LREC 2018 Workshop

Natural Language Processing meets Journalism III

PROCEEDINGS

Edited by

Octavian Popescu and Carlo Strapparava

ISBN: 978-0-306-40615-7
EAN: 4 003994 155486

12 May 2018

Proceedings of the LREC of Workshop
Natural Language Processing meets Journalism III

12 May 2018 – Miyazaki, Japan

Edited by Octavian Popescu and Carlo Strapparava

http://nlpj2018.fbk.eu

Organising Committee

- Octavian Popescu, IBM Research, US
- Carlo Strapparava, FBK-irst, IT

Programme Committee

- Joachim Bingel, University of Copenhagen, DK
- Peter Bourgonje, DFKI, DE
- Tommaso Caselli, Rijksuniversiteit Groningen, NL
- Maria Pia di Buono, University of Zagreb, HR
- Elena Erdmann, TU Dortmund, DE
- Lorenzo Gatti, FBK-irst, IT
- James Hamilton, Stanford University, US
- Daisuke Kawahara, Kyoto University, JP
- Kristian Kersting, TU Dortmund, DE
- Shervin Malmasi, Harvard University, US
- Rada Mihalcea, University of Michigan, US
- Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar
- Vivi Nastase, University of Heidelberg, DE
- Gözde Özbal, FBK-irst, IT
- Dina Pisarevskaya, Russian Academy of Sciences, RU
- Georg Rehm, DFKI, DE
- Mattia Rigotti, IBM Research, US
- Paolo Rosso, Universitat de Valencia, Spain
- Anna Rumshisky, University of Massachussetts, US
- Jan Šnajder, University of Zagreb, HR
- Xavier Tannier, Sorbonne Université, FR
- Serra Sinem Tekiroglu, FBK-irst, IT
- Paola Velardi, University of Roma “La Sapienza”, IT
- Ngoc Phuoc An Vo, IBM Research, US
- Kun Xu, IBM Research, US
- Marcos Zampieri, Saarland University, DE

Preface

In the third edition of Natural Language Processing Meets Journalism, held in conjunction with LREC 2018, the main four trends that we observed previously are also represented. We received (i) papers that analyze the language used in a well defined period by specific journals, (ii) papers that offer solutions and tools to manage large news corpora and (iii) papers that investigate the events reported in news and (iv) sentiment analysis.

In this workshop we see two papers discovering the social implications of the linguistic register used in news. By corroborating one analysis of two decades of Greek journal news and one of verbal aggressiveness on Twitter two decades later, the first paper proves the connection between the two analyses. It is only for the better that irrefutable statistical claims are produced by this type of analysis. The second paper shows that the implications of controversies in news are far richer than simply analyzing words. By focusing on a set of a few hundred words that are analyzed via the mutual information technique, the authors draw interesting conclusions, challenging some of the previous findings reported elsewhere.

There are two papers presenting tools and systems that can be used not only by journalists to navigate via smart connections among news in corpora. The first system presents a platform where the information between different types of media, video and text, is matched. The second system is designed to monitor news agencies and to automatically extract and link news that are related.

In the event extraction field, we see three papers at this workshop. The first one discusses the interesting experiment of learning the tacit knowledge regarding the weak signals that are present in scientific news. While there cannot be a formal definition of weak signals, the paper proves that there is consistent tacit knowledge among annotators on deciding which news contain weak signals. The paper also reports a supervised learning of the tacit knowledge. The second paper analyses the type of narratives in the international relations reporting in mass media. The authors also report a semi supervised experiment in learning the layers/templates of these narratives.

Finally, one paper does sentiment analysis on comments on Yahoo Message Board comments. State of the art methods are used to extract the positive and negative comments that influence the trade.

All this proves that many natural language researchers become interested in journalism and their work crystalizes around important social aspects of mass media. It is rewarding for the organizers of this workshop to see that kind of convergence in interests and research. The common tread of all these papers is that they present research that is ready to take the next step: prediction and meta comments on the news itself.

Octavian Popescu and Carlo Strapparava, May 2018

Programme

Morning

09.00 – 09.10 Introduction

09.10 – 10:10 Daisuke Kawahara, Kyoto University, Japan
Invited talk

10:10 – 10.30 Chris Leberknight, Kateryna Kaplun and Anna Feldman
A Comparison of Lexicons for Detecting Controversy

10:30 – 11:00 *Coffee break*

11.00 – 11.20 Mohammad Taghipour, Foad Aboutorabi, Vahid Zarrabi and Habibollah Asghari
An Integrated text mining Platform for Monitoring of Persian News Agencies

11.20 – 11.40 Maria Pontiki, Konstantina Papanikolaou and Haris Papageorgiou
Exploring the Predominant Targets of Xenophobia-motivated behavior: A longitudinal study for Greece

