Analyzing Discourse Communities with Distributional Semantic Models

Igor Brigadir
Insight Centre
University College Dublin
Ireland
igor.brigadir@insight-centre.org

Derek Greene
Insight Centre
University College Dublin
Ireland
derek.greene@ucd.ie

Pádraig Cunningham
Insight Centre
University College Dublin
Ireland
padraig.cunningham@ucd.ie

ABSTRACT

This paper presents a new corpus-driven approach applicable to the study of language patterns in social and political contexts, or Critical Discourse Analysis (CDA) using Distributional Semantic Models (DSMs). This approach considers changes in word semantics, both over time and between communities with differing viewpoints. The geometrical spaces constructed by DSMs or "word spaces" offer an objective, robust exploratory analysis tool for revealing novel patterns and similarities between communities, as well as highlighting when these changes occur. To quantify differences between word spaces built on different time periods and from different communities, we analyze the nearest neighboring words in the DSM, a process we relate to analyzing "concordance lines". This makes the approach intuitive and interpretable to practitioners. We demonstrate the usefulness of the approach with two case studies, following groups with opposing political ideologies in the Scottish Independence Referendum, and the US Midterm Elections 2014.

1. INTRODUCTION

In linguistics and social sciences, Discourse Analysis is concerned with analysis of naturally occurring language use and patterns. Van Dijk [32] defined *Critical* Discourse Analysis (CDA) as:

> "...a type of discourse analytical research that primarily studies the way social power abuse, dominance, and inequality are enacted, reproduced, and resisted by text and talk in the social and political context."

This paper presents techniques for Critical Discourse Analysis applicable to opposing communities. A *discourse community* is a group of people sharing a set of basic values, assumptions and ways of communicating. Porter [27] offers a definition of a discourse community as:

> "...a local and temporary constraining system, defined by a body of texts (or more generally, practices) that are unified by a common focus. A discourse community is a textual system with stated and unstated conventions, a vital history, mechanisms for wielding power, institutional hierarchies, vested interests, and so on."

We are primarily concerned with the changes between, and within discourse communities over time. We explore how different political groups unified by a common focus (*i.e.* discourse communities) present themselves as defined by the body of text they generate on Twitter. This work proposes using word similarities from statistical semantics grounded in the Distributional Hypothesis, popularized by Firth [12] and adopts elements of the discourse-historical approach [33], a methodology that is problem-oriented, interdisciplinary, and recommends movement back and forth between theory and empirical data.

Current methods drawn from corpus linguistics that are used in critical discourse analysis, usually rely on keyword extraction and manual examination of "concordance lines" or "key words in context" (i.e. sorted and aligned lists of words with their surrounding contexts) but can also include topic modeling approaches. Emphasis is placed on generating insights into the ways in which the structures of text or speech relate to social and political contexts rather than on any particular approach.

The nature and volume of tweet text makes these approaches challenging for several reasons. Firstly, poor sampling can lead to raw frequency counts of words to be skewed. Terse style and Twitter-specific use of user account mentions, hashtags and other media entities can also cause problems for methods that rely on frequency counts. Secondly, the sheer number of tweets available often makes close reading intractable, while distant reading techniques that look at the entire corpus can hide interesting periods of change and dynamics between communities.

The changes between and within communities with opposing political ideologies, manifest themselves as shifting distributional semantic similarities between words. We suggest that these changes can be quantified using Distributional Semantic Language Models (DSMs). The differences between semantic models derived from text produced by certain discourse communities can offer practitioners useful tools for CDA. These tools are more aligned with what Fairclough [11] calls *textually oriented discourse analysis*, examining how "the mode of language...identified as constitutive of

WebSci '15, June 28—July 01, 2015, Oxford, United Kingdom

ACM 978-1-4503-3672-7/15/06$15.00
DOI: http://dx.doi.org/10.1145/2786451.2786470.

power in modern society... is received and appropriated by those who are subjected to it". Concretely, language models are used in this work as a supporting corpus-driven technique, providing entry points for further, more detailed analysis, allowing researchers to investigate how language use is shaped by political objectives.

As the primary contribution of this paper, we propose a novel application of distributional semantic models for CDA, where constructing a DSM or word space model can be related to Key Words in Context (KWIC) analysis—a well-established, qualitative approach familiar in corpus-assisted CDA. We evaluate the approach on two case studies from Twitter, each involving two distinct communities with differing political viewpoints—the 2014 Scottish Independence Referendum and the 2014 US Midterm Elections. In addition, we also provide reusable data sets of tweets for these case studies.

2. RELATED WORK

A framework for using Corpus Linguistic methods for Critical Discourse analysis is presented by Baker et al. [3]. This paper takes a similar position, arguing that since CDA lacks a concrete set of techniques for performing analysis, novel approaches can be made available to practitioners.

Social scientists and CDA practitioners are increasingly looking to social media as a rich source of data. Current corpus-based approaches and tools involve manual inspection of keyword frequency lists and reading concordance lines. Collocation analysis offers "a suitable vehicle for the discoursal presentation of a group" [4] but using plain frequency for collocation extraction yields general, uninteresting terms [30]. The methods proposed in this paper are not related to Rhetorical Structure Theory (RST), a method for discourse *parsing* [23], concerned with coherence of multi-sentence texts. In contrast, we consider similarities at the word level, rather than sentence level.

2.1 Political Discourse on Twitter

An analysis of political discourse on Twitter by Zappavigna [34] suggests that users appear to bond around the act of collectively witnessing moments they perceive to be important to their cultural history, while politicians often use Twitter as a means of fostering engagement with others, offering positive evaluations of themselves and their parties. In our case studies, this promotional style adopted by official campaign accounts is also evident.

