne, S. Petrovic, and R. McCreadie, "Bieber no more: First Story Detection using Twitter and Wikipedia," *SIGIR 2012 Workshop on Time-aware Information Access*, 2012.

[9] J. Allan, V. Lavrenko, and H. Jin, "First story detection in tdt is hard," in *Proceedings of the Ninth International Conference on Information and Knowledge Management*, ser. CIKM '00. ACM, 2000, pp. 374–381.

[10] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Data Mining and Knowledge Discovery*, vol. 7, pp. 373–397, 2003.

[11] K. K. Mane and K. Börner, "Mapping topics and topic bursts in PNAS." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl, pp. 5287–5290, 2004.

[12] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the Dynamics of the News Cycle," *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 497–506, 2009.

[13] R. Swan and D. Jensen, "TimeMines: Constructing Timelines with Statistical Models of Word Usage," *In ACM SIGKDD 2000 Workshop on Text Mining*, pp. 73–80, 2000.

[14] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, pp. 61–74, 1993.

[15] D. Greene and P. Cunningham, "Producing accurate interpretable clusters from high-dimensional data," *Knowledge Discovery in Databases: PKDD 2005*, 2005.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[17] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization," *Information Retrieval*, vol. 1, no. 1, pp. 69–90, 1999.