11.40 – 12.00 Marcelo Sardelich and Dimitar Kazakov
Extending the Loughran and McDonald Financial Sentiment Words List from 10-K Corporate Fillings using Social Media Texts

12:00 – 13.30 *Lunch Break*

Afternoon

13.30 – 13.50 VenuMadhav Kattagoni and Navjyoti Singh
IREvent2Story: A Novel Mediation Ontology and Narrative Generation

13:50 – 14.10 Delphine Charlet and Géraldine Damnati
Linking written News and TV Broadcast News topic segments with semantic textual similarity

14.10 – 14.30 Alina Irimia, Punguta Paul and Radu Gheorghiu
Tacit Knowledge - Weak Signal Detection

14.30 – 15:30 to be announced
Invited talk

15.30 – 15:40 Best paper announcement

15.40 – 16:00 Discussion and Closing

Table of Contents

A Comparison of Lexicons for Detecting Controversy
Chris Leberknight, Kateryna Kaplun and Anna Feldman 1

An Integrated text mining Platform for Monitoring of Persian News Agencies
Mohammad Taghipour, Foad Aboutorabi, Vahid Zarrabi and Habibollah Asghari 6

Exploring the Predominant Targets of Xenophobia-motivated behavior: A longitudinal study for Greece
Maria Pontiki, Konstantina Papanikolaou and Haris Papageorgiou 11

Extending the Loughran and McDonald Financial Sentiment Words List from 10-K Corporate Fillings using Social Media Texts
Marcelo Sardelich and Dimitar Kazakov 16

IREvent2Story: A Novel Mediation Ontology and Narrative Generation
VenuMadhav Kattagoni and Navjyoti Singh 22

Linking written News and TV Broadcast News topic segments with semantic textual similarity
Delphine Charlet and Géraldine Damnati 26

Tacit Knowledge - Weak Signal Detection
Alina Irimia, Punguta Paul and Radu Gheorghiu 31

A Comparison of Lexicons for Detecting Controversy

Kateryna Kaplun, Christopher Leberknight, Anna Feldman
Montclair State University
1 Normal Avenue, Montclair, New Jersey 07043
{kaplunk1, leberknightc, feldmana}@montclair.edu

Abstract

We collect a corpus of 1554 online news articles from 23 RSS feeds and analyze it in terms of controversy and sentiment. We use several existing sentiment lexicons and lists of controversial terms to perform a number of statistical analyses that explore how sentiment and controversy are related. We conclude that the negative sentiment and controversy are not necessarily positively correlated as has been claimed in the past. In addition, we apply an information theoretic approach and suggest that entropy might be a good predictor of controversy.

Keywords: controversy, online news, sentiment analysis

1. Introduction

In many countries around the world access to online information is strictly regulated. The news is a large part of our everyday lives. News media brings social, economic, political, and all other issues to the forefront to facilitate discussions about these topics. Some of these topics may be considered controversial in that they spark debate among those with firm opposing beliefs. It is also important to know what kind of sentiment these topics emote for people. This can help determine if an article is controversial through the positive or negative words that occur in it. By studying the sentiment and controversiality of articles, we can better understand how news is censored and how news sources and people in general use language to share and promote certain ideas. In this paper, we perform a statistical analysis of sentiment and controversiality.

1.1 Previous Work

Mejova et al. (2014), Pennacchiotti et al. (2010), Dori-Hacohen & Allan (2015) and Jung et al. (2010) use a logistic regression classifier, a support vector machine classifier, a nearest neighbor method, and a probabilistic approach, respectively, to detect controversial content. There is also work on censorship tracking that uses bag-of-words models (e.g., Crandall et al. 2007). Mejova et al. (2014) conduct an experiment in which they quantify emotion and bias in language through news sources. To establish a baseline for measuring controversy, they use a crowdfunding technique with human annotators whose task was to classify controversial articles. They develop a list of strongly controversial, somewhat controversial, and non-controversial terms. They use their new lexicon to analyze a large corpus of news articles collected from 15 US-based news sources. They compare controversial and non-controversial articles in terms of a series of bias and sentiment lexicons and discuss the differences in the strength with which annotators perceived a topic as controversial and how it was perceived in the media. Mejova et al. (2014) report that in controversial text, negative affect and biased language prevail. While the results of this experiments are definitely interesting, the researchers use a relatively small number of annotators.

The size of their new dataset is small, too. They classify 462 words in their experiments. Such a small sample size adversely impacts discrimination quality and classification accuracy. To investigate the reliability of their results, we reproduce their experiment to evaluate predictive accuracy for potential use with other datasets. However, since we could not gain access to their dataset, we could not reproduce the experiment used to classify documents using logistic regression.