An in-depth study concentrating on politicians on Twitter is presented in [22]. Methods common in Information Retrieval have been applied to theoretical sociological constructs, deriving measures of "Cultural Similarity", Rank Biased Overlap measures for "Cultural Reproduction" and several others. The type of conversational practice (or discourse) examined included analysis of hashtags, retweets and mentions. Political polarization on Twitter is investigated by Conover et al. [10] through the analysis of mention and retweet interactions in the previous 2010 US Midterm elections. The notion of "content injection" is also revealed using our proposed methods, although this is more pronounced in groups containing regular supporters of a particular ideology, rather than official function accounts such as campaign accounts of prominent politicians.

Related work that does not use Twitter data but deals with similar themes, includes: summarizing contrastive views with augmented summarization techniques [26], performing comparative text mining and ideology classification with a topic modeling approach [8], and a network analysis approach for quantifying political polarity of individuals [1]. The problem of political alignment on policy issues, which is often cast as a classification task, is outside the scope of this work.

2.2 Distributional Semantic Models

Recently, *word2vec* [24] has been widely used to generate useful representations of words using a Neural Network Language Model (NNLM). This distributional semantic model offers efficient training times and performs well on a variety of semantic and syntactic word similarity tasks. A comparison of distributional semantic models that involve context prediction and context counting is performed in [5]. Models were compared using a number of widely-used syntactic relatedness, synonym, concept categorization and analogy tasks. Context-predicting models, such as *word2vec*, were shown to perform better than context-counting variants.

2.3 Linguistic Shift

Measuring linguistic shift with an information theoretic approach is explored by Juola [18]. Using a corpus of several decades of National Geographic publications, changes in language were not only perceptible algorithmically, but are also not uniform over time, suggesting that some periods of time are more actively changing than others.

Kulkarni et al. [20] developed a method for detecting significant linguistic shift in the meaning and usage of words, employing a DSM to construct a time series of word usage and a mean shift change point detection algorithm to estimate when this change occurs. In contrast to our work, the timespan involved is larger, covering two years for Twitter data and several decades for the Google n-gram set. However, a comparison between different clusters or communities is not considered by the authors. In [15], a distributional similarity approach is compared to a relative frequency based approach, using two Google n-gram corpora from 1960s and 1990s.

Another relevant approach that explores changing word meanings over a long period of time is described by Basile et al. [6]. Neighborhoods of words are examined across several decades of Italian books and the ACL Anthology Network data set. Unlike previous work, in this study we focus on shorter time spans, seeking to identify changes between and within communities, as opposed to simply looking at a changes across an entire corpus.

3. PROPOSED METHOD

Common corpus-assisted techniques for discourse analysis include comparative word frequency lists, keyword extraction, and concordance lines or KWIC—showing the surrounding context of a keyword of interest (See Table 1). Typically, results are presented as raw or normalized counts derived from the corpus, along with a qualitative assessment that involves close reading of a selection of material. In collocation analysis [4] the most frequent co-occurrences may not be the most useful for CDA. To address the drawbacks of frequency-based approaches [30], we propose the use of a distributional semantic model that computes vector representations of words. The rationale here is that DSMs reveal different types of similarity and relatedness useful for CDA.

...	#nothanks #indyref foremost authority on north sea oil throws doubt over snp prediction for the future of ...
...	foremost expert felt he had to speak out to warn of oil depletion #indyref cant rely on oil to deliver public ..
...	points out what john swimney said about volatility of oil @sygazette debate an injustice in one part of the uk ...
..	welfare health education and pensions costs 40 billion oil revenue 3 billion #no gb on the #nhs the ties that bind

Table 1: Example concordance lines, from Scottish Referendum tweets containing the word "oil".

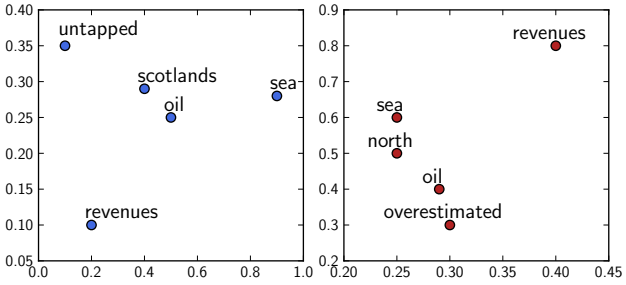


Figure 1: An illustrative example of word spaces with only 2 dimensions. Words are represented by vectors with 2 components, x and y values.

Motivated by results from Baroni et al. [5], we use a context-*predicting* distributional semantic model as opposed to a context-*counting* model. The task requires a good estimation of word similarity, as well as *association*. As a concrete example, the words "fields" and "oil" are not synonymous, describing two different concepts, but are *related* in the context of fossil fuels. Likewise, "oil" and "crude" are synonymous[1]. Both similarity and relatedness are useful to consider for CDA. It is important to note that DSMs have previously been evaluated with this distinction [16].

Figure 1 shows a toy example of two word spaces. In the space on the left, the word "oil" is represented by a vector $[0.50, 0.25]$, and on the right $[0.29, 0.40]$. The components of the vectors do not represent actual counts or occurrences, but after training a DSM, words that are more *related* are closer to one another. We can compare the two spaces by looking at what words are closest to one another in each space. In practice, words are represented by hundreds of dimensions.

3.1 Discourse-historical Approach

A suitable framework for CDA is Wodak's Discourse–Historical Approach (DHA)[33]. DHA emphasizes the interpretation of discourse in its historical and cultural contexts. Four levels of context are suggested:

1. The immediate *co-text* for a particular linguistic feature found in the text: Involves analysis of the text itself. Our proposed approach provides most benefit at this level, suggesting entry points for further, more in-depth analysis.
2. Other texts and discourses that the text draws upon: In a corpus of tweets, this level would include analysis of media linked and referred to in tweets.
3. Conditions of text production, distribution and reception. This level is constrained by the platform, particularly in terms of distribution and reception with use of hashtags and mentions [34].
4. Wider social, political, economic and cultural contexts.

[1]The word "oil" is frequently dropped from the phrase "crude oil", especially in length-restricted posts like Twitter.