Other approaches include Dori-Hacohen & Allan (2015) who use a nearest neighbor classifier to map webpages to the Wikipedia articulates related to them. The assumption is that if the Wikipedia articles are controversial, the webpage is controversial as well. Dori-Hacohen & Allan (2015)'s algorithm depends on Wikipedia controversy indicators, produced from Wikipedia specific features (Jang et al. 2015). Searching for k nearest neighbors for each document is non-trivial and therefore this could be practically inefficient (Jang et al., 2015). Another limitation is that it is necessary for the topic to be covered by a Wikipedia article (Jang et al., 2015). There are also generalization limitations with domain specific sources such as Wikipedia's edit history features and Twitter's social graph information (Jang et al., 2015).

Jang et al. (2015) extends Hacohen and Allan (2015)'s work by introducing a probabilistic method for detecting controversy based on the kNN-WC algorithm. Their approach uses binary classification and a probabilistic model that can be used for ranking (Jang et al., 2015). Their approach also uses Wikipedia, since it has domain specific features.

There has been work on controversy detection that explores sentiment. Jang et al. (2015), for example, demonstrate that utilizing sentiment for controversy detection is not useful. However, Choi, Jung, and Myaeng (2010) detect controversy using a mixture model of topic and sentiment and report good results. Pennacchiotti et al. (2010) detect controversies surrounding celebrities on Twitter. They use features such as the presence of sentiment-bearing words, swear words, and words in a list of controversial topics that come from Wikipedia. In our

work we investigate the relationship between controversial content and sentiment.

2. Experiments

Three experiments investigate the potential for using an existing annotated corpus of controversial and non-controversial terms (Mejova et al, 2014) to detect controversy in online news articles. Experiments use data comprised of Montclair Controversy Corpus (see section 3.1.1) of 1,554 online news articles collected from 23 RSS feeds with four lexical resources (Table 1). If lexicons of controversial words exist, can they be used to detect controversy in online news articles? We explore this question with a series of experiments inspired by previous research (Mejova et al, 2014) that suggests a positive correlation between negative sentiment and controversy in several news articles.

1. *Experiment I* aims to test the reliability of previously annotated controversial words (Mejova et al, 2014) for detecting controversy in unlabeled documents. This was done using existing lexical resources and 19 subjects who annotated the set of words from previous research (Mejova et al, 2014). We claim that the frequency of controversial terms in the MCC can be used to partition the data into controversial and non-controversial sets. We do not believe the lexicon can be used to detect controversy for individual documents, but believe it can be used to describe an aggregate view of the data.
2. *Experiment II* provides a descriptive analysis comparing the frequency of positive and negative words in our dataset compared to previously annotated sentiment datasets (Choi et al, 2010 and Chimmalgi 2010). The claim is that negative sentiment will be correlated to controversial documents and positive sentiment will be correlated to non-controversial documents.
3. Experiment III statistically tests the proportion of negative sentiment in controversial text will be higher than the proportion of positive sentiment in non-controversial text.
    a. **H1:** The proportion for overlapping words between the negative sentiment and controversial datasets is greater than the proportion for overlapping words between the negative sentiment and noncontroversial datasets.
    b. **H2:** The proportion for overlapping words between the positive sentiment and noncontroversial datasets is greater than the proportion for overlapping words between the positive sentiment and controversial datasets

3. Lexicons

As seen in Table 1, we use several resources for our experiments. MicroWNOp and General Inquirer are sentiment dictionaries. The Mejova lexicon is a set of words labeled with controversial/non-controversial categories. Finally, we use a set of words extracted from a list of controversial Wikipedia topics (Pennacchiotti and Popescu, 2010).

Lexicon	Type of Lexicon	# of Words
MicroWNOp	Positive	418
MicroWNOp	Negative	457
General Inquirer	Positive	1628
General Inquirer	Negative	2000
Mejova	Controversial	145
Mejova	Not Controversial	272
Wikipedia	Controversial	2133

Table 1: Lexicons with the number of words

This Wikipedia terms are deemed controversial because they appear in articles that are constantly being re-edited in a cyclic way, have edit warring issues, or article sanction problems (Wikipedia: List of Controversial Issues, 2018).

3.1.1 Montclair Controversy Corpus

The Montclair Controversy Corpus (MCC) contains 1554 news articles collected from 23 RSS feeds and has 317,361 word tokens in total after the stopwords (function words and punctuation) have been removed.