For analysis of groups with opposing political ideologies, DHA recommends six *discursive strategies* for identifying ideological positioning, summarized with questions below:

1. Nomination: Constructing in-groups and out-groups via membership categorization. How do different groups categorize themselves and opposing groups? Do these change over time?
2. Predication: Labeling social actors positively or negatively. How are key individuals represented by different groups? How is this reproduced in tweet text?
3. Argumentation: Justifying positive or negative attributions, political inclusion or exclusion.
4. Framing: Expressing involvement through reporting, narration of events and utterances.
5. Intensification / Mitigation: Modifying a proposition by intensifying or mitigating force of utterances.

As a motivating example, consider discursive strategies used in the following tweets:

> Why do Nats want Scotland to be one of Europes vulnerable, marginal economies? We truly are #BetterTogether #IndyRef
>
> Nationalist lies over oil @YesScotland @UK_Together #idyref No Boom No Oil Bonanza #ProjectFear

We can manually identify several interesting keywords in the text of these tweets: "nats"—nationalists (nomination, membership categorization), "lies" (predication, labeling social actors negatively), "bonanza" (intensification).

Closely examining a large volume of tweets this way is impractical, and while some levels of context require a close reading, analysis of the text itself can be performed at scale, using of corpus driven approaches. A concrete example of where DSM approach can help DHA, is in exploring *discursive strategies* in racial, national and ethnic issues. Questions like "How are persons named and referred to linguistically?", "What traits, characteristics, qualities and features are attributed to them?" prescribed in DHA to explore "positive self" and "negative other" presentations can be answered with a combination of examining nearest neighborhoods of words, and closer reading of selected tweets.

As a starting exploratory step for our analysis, we examine the different communities using a small selection of words representing topics of interest which are known *a priori*. We then examine some *discursive strategies* in the communities. This is firstly performed across communities over the entire period, and then in more detail, looking both within and across communities over shorter time periods. We then expand this set of words, by examining the k-nearest neighboring words for the communities, in order to discover interesting commonalities or differences between them. Restricting the nearest neighbor search to consider either words or hashtags alone could potentially provide alternative lines of inquiry [10]. However, we follow a more general approach, allowing for a mix of words, mentions and hashtags to appear, but excluding URLs which appear in tweets. The method is

an iterative, word-level approach to critical discourse analysis that alternates between exploration, and close reading.

To summarize the prescribed process: 1) select initial candidate words of interest, 2) examine the word change "profile" visualized as a trend, 3) examine word neighborhoods, 4) retrieve relevant tweets for a more qualitative, closer reading of texts, and 5) repeat the process armed with new keywords or hypotheses.

3.2 Distributional Semantic Approach

Mikolov et al. proposed a DSM that performs well on a variety of syntactic and semantic relatedness tasks [25]. The *skip-gram* training process learns word representations that are useful for predicting the nearby words (the context). From a sequence of words $(w_1, w_2, \ldots w_T)$, the objective maximizes the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

where the context size c is the number of words before and after the target word w_t. It is interesting to note that the mechanics of "key words in context" analysis roughly maps to the training objective of the DSM approach.

When analyzing *collocates* (*i.e.* words that co-occur more often than would be expected by chance), Mutual Information (MI) is a commonly used association measure. For instance, the popular *AntConc* concordance tool [2] uses Mutual Information to rank collocates.

Pointwise Mutual Information (PMI), proposed by Church and Hanks [9] is another widely-adopted association measure. If two outcomes x and y have probabilities $P(x)$ and $P(y)$, then their Pointwise Mutual Information $PMI(x, y)$ is defined as:

$$PMI(x; y) = \log \frac{P(x, y)}{P(x)P(y)}$$

Levy and Goldberg show that the *skip-gram* training process implicitly factorizes a word-context matrix, the cells of which are Pointwise Mutual Information (PMI) of word and context pairs, shifted by a global constant[21].

The *word2vec* model, trained with *skip-gram* is a good choice for the purpose of quantifying semantic & syntactic similarity between words. Similar models can also be used, but *word2vec* was chosen for having a good trade-off between performance and training time.

3.3 Comparison of KWIC & k-NN Lists

Manually examining KWIC lists will often be unfeasible for larger corpora, such as a collection of tweets. Sampling a selection can introduce bias, while a *close reading* of each and every document in a collection is impractical. Collocates are generally useful for "a semantic analysis of a word" [28]. Examining these collocation patterns can provide a corpus driven tool for CDA, as used in an analysis of the representation of refugees, asylum seekers, and immigrants in the UK press [4].

Our approach can be used to discover similar patterns, where predicting a set of contexts given a word can be interpreted as an aggregation of concordance lines, drawing an analogy between the training objective and KWIC analysis familiar to practitioners. Table 2 shows how different measures of association can provide different collocates.

Rank	*Frequency*	*Mutual Information*	*word2vec*
1	important	scotenergynews	untapped
2	#indyref	pegging	revenues
3	tank	kuwaits	pouring
4	oil	headlined	#clairridge
5	#scotdecides	@conhome	recoverable
6	#yes	kindness	300bn
7	#westcoastoil	exploration	rig
8	thousands	@yuillnoodz	bonanza
9	#voteyes	@wynnscottishsun	#northseaoil
10	north	@wulliekane	asianomics
	...	...	...

Table 2: Comparison of collocates considering 5 words before and after the word "oil" from 7 days of 'Yes' vote supporters using raw frequency, Mutual Information, and word2vec.

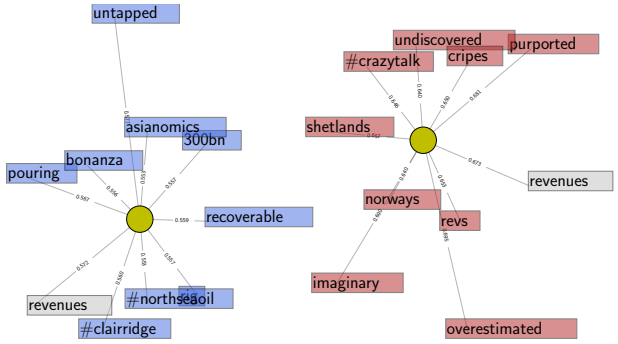


Figure 2: A sample k-nearest neighbor graph for the word "oil" for Scottish 'Yes' and 'No' voters. Words unique to a community are colored, words in gray are common to both communities.