To create the MCC, we start by generating a dataset of hundreds of English-language articles through 23 RSS feeds. We then remove the stopwords from the MCC. We used crowdsourcing to label the words from the collected corpus as controversial, somewhat controversial, and not controversial. The final category was determined using a majority vote rule. We use these categories to make our new resource comparable to that of Mejova et al. (2014). We use a set of controversial terms, somewhat controversial terms, and not controversial terms that were also used in Mejova et al (2014) to test against the MCC. In testing their terms against our dataset, we used a set of lexical resources as seen in Table 1. Using different lexicons, we determine whether our dataset has sufficient terms that can be classified as controversial, somewhat controversial, and non-controversial. We also compare our dataset with Wikipedia words extracted from a list of controversial topics. (Pennacchiotti and Popescu, 2010).

Our goal is to determine how well Mejova et al. (2014)'s results generalize. To determine the sentiment of words included in the MCC, we apply two sentiment lexicons to it: MicroWNOp (Choi et al, 2010) and General Inquirer (Chimmalgi 2010).

4. Results

4.1 Experiment I: Controversy

We compare Mejova et al 2014's controversial and non-controversial words, Wikipedia 2018 (only controversial words) and our new corpus. The results are summarized in the Table 2. The normalized proportion is calculated by taking the frequency of each word found the MCC and dividing it by the total number of words in the corpus in this case, 317361. Results in Table 2 suggest the Wikipedia list has a better coverage than the Mejova list of controversial terms. Overall, the results demonstrate that the controversial terms represent only a small fraction of words in the MCC, but it does appear that the MCC is

biased more toward controversial documents compared to non-controversial documents. This supports C1, but the small fraction of controversial words represented in the MCC suggests that controversial terms are not the only indicators of controversial documents.

Dataset	Type	Normalized Proportion
Mejova	Controversial	0.06130873
Mejova	Non-Controversial	0.054017349
Wikipedia	Controversial	0.193227901

Table 2: Normalized proportion of words vs lexicon

4.2 Experiment II: Sentiment

Experiment II evaluates previous claims that sentiment can help to identify controversial documents (Mejova et. al 2014). We matched the words in the MCC against the sentiment lexicons described above. The General Inquirer was approximately three or four times larger than the MicroWNOp dictionary and the frequencies were also approximately three or four times higher unlike in the controversy lexicons.

Figure 1 suggests that positive sentiment is found more in non-controversial documents across both lexicons compared to the fraction of words that emote negative sentiment in controversial documents. This appears consistent with previous results (Mejova et al, 2014). However, since results from experiment I suggest the proportion of MCC data contains controversial words we would expect that negative sentiment would also be more frequent in the MCC data. Results in Figure 1 indicate this is not the case. The statistical analysis in experiment III statistically evaluates this result.

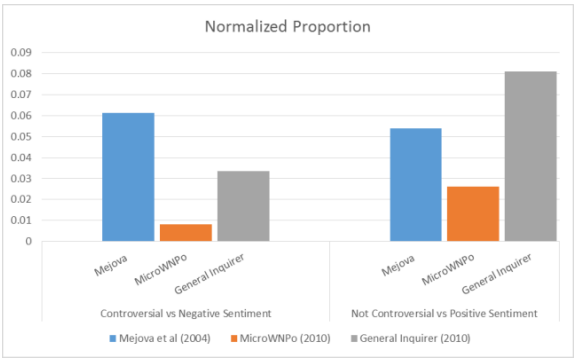

Figure 1: Normalized proportion of positive and negative sentiment

4.3 Experiment III: Controversy and Sentiment

Baseline lexicons are evaluated against each other in order to see how sentiment and controversy relate to each other. We ran four two proportion z tests to determine if words that indicate negative sentiment are more likely to appear in the lexicon of controversial terms. The following tests are summarized in Table 3.

1. Test 1 compares the controversial and non-controversial words from the Mejova lexicon in terms of the negative sentiment derived from the MicroWNOp dictionary.
2. Test 2 compares the controversial and non-controversial words in terms of the negative sentiment General Inquirer dictionary.
3. Test 3 analyzes the Wikipedia controversial words in terms of the negative sentiment obtained from the General Inquirer.
4. Test 4 compares Wikipedia list of controversial words against the Mejova non-controversial words in terms of the negative sentiment MicroWNOp dictionary.

Our alternate hypothesis (H1) is that the proportion for overlapping words between the negative sentiment and controversial datasets is greater than the proportion for overlapping words between the negative sentiment and noncontroversial datasets.

Test	z-statistic	p-value
Test 1	3.37927	0.000363
Test 2	2.00707	0.022371
Test 3	-1.45454	0.927101
Test 4	3.4985	0.000234

Table 3: Negative sentiment tests with z stats and p-values

Based on the frequencies and proportions, there is not a lot of overlap between the sentiment dictionaries and controversial lexicons.

We test the difference between the proportion of controversial words (first proportion) and non-controversial words (second proportion) in our dataset. The test evaluates if the first proportion is higher than the second proportion. For example, in Test 1, the two proportions that are being tested are 0.0414 and 0. Tests 1 and 4 are statistically significant as they have a p-value less than the 0.01 significance level. However, in these two cases the sample proportion that is being tested against is 0 because there was no overlap between the noncontroversial Mejova dataset (2014) and the negative sentiment MicroWNOp dataset (Choi et al, 2010). Therefore, the test is not particularly useful in determining if negative sentiment is more likely in a controversial dataset than a non-controversial dataset. Test 2 is significant at the 0.05 significance level but not strongly significant at the 0.01 significance level. Test 3 is not significant at all with a p-value of 0.927101. This test would actually have a better p-value if our assumption was that the non-controversial dataset had more overlap than the controversial dataset. Even still, this p-value would be 0.072899, which is still not significant.

This is an intriguing result. Unlike previous research (see e.g., Mejova (2014)) that has shown a positive correlation between controversy and negative sentiment, our statistical tests do not provide strong evidence to support this hypothesis and therefore *H1 is not supported*. In addition, we also ran four two proportion z tests to determine if words that indicate positive sentiment are more likely to appear in the noncontroversial dataset than the controversial dataset.

1. Test 5 analyzes the Mejova controversial words in terms of the positive sentiment derived from MicroWNOp dictionary.

2. Test 6 analyzes the Mejova controversial words in terms of the positive sentiment derived from the General Inquire dictionary.
3. Test 7 analyzes the Wikipedia controversial words in terms of the positive sentiment derived from the General Inquirer dictionary.
4. Test 8 analyzes the Wikipedia words in terms of the positive sentiment obtained from the MicroWNOp dictionary.

Test	z-statistic	p-value
Test 5	-1.09122	0.862412
Test 6	0.115636	0.453971
Test 7	-0.875432	0.809331
Test 8	-0.896901	0.815114

Table 4: Positive sentiment tests with z stats and p-values

Our alternate hypothesis is that the proportion for overlapping words between the positive sentiment and noncontroversial datasets is greater than the proportion for overlapping words between the positive sentiment and controversial datasets (H2). None of these four tests are significant indicating that there is no evidence that words that express positive sentiment occur more in noncontroversial data than in controversial data. Overall, results from our lexical resources suggest there is not enough conclusive evidence to determine that negative words are more likely in controversial words than noncontroversial words or that positive words are more likely in noncontroversial words than controversial words. *H2 is therefore not supported*. We hypothesize that it is the intensity of emotion rather than valence that correlates positively with controversy. We will address this issue in future work.

5. Entropy

Next, we ran an experiment in which we asked human subjects to rate all of the words in the Mejova lexicon (2014) in terms of controversy. The analysis is based on responses from 19 subjects. The subjects were presented with a single word and asked to label each individual word as controversial, somewhat controversial, or not controversial. The final category was determined based on which category had the majority of votes. Entropy is used to measure the amount of disorder or randomness in responses for each word categorized. Entropy is computed using the following formula:

$$Entropy = \sum_{i=1}^{n} \frac{(p(x_i, y_1))(\log(x_i, y_1))}{p(y_i)} \quad (1)$$

Word	Classification	Normalized Standard Deviation	Normalized Entropy
abuse	Controversial	0.00226879	0.00281437
aid	Not Controversial	0.00097444	0.00303359
america	Controversial	0.00168772	0.00289315
american	Controversial	0.00102444	0.00301502

Table 6: A small portion of the data with normalized standard deviation and normalized entropy

Table 6 presents a small portion of the data, which includes the word, its classification, its normalized standard deviation, and its normalized entropy

Entropy was normalized to account for words with a different number of responses but the same entropy value. Higher values of entropy suggest more randomness and hence more difficult to predict compared to low values of entropy. Entropy was computed for each word in the Mejova et al (2014) lexicon, the controversial, somewhat controversial, not controversial, and unknown categories, were plotted in Figure 2. The *unknown* category corresponds to words that did not have a majority vote. Figure 2 demonstrates that the entropy of words that are not controversial is lower than words that are controversial, somewhat controversial, and unknown. This indicates that entropy may be a useful feature in predicting controversial words. Additionally, of the 16 words that were classified as unknown within the survey, the classifications done by Mejova et al (2014) showed that 13 of the words were controversial and 3 of the words were somewhat controversial. Since they were controversial and somewhat controversial, they have a higher entropy just like the other controversial and somewhat controversial words. The main challenge this paper explores is how to detect controversial documents when the "ground truth" is unknown. We have shown existing controversial lexicons can be used to gain a global understanding of the degree of controversy and entropy can be used to cluster articles with the same label. For example, Figure 2 illustrates that articles with high entropy tend to fall within the controversial category. The differences between categories was shown to be statistically significant at the 0.05 significance level.