The advantage of using a *word2vec* model in this case is that searching the trained model for nearest neighboring words is extremely fast, and provides meaningful results, without over-promoting highly-rare or highly-frequent terms. A drawback to our approach is that it requires a relatively large, pre-processed corpus. Applying these corpus-assisted techniques over a stream of documents can reveal more nuanced changes in discourse. These changes can potentially be related to external events, or can serve to quantify the evolution of a discourse community over time. Rather than examining the social network structure of different communities, k-nearest neighbor graphs (k-NNG) can be used to examine distinctive linguistic similarities and differences between discourse communities. As an example, Figure 2 shows the word neighborhoods of the word "oil" from two different communities in the Scottish Referendum campaign. Nodes are other related words, and edge lengths are inversely proportional to the cosine similarity of each word to "oil" in the community-specific word space.

3.4 Accounting for Change Over Time

Since discourse communities are temporary systems defined by a body of texts [27], we must account for time in the models. As *word2vec* does not account for a temporal dimension in texts, we propose splitting the data set of each community into windows, each covering different time periods. The conversion of continuous data streams, such as

content from social media platforms, to discrete windows has been a common strategy in the analysis of online communities [29].

Separate models are then trained for a number of fixed length windows of tweets, creating different models for each window and each community. Models are trained using the *skip-gram* architecture, with vector size 300, context window of 5 words, for 30 iterations on time windows spanning 7 days. Due to the stochastic nature of *word2vec* training, and the different training sets, the various word spaces created are not directly comparable. However, we are not interested in the resultant model representation, but rather the relationships between words that can be interpreted by practitioners. Therefore, for a given word, we retrieve its k nearest neighbors in the model, and present these for consideration. This is analogous to the way in which a KWIC analysis presents concordance lines. From this, quantifying the changes between time windows and communities can now be accomplished by comparing the *word neighborhood* of a particular word in different spaces—*i.e.* the similarity of the word's k nearest neighbor lists in each model.

3.5 Average Jaccard for k-NN

To compare word neighborhoods, we require a suitable similarity measure. The k-nearest neighbors of words in a word space, when viewed as rankings, are incomplete (*i.e.* not all words are covered), top-weighted (*i.e.* top ranked words are more important), and indefinite (*i.e.* choice of k is arbitrary). A desirable measure should account for these properties.

The *Average Jaccard* (AJ) measure from Greene et al. [14] used for comparing ranked lists has the required properties. Though distinct, AJ can be related to *cultural reproduction* [22], as both measure a form of rank-biased overlap. We calculate the AJ scores between the k nearest neighboring words from two spaces. The two spaces can either be two time windows from the same community, or the same time window from two different communities.

As AJ is top-weighted, increasingly higher values of k have a decreasing influence on the overall score. The choice of k is largely influenced by the need for rankings to be examined by practitioners. In both case studies presented later, we set $k = 30$, but this parameter can be varied to consider more distant words.

Jaccard similarity between two sets is defined as the size of the intersection divided by the size of the union. The Average Jaccard (AJ) between two ranks A and B to depth k is defined as the average Jaccard scores between subsets of d top ranked words in two rankings, where d is $d \in [1, k]$.

$$AJ(A, B) = \frac{1}{k} \sum_{d=1}^{k} J_d(A, B)$$

where $J_d(A, B) = \frac{|A_d \cap B_d|}{|A_d \cup B_d|}$ and A_d,B_d are the heads of lists up to depth k. See Table 3 for a worked example.

4. DATA SETS

Various selection criteria exist for gathering Twitter corpora for the study of politics and political discourse. A recent survey [17] offers a comprehensive overview of data sources and collection techniques. Gathering all tweets from a subset of users was shown to provide a much richer and trustworthy source of data as opposed to a random sample of tweets from all users [13]. Problems arising from monitoring social media based on the pre-selection of specific hashtags or keywords have also been discussed in the literature [31].

k	*Rank A*	*Rank B*	*Jaccard at k*	*AJ at k*
1	untapped	untapped	1.00	1.00
2	revenues	yada	0.33	0.66
3	pouring	reserves	0.20	0.51
4	#clairridge	bonanza	0.14	0.42
5	300bn	revenues	0.25	0.39
	...	...	...	...

Table 3: Average Jaccard values at different values of k, comparing the word neighborhoods of "oil" from two consecutive weeks from 'Yes' voter tweets.

Rather than relying on keyword or hashtag searches, for our experiments we gathered all available tweets for a fixed subset of users. For both the Scottish Referendum and US Midterm Elections, users were first selected by their "official function"—*i.e.* politicians, campaign accounts, political organizations. Additional accounts included in each set are detailed in 5.1 and 6.1. During data collection, users are automatically notified when added to a Twitter list, and have the ability to remove themselves by "blocking" the account used to add them, or by making their account private. Several accounts were either deleted or made private during and after the data collection period. For pre-processing, common stop words, those words occurring less than twice, and URLs are removed from tweet text. The default NLTK English stop word list[2] was expanded to include several Twitter-specific function words such as "ht", "via", "mt". The data set was post-processed to honor deletion requests and user privacy settings. A summary of the data is shown in Table 4.

Community	*Users*	*Tweets*	*Total Words*	*Date Range*
Scotland Yes	618	799,096	12,551,654	11-Aug to 19-Oct
Scotland No	610	570,024	8,957,721	11-Aug to 19-Oct
Democrat	942	89,296	1,404,737	10-Oct to 20-Nov
Republican	997	80,840	1,209,197	10-Oct to 20-Nov

Table 4: User, tweet and word counts for Scottish and US data sets. Date ranges are in 2014.

The sets of tweet IDs and users are available, together with tools for retrieval to reconstruct the data set[3]. While classifying polarity and party affiliation is outside the scope of this study, this data set potentially offers a useful ground truth for such tasks.

5. CASE STUDY: 2014 SCOTTISH REFERENDUM

The Scottish Independence Referendum, which took place on 18th of September 2014, decided Scotland's membership in the United Kingdom political union. The single question posed by the referendum—"Should Scotland be an independent country?"—generated considerable debate on social media platforms in the weeks before the vote. Both the official *Yes Scotland* and *Better Together* (No vote) campaigns were established in 2012, with the date of the referendum set in March 2013, and legislation passed in November 2014.

[2] `http://www.nltk.org/nltk_data/`

[3] `http://dx.doi.org/10.6084/m9.figshare.1430449`

While the lifetimes of these campaigns were long, the majority of activity occurred within weeks of the referendum. We consider tweets over a time span of 10 weeks (11 August to 19 October 2014), for communities of 'Yes' and 'No' supporters.

5.1 Scottish Voter Communities

An initial seed list of Twitter accounts belonging to "registered campaigners" on the Scottish Independence Referendum Electoral Commission was built. As the number of these *official function* accounts was small, additional accounts were added to the set based on public Twitter lists the seed accounts were members of. Parody accounts, nonpartisan organizations, and users with private accounts were removed. To be included as a "Yes" or "No" supporter, users had to self-identify through prominent use of campaign profile banners (party logos and campaign icons in profile images were popular with both sides), explicitly stating an affiliation in their user descriptions (*e.g.* using `#BetterTogether`, `#iVotedYes` etc.), and actively engaging with referendum topics.

Data from Twitter showed the 'Yes' campaign was dominant in terms of volume and participation, skewing some predictions and on-line opinion polls in their favor[4]. Polls that relied on interviews showed more support for a 'No' vote[5]. Ultimately, Scotland remained part of the United Kingdom, the 'No' vote gathering 55.3% and 'Yes' 44.7%, with a turnout of 84.6%, one of the the highest recorded for a referendum or election in the UK.

Key issues in the campaign included: EU membership and currency, health care, education and research funding, Scotland's renewable energy and north sea oil revenue, NATO membership, and the issue of British Trident nuclear missile system on Scottish territory. Using the analysis methodology proposed in Section 3, an initial set of words relating to these issues was selected. This was followed by an exploration step, adding related words, and removing those words that did not feature prominently in either community.

Yes Week 5	*No Week 5*	*Yes Week 6*	*No Week 6*
untapped	fields	norways	revenues
yada	rigs	revenues	ridge
reserves	rosy	reserves	**bonanza**
bonanza	revenues	147bn	sea
discoveries	downgrade	discoveries	4bn
bloddy	**bonanza**	chevron	offset
1billion	inflated	**bonanza**	estimates
...	...	...	...

Table 5: Yes and 'No' Voter word neighborhood of "oil" corresponding to Figure 3 in weeks 5 and 6. Top 7 words are shown from the 30 used in analysis.

While neighbor graphs such as Figure 2 can be illustrative for small examples, a network visualization of larger word spaces will quickly become an uninterpretable "hairball". The differences between time windows are also not evident. As an alternative, we suggest a trend visualization to compare the similarities between communities and time windows. Figure 3 provides a sample comparison of word neighborhoods between 'Yes' and 'No' voters over 10 weeks. A point in the trend is the *AJ* similarity between

[4] `http://blog.twitter.com/en-gb/2014/indyref-at-the-polls`

[5] `http://survation.com/?s=Scottish+Referendum`

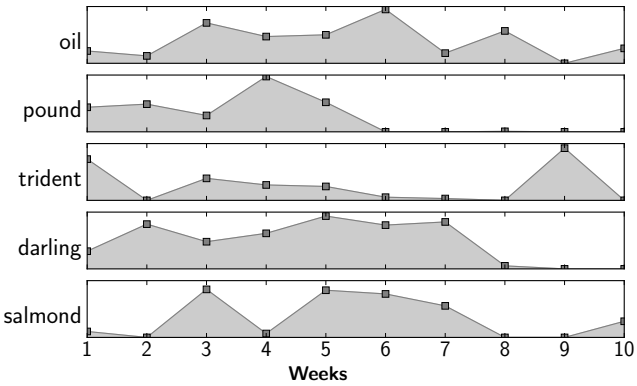


Figure 3: Trends illustrating the Average Jaccard similarity changes between word neighborhoods for the 'Yes' and 'No' communities, over 10 weeks of the Scottish Referendum campaign. High similarity indicates agreement between communities, while low similarity indicates greater difference in how a word is used.

word neighborhoods from different communities, for a window of a single week. A more detailed understanding of a given point can be supported by looking at the corresponding word neighborhoods for each community for that the time window, as illustrated by the ranked lists of words in Table 5. This can subsequently be used to retrieve the relevant tweets for a closer reading and analysis.

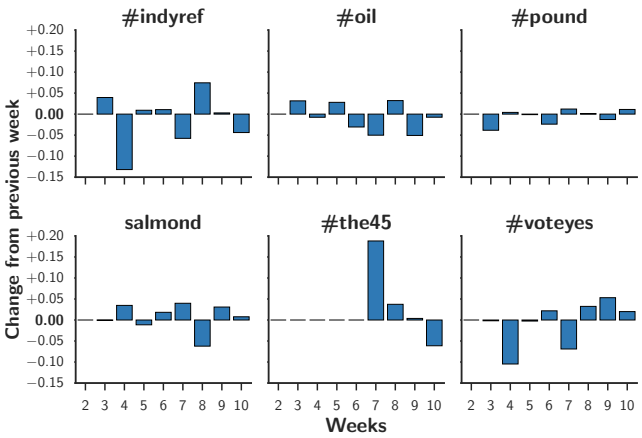
5.2 Discursive Strategies

Predication is an important discursive strategy with the objective of labeling social actors, used for reinforcing the construction of "us" and "them" between the 'Yes' and 'No' voter communities. These "positive self" and "negative other" presentations can be extracted from the nearest neighboring terms used to refer to political leaders. Table 6 shows a selection of nearest neighboring words from 'Yes' and 'No' voters, for terms that refer to political figures central to the campaigns. The k nearest neighbors for Table 6 are derived from a model trained on tweets in the entire date range, before during and after the referendum.