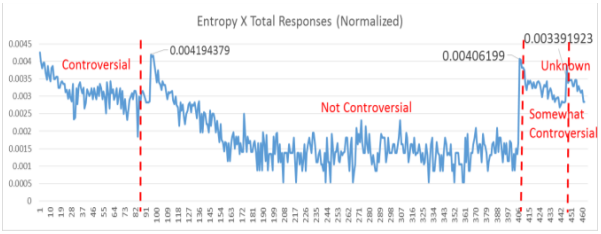

Figure 2: Normalized entropy by category

6. Limitations

There are some limitations with this study that are important to note. Our article dataset may not be fully representative of a variety of controversial topics. Most of the lexicons have a counterpart such as the Mejova noncontroversial words and the controversial words. However, there is no noncontroversial lexicon for Wikipedia, which hinders our ability to test that against the Wikipedia controversial set or compare to our sole noncontroversial dataset. Another limitation is that the negative sentiment MicroWNOp dictionary and the Mejova noncontroversial list have zero words that overlap, making it difficult to analyze these two together as well as run any tests. Also, some words are marked both as positive and negative by the sentiment dictionaries we experimented with affecting the results. This is because these words can be subject to interpretation and depending

on their context could be negative. For example, the word *help* can be positive when it is used in the sense that someone is assisting someone else with something whereas it can be seen as negative if someone is calling out for help because they are in trouble.

The final category that was assigned to each word using 19 annotators was determined with a majority vote rule. Out of 462 words, 180 had inter-rater agreement < 60%. We instead used a majority vote and results were very consistent with previous results (Mejova et al, 2014). Furthermore, entropy values need to be evaluated against "ground truth" to fully understand the benefit and reliability of this metric.

7. Future Work

Future research will investigate the potential of sentiment for controversy detection with larger news datasets and explore other methods and features for identifying controversial news. We will also build an automatic classifier to detect controversy using sentiment and entropy features. Determining the controversiality of news articles can assist future research by providing a predictor for censorship. If censorship can be predicted, a system can be designed to circumvent censorship allowing citizens to openly and freely communicate on the Internet.

8. Acknowledgements

This work has been supported by the National Science Foundation (NSF) under grant 1704113.

9. References

Choi Y., Jung Y., and S. H. Myaeng. (2010). Identifying Controversial Issues and Their Sub-topics in News Articles. Intelligence and Security Informatics, 140.

Crandall J. R., Zinn D., Byrd M., Barr E., and R. East. (2007). Concept Doppler: A Weather Tracker for Internet Censorship. *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, 352–365.

Dori-Hacohen S. and J. Allan. (2015). Automated Controversy Detection on the Web. Advances in Information Retrieval 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 – April 2, 2015. Proceedings, 423.

Garimella K., De Francisci Morales, G., Gionis A., and M. Mathioudakis. (2015). Quantifying Controversy in Social Media.

Jang M., Foley J., Dori-Hacohen S., and J. Allan. (2016). Probabilistic Approaches to Controversy Detection. Conference on Information & Knowledge Management, 2069.

Wikipedia: List of Controversial Issues. (2018).

Mejova Y., Zhang A. X., Diakopulos N., and C. Castillo. (2014). Controversy and Sentiment in Online News.

Pennacchiotti M. and A.M. Popescu. (2010). Detecting Controversies in Twitter: A First Study. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, 1873-1876.

R. V. Chimmalgi. Controversy trend detection in social media. Master's thesis, Louisiana State University, May 2010.

An Integrated Text Mining Platform for Monitoring of Persian News Agencies

Mohammad Taghipour[1], Foad Aboutorabi[2], Vahid Zarrabi[3], Habibollah Asghari[4]
[1,2,3,4] ICT Research Institute, Academic Center for Education Culture and Research
Tehran, Iran
{taghipour, foad.aboutorabi, vahid.zarrabi, habib.asghari}@ictrc.ac.ir

Abstract

There is an increasing trend in design and development of standalone tools and techniques for natural language processing (NLP) purposes. In recent years, the news agencies have focused on extracting knowledge from a huge amount of pile text news from various media. However, little work has been done to develop a unified platform for mining and monitoring the news agencies in Persian. In this paper, we present an integrated platform for monitoring the Persian news agencies. This platform consists of four main blocks including the web/social media crawler, feature extraction, impact analysis and the visualizer. Various open source tools and techniques have been employed in order to design and implement each of the mentioned segments. The final platform has been deployed in one of the most influential Iranian news agencies as a decision support system for comparing the position and rank of the news agency with respect to the other competitors.