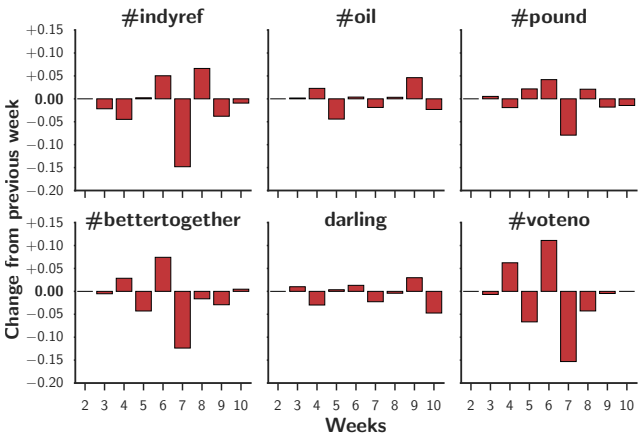
Alex Salmond		**Alistair Darling**	
Yes Voters	*No Voters*	*Yes Voters*	*No Voters*
lucid	frantical	adversarial	commanding
authoritative	misdirection	bluffing	principled
statesman	fraudster	dismissive	quizzing

Table 6: Sample nearest neighbor words for "salmond", @alexsalmond, #alexsalmond, and "darling", @togetherdarling, #alistairdarling

The *Referential / Nomination* strategy in the discourse historical approach is used for constructing in-groups and out-groups, and categorizing memberships. A key advantage of using the DSM in this task, is that all tokens (individual words, hashtags, mentions) are in the same "word space" and their similarity can be compared—however, this process requires a practitioner to perform several searches: first to identify which nearest neighboring terms are used to refer to a social actor ("salmond", "@alexsalmond", "#alexsalmond")



(a) Selected "Yes" community word neighborhoods.



(b) Selected "No" community word neighborhoods.

Figure 4: Temporal changes in within-community AJ similarity

and then retrieve some sample tweets for context.

In terms of *argumentation & framing*, there is evidence for *content injection* [10] in the time windows with highest similarity between the two groups. Our method suggests that this strategy is effectively reproduced in tweet text with hashtags, evidenced by the appearance of hashtags from the opposition in the nearest neighboring term lists.

Twitter users would temporarily adopt hashtags popular with ideologically opposing groups, in order to spread and reinforce their political views. Below are examples of content injection from 'No' supporters, using `#yesscot`, and 'Yes' supporters using `#bettertogether`:

```
independence would bring a new wave of austerity for
families in scotland #indyref #yesscot #nothanks
```

```
why voting 'No' is a huge mistake #bettertogether #yess-
cotland #indyref [link]
```

The debate around North Sea oil revenues featured frequently in Twitter discussions on both sides. In Figure 3, the "oil" row in Week 6 has a high AJ similarity. Both groups had "bonanza" in the word neighborhoods, listed in Table 5. This revealed an interesting case, where 'Yes' voters were sharing an old article from 2013[6], while 'No' voters were quoting a correction to another news article from Prof. Alex Kemp, director of Aberdeen Center for Research in Energy Economics:

```
There will be 'No' oil bonanza - respected researcher
Prof Alex Kemp P&J 12/9/14 #indyref
```

Manually examining concordance lines for "oil", such as those illustrated in Table 1, would entail reading through thousands of entries, whereas the proposed method offers an immediately useful starting point for further exploration into how different groups appeal to authority in order to disseminate their ideas and exert power over one another.

5.3 Variation Within Communities Over Time

Looking at each community individually, we can begin to formulate an explanation for why and how these variations emerged in terms of social, political, internal and external influences. Figure 4 show temporal changes between word neighborhoods for the two respective communities. These are derived by calculating the AJ similarity between neighbourhoods for each week with those from the previous week, for the same words and within the same community. Bars above 0.00 indicate increasing agreement within a community, while bars below 0.00 show a decrease in AJ scores between two consecutive weeks, indicating larger difference between word neighborhoods.

The impact of the day of the vote, and the winning announcement can be seen within the communities between weeks 5 and 6 on x-axis in Figure 4 *a* and *b*. Naturally, there is an upsurge in agreement for `#bettertogether` & `#voteno` within the community as 'No' supporters celebrate the result. Zappavigna in [34] describes how Twitter users bond around a moment they perceive to be important to their cultural history.

After the vote results are announced, the changes within the 'No' community diminish, while the 'Yes' voters, form a brand new label (*nomination* strategy). The 'Yes' voters rapidly adopted `#the45` hashtag, rallying supporters around a new in-group. `#the45` refers to a rounded figure of 45% counted for the 'Yes' vote. The 'Yes' voters were much more active and engaged on social media, but this activity did not seem to translate into higher turnout for the 'Yes' campaign. In a meta-analysis of social media usage [7], while there may be a positive relationship between social media use and voter participation, whether or not this relationship is causal and transformative is questionable.

There are many other examples where interesting deviations in discourse between and within communities can serve as a guide for further, more qualitative interpretation.

6. CASE STUDY: 2014 US MIDTERM ELECTIONS

Midterm elections in the United States are held near the midpoint of the four year presidential elections. In 2014, elections were held on November 4th, involving seats being contested for the House of Representatives and Senate, along with governorships and a variety of local positions. Several key topics dominated the elections: immigration, national debt, jobs and minimum wage, and fears of an Ebola outbreak in the US.