Keywords: Persian Language, News Monitoring, Near Duplicate Detection

1. Introduction

In today's world, the news agencies play an important role in shaping the mindset of people and their perception as well as the image of the governments in international media coverage. Surviving and having a strong presence in today's competitive world requires employing of strong tools and techniques and news agencies are not exempted from this principle. Natural Language Processing (NLP) methods can greatly help to improve the impact and performance of the news agencies. A news monitoring system tries to analyze and discover hidden patterns in data and present them in the form of key performance indices (KPI's) and also to visualize them for managerial purposes.

In this paper, we focus on a Persian news monitoring system for extracting knowledge from news agencies to properly demonstrate their position in a media ecosystem in the country. Tackling the problematic characteristics of Persian as an Arabic script-based language is one of the most challenging tasks of this research. Moreover, Persian is a low-resource language with little tools and data available in digital format. So, in order to overcome these problems and to improve the quality of the platform, some custom tools and corpora were also developed and implemented.

To evaluate the power and influence of each news agency, a collection of standard measures alongside several custom features were defined and implemented. Furthermore, various methods and techniques were developed in order to demonstrate the unusual changes in the news by some reporters and agencies. A comprehensive set of visualizations were also presented based on the data and defined measures.

The paper is organized as follows: section 2 describes the related work. Section 3 comes with system design and architecture. In section 4 we describe the implementation process. Conclusion and recommendations for future works are presented in the final section.

2. Related Work

For many languages around the world, various tools and techniques are available for NLP related systems and applications, but little work have been done for monitoring of news agencies in Persian. In a research accomplished by Volkovich et al., (2016), they have proposed a novel method for analyzing Arabic media using some quantitative characteristics. Their methods try to demonstrate the ways in which important social events can be recognized by analyzing two well-known Arabic daily newspapers Al-Ahraam and Al-Akhbaar. (Volkovich et al., 2016).

In a framework proposed by (Martins et al., 2016) they mine the behavior of the crowds for temporal signals. This new time aware ranking method integrates lexical, domain and temporal evidences from multiple Web sources to rank microblog posts. Their system explores the signals from Wikipedia, news articles, and Twitter feedback to estimate the temporal relevance of search topics.

Text mining technics have been used in (Nassirtoussi et al., 2015) to predict FOREX market based on news-headlines. They proposed a novel approach to predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines. They have addressed accessing the fundamental data hidden in unstructured text of news as a challenge in a specific context by bringing together natural language processing and statistical pattern recognition as well as sentiment analysis to propose a system that predicts directional-movement of a currency-pair in the foreign exchange market based on the words used in adjacent news-headlines in the past few hours.

3. System Design and Architecture

The architecture of the media monitoring system along with the main blocks of the system is described in detail in the following subsections.

3.1 Web/Social media Crawler

Web Crawler: One of the most challenging parts of any text processing platform is the way that the data is collected from the web and stored. In this research, we have used web crawlers for gathering news data through the web. The available open source crawlers have various features and characteristics that make the benchmarking process essential for choosing the most appropriate one.

It is critical for a crawler used in a news monitoring system to be scalable and robust. Moreover, Quality, Freshness and Coverage are the other important features that should be covered by a crawler. Table 1 shows a comparison between the features of some available popular Web crawlers.

Feature / Platform	language	Operating System	License	Parallel
Scrapy	python	Linux/Mac OSX/Windows	BSD License	Yes
Apache Nutch	Java	Cross-platform	Apache License 2.0	Yes
Heritrix	Java	Linux/Unix-like/Windows Unsupported	Apache License	Yes
Web-Sphinix	Java	Windows/Mac/ Linux/Android/ IOS	Apache Software License	Yes

Table 1: Comparison of various open-source web crawlers (Yadav and Goyal, 2015)

We used Apache Nutch™ web crawler toolkit[1] for crawling the news because it satisfies the mentioned criteria. Apache Nutch™ also has some other features makes it outstanding among the other crawlers such as compatibility with Apache Hadoop™. It is also compatible with other Apache frameworks like Tika and Solr.

Using Alexa[2] site ranking system and consultation with media experts, 200 top ranked major news agencies were hand-picked for the project. It's also worth mentioning that the number of news crawled are estimated around 100,000 records per day. This data is stored in database and used for visualization.

Since some news agencies change the contents of the news after publishing, so all the crawled news are re-crawled again after 24 hours of their release time in order to find the unusual changes after the release.