[6] `https://twitter.com/BizforScotland/status/510166055659270144`

6.1 Midterm Elections Communities

Several official and unofficial sources listing Twitter accounts of incumbent and challenger campaigners were merged and segmented into Republican and Democrat groups. Third parties were not included in this case study. Official sources included verified government accounts listed by the `@gov` Twitter account, and accounts linked from the House of Representatives[7] and Congress member pages[8]. Twitter accounts advertised on these pages were included in the set. The majority of these accounts were verified by Twitter, and were either official campaign accounts of representatives run by staff, or their personal accounts which in many cases were also run by staff for the duration of the campaign.

While Midterm elections do not involve the same level of activity as presidential elections, a qualitative analysis of Twitter feeds and interviews with campaign staff from the 2012 Presidential Election by Kreiss [19] offers an insight into the use of Twitter by campaign staff to frame an agenda and engage with supporters.

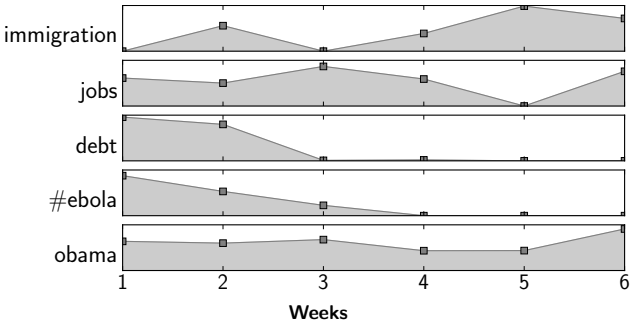


Figure 5: Trends illustrating AJ similarity changes between word neighborhoods for Democrat and Republican communities. High similarity indicates agreement between communities, while low similarity indicates greater difference in how a word is used.

For initial exploration, words associated with issues outlined by The Brookings Institution[9] were used. The issue of immigration revealed an important difference between Republican and Democrat candidates. Examining the nearest neighbors of "immigration", in word spaces built on Republican accounts, there were many more hashtags such as `#noamnesty` and `#amnesty` in contrast to Democrat word spaces, where "immigration" was associated with "reform" and "senate".

6.2 Discursive Strategies

The *framing* strategy was by far the most prominent in this case. Official campaign accounts rarely expressed or argued a stance on an issue, but did reference content elsewhere - manifestos on websites, interviews, etc. The majority of contentious conversation happened away from official accounts, among supporters and journalists. Finding this type of political discourse supports findings in [34], where politicians mainly use Twitter to foster engagement with their supporters, offering positive evaluations of themselves with a promotional style. In this study, we did not consider these other texts, restricting the corpus to tweets alone.

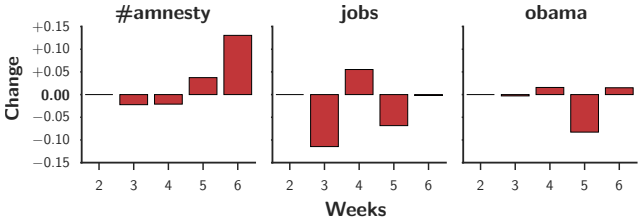


Figure 6: Temporal changes for selected Republican community word neighborhoods.

Week 3	*Week 4*	*Week 5*	*Week 6*
create	create	#energy	#jobs
textile	sector	independence	american
#madeinusa	private	lowers	create
kills	created	fewer	#energy
manufacturing	180	kill	project
creating	manufacturing	create	approve
remark	kill	gas	creating
amortization	scientific	lowering	#yes2energy
1k	generated	#jobs	supports
...	...	...	...

Table 7: Sample top words from word neighborhoods for "jobs" in the Republican community, over 4 weekly time windows.

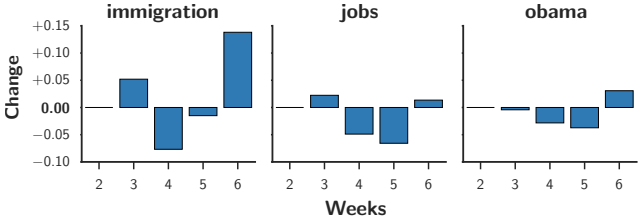


Figure 7: Temporal changes for selected Democrat community word neighborhoods.

Week 3	*Week 4*	*Week 5*	*Week 6*
creating	outsourced	000	creating
add	overseas	added	celebrate
manufacturing	paying	214	lets
overseas	created	economy	create
adding	#dayonenc	manufacturing	entrepreneurs
create	1943	adds	would
ship	300k	breaking	thousands
ca52	shipped	news	pass
bringing	rebuilding	214k	#kxl
...	...	...	...

Table 8: Sample top words from word neighborhoods for "jobs" in the Democrat community, over 4 weekly time windows.

As an alternative to selecting words of interest, a full search through the word spaces using the AJ similarity measure can rank the most similar word neighborhoods discussed by Republicans and Democrats during the election period. This has the advantage of discovering surprising instances of *mitigation* or intensifying utterances reproduced in text. The top ranked terms before and during elections included "#ebola", "jobs", and "halloween". Just after the elections, the most similar words included "birthday" associated with the 239th Birthday of the Marine Corps. Neither

[7] `http://house.gov/representatives`
[8] `https://www.congress.gov/members`
[9] `http://www.brookings.edu/research/flash-topics/flash-topic-folder/2014-midterm-elections#state`

community attempted to steer the conversation into some of the more contentious topics relating to veteran care or troops overseas around the time of the Marine Corps birthday celebrations. The *framing* strategy used by both sides amounted to sharing the same video messages and congratulations.

6.3 Variation Within Communities Over Time

Plots of the within-community temporal changes for word neighborhoods (see Figures 6 and 7) show a large spike in similarity within the Republican community for "amnesty", and Democrat community for "immigration". This may be largely due to Obama's immigration reform speech that aired on November 20th. Both communities, individually, expressed support for their official party line, as evidenced by a high similarity over time within each community. Overall, the majority of tweets produced by campaigns rarely discussed political issues on Twitter, instead the platform was utilized for general announcements and advertising positive feedback from supporters. Both Democrats and Republicans on Twitter steered away from contentious topics, opting to share generic calls to action or announcement updates about their campaigning activities. Some examples are highlighted below:

```
wow based on the turnout tonight voters are fired up
```

```
looking forward to discussing my work in congress on
wnri tune in at 8 am [link]
```

```
this isnt just an election we can win its an election we
must win
```

Discussions around jobs and employment featured frequently in both Republican and Democrat campaign accounts (See Tables 8 and 7). Both parties brought out announcements that thousands of new jobs need to be created, and frequently cited legislation on which they either voted, or will vote if re-elected.

Republicans tended to promote energy sector growth, while democrats tended towards "entrepreneurs" in the context of creating jobs. Both are similar in using words like "approve" and "pass" referring to their party proposed legislation targeting job creation.

In general, for this case study we observed that candidates did not directly engage in debates with one another, and typically kept expressions of their political positions to a minimum. This curiosity is perhaps explained by an overly cautious approach to Twitter as a medium of communication for politicians.

7. CONCLUSION

In this paper, we have proposed an approach to Critical Discourse Analysis that uses distributional semantic models to explore variations between online communities, and over time, at a word level. The approach is applicable to large social media data sets, where frequency-based approaches fail to adequately capture discourse variations between communities and over time. We evaluated our approach using two quite different political case studies, each with distinct communities active on Twitter. These case studies illustrate that analyzing discourse communities over short periods of time can highlight interesting dynamics both between and within communities, as they react to external influences that shape their discourse. While we have focused on cases involving two communities on Twitter, the approach naturally generalizes to scenarios involving multiple distinct communities, and longer pieces of text such as party manifestos, news articles, and blog posts.

In general, we suggest that DSMs offer CDA practitioners a useful exploratory tool that can be used in conjunction with existing qualitative and quantitative approaches. The ability of the DSM approach to produce word neighborhoods with both semantic and syntactic similarities could also be employed in downstream applications, such as estimating party or candidate positions on certain key issues. The effectiveness of using features derived from DSMs for these tasks is currently being investigated.

Potentially, multi-disciplinary methodologies can benefit from both quantitative and qualitative methods. A purely quantitative approach can be backed by a large body of social science and literary theory, while traditional qualitative approaches to discourse analysis can be guided and supported with quantitative techniques that are familiar in text mining, but remain underutilized by CDA practitioners. The DSM-supported process of initial word selection, followed by expansion, and examination of word neighborhoods can yield support to, or inform hypotheses about mechanisms for wielding power, hierarchies, vested interests and other aspects of critical discourse analysis.

Acknowledgments. This publication has emanated from research conducted with the support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

8. REFERENCES

[1] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA*, 2014.

[2] L. Anthony. Antconc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 2005.

[3] P. Baker, C. Gabrielatos, M. Khosravinik, M. Krzyżanowski, T. McEnery, and R. Wodak. A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press. *Discourse & Society*, 19(3):273–306, 2008.

[4] P. Baker, T. McEnery, and C. Gabrielatos. Using collocation analysis to reveal the construction of minority groups: The case of refugees, asylum seekers and immigrants in the uk press. In *Corpus Linguistics*, 2007.

[5] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 238–247, 2014.

[6] P. Basile, A. Caputo, and G. Semeraro. Analysing word meaning over time by exploiting temporal random indexing. In *CLIC 2014, The Italian Conference on Computational Linguistics*, pages 38–42, 2014.

[7] S. Boulianne. Social media use and participation: a meta-analysis of current research. *Information, Communication & Society*, 18(5):524–538, 2015.

[8] C. Chen, W. L. Buntine, N. Ding, L. Xie, and L. Du. Differential topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):230–242, 2015.

[9] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[10] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain*, 2011.

[11] N. Fairclough. *Critical Discourse Analysis: Papers in the Critical Study of Language*. Language in social life series. Longman, 1995.

[12] J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

[13] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1739–1744, New York, NY, USA, 2013. ACM.

[14] D. Greene, D. O'Callaghan, and P. Cunningham. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD '14, Nancy, France*, pages 498–513, 2014.

[15] K. Gulordava and M. Baroni. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP '11*, pages 67–71. Association for Computational Linguistics, 2011.

[16] F. Hill, R. Reichart, and A. Korhonen. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *ArXiv e-prints*, Aug. 2014.

[17] A. Jungherr. Twitter in politics: a comprehensive literature review. *Available at SSRN 2402443*, 2014.

[18] P. Juola. The time course of language change. *Computers and the Humanities*, 37(1):77–96, 2003.

[19] D. Kreiss. Seizing the moment: The presidential campaigns' use of twitter during the 2012 electoral cycle. *New Media & Society*, 2014.

[20] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically significant detection of linguistic change. In *Proceedings of the International Conference on World Wide Web*, WWW '15, pages 625–635, Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[21] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185, 2014.

[22] H. Lietz, C. Wagner, A. Bleier, and M. Strohmaier. When politicians talk: Assessing online conversational practices of political parties on twitter. In *Proceedings of the International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA*, 2014.

[23] W. C. Mann and S. A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text - Interdisciplinary Journal for the Study of Discourse*, 3(8):234–281, 1988.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[26] M. J. Paul, C. Zhai, and R. Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76, Stroudsburg, PA, USA, 2010.

[27] J. Porter. *Audience and rhetoric: an archaeological composition of the discourse community*. Prentice Hall studies in writing and culture. Prentice Hall, 1992.

[28] J. Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford, 1991.

[29] R. Sulo, T. Berger-Wolf, and R. Grossman. Meaningful selection of temporal resolution for dynamic networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, MLG '10, pages 127–136, New York, NY, USA, 2010.

[30] A. Thanopoulos, N. Fakotakis, and G. Kokkinakis. Comparative evaluation of collocation extraction metrics. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[31] Z. Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA*, 2014.

[32] T. A. Van Dijk. Critical discourse analysis. *The handbook of discourse analysis*, 18:352, 2003.

[33] R. Wodak. The discourse-historical approach. In *Methods of critical discourse analysis*, pages 63–94, 2001.

[34] M. Zappavigna. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Bloomsbury Academic. 2012.