A custom python-based RSS crawler was designed and developed for gathering the recent news. This RSS crawler has trained and scheduled using one of the well-known Reinforcement Learning (RL) algorithms. Q-Learning method was introduced by Watkins (1989). This model consists of states and set of actions per state, where each agency was considered as a state. The agent checks the RSS feed of the news agency while visiting the corresponding state. There are some times like at midnight, which the crawler should check the RSS feed less frequently, therefore, we considered a special state to address these cases, and when the agent visits it, the crawler does nothing. Each transition between states i and j was considered as an action between them. Q-Learning finds an optimal action-selection policy for optimum crawling. The number of news crawled in each iteration, the time interval between the iterations, and the capacity of each agency RSS feed are the features exploited for training this crawler. The aim of the method is to minimize the number of RSS feed check, while maximizing the number of crawled news in each check as well.

Social media crawler: Telegram™ is a well-known instant messaging service which is widely used in Iran. Unofficial statistics show that more than 40 million people use the services provided by this social media messaging service. Thousands of official and unofficial active channels are providing news to the masses. Most of the well-known agencies are also providing news through their own channels in Telegram. The above-mentioned points depict that this social media has a great value for monitoring and analyzing the media. It's also worth mentioning that the Telegram-API provides us some parameters such as view count for each post and participant count for each channel which are invaluable information to methods like Hot Topic Detection and so on. We developed a custom Python-based crawler to fetch the posts of about 5000 Telegram channels.

For extracting the Telegram data, two distinct custom crawlers were developed. The first one constantly checks the Telegram channel feeds and stores new posts in raw repository. The second crawler scheduled to check the recent posts constantly and update the view count of each post.

3.2 Feature extraction

HTML/JSON Parser: As the raw news data fetched by the crawlers couldn't be used directly, several custom parsers are developed to extract the features for next blocks of the project. We also developed a special parser for 14 more influential agencies to extract more detailed features, while a fast one is implemented for the others. The most important features which are extracted from the news in this step includes: the title, URL, release date and time, body of news, news category, report information and so on. The Telegram-API provides the data in neat format, so we easily extracted and parsed the required data.

Preprocessing: As the text extracted from the Web are written by different authors with various types of writing and encoding styles, a preprocessing step is required before applying any NLP task. The text extracted by the parser is fed into the Parsivar preprocessing tool (Mohtaj, et al, 2018) and then stored in a news repository. After applying a normalizer, each word in the text is marked

[1] http://nutch.apache.org/
[2] https://www.alexa.com/

corresponding to its particular Part of Speech (PoS) by its PoS tagger. Furthermore, named entities are labeled using Stanford Named Entity Recognizer (Finkel, J. R., et al., 2005). For improving the results of the NER algorithm, a customized Persian NER algorithm was also employed (Zafarian, A., et al., 2015). In the NER module, we extract the Names, Places and Time from the body of the news.

Similarity detection: Near-duplicate detection in news is the primary step to calculate the other features. We need a fast method to find similar news and to compute the news and agency characteristics, assuming that a great deal of news is released daily. Min-hash is a popular algorithm for similarity search, especially for high dimensional data in which it converts large sets into short signatures while preserving similarity (Theobald, et al 2008). We used Min-hash to extract short signatures and Jaccard similarity for computing similarity between signatures. Finally, we clustered all pairs of news which are highly similar.

Automatic news labeling: All of the news agencies are hand-labeled in some categories. Since the default categories used by each agency are different from the other agencies, so we should generate a unified class of category and label all the news according to it. We select a base category from one of the top ranked news agencies. Using the SVM method, a classifier was trained and employed for mapping the news to our standard category.

News topic detection: In order to get a deeper insight from the data, a multi-level clustering engine based on Latent Dirichlet Allocation (LDA) method was implemented on the crawled data (Blei et al. 2003). LDA is one of the most powerful methods for modeling topics of the documents which is able to properly extract the topics and provide a deep insight to the users. As the original LDA method is inherently not incremental, so we implemented another version of LDA which is capable of dealing with news streams (Hoffman, et al., 2010).

3.3 Impact analysis

Using the data and features collected from the previous steps, several quantitative characteristics or indicators are derived and calculated in this step. These characteristics are especially useful when the users want to know about the influence and penetration of the news, agencies and reporters. In order to describe the indicators, some definitions should be clarified as follow:

Media: The web based infrastructure of news platforms including news agencies, news websites, newspapers and social media.

Target News: The current news that the other indicators are calculated based on it.

Target Media: The current media that released the Target News.

Release time: The time that the Target News is released by an agency.

Release Chain: All the similar news copied and released by the other agencies are put in a time-line which is called the Release Chain.

Starter Media: A news agency or media that is the starter of a release chain

Starter News: A news that is the starter of a release chain

Table 2 illustrates some of the indicators used in the system.

Indicator	Calculation Method
Absolute delay	The time interval between the release of target news and the release of the starter news in the release chain
Relative delay	The average time interval between the release of target news and the release of all previous news in the release chain
Release rank	The rank of the current agency in the release chain (number of released news before the target news in the release chain)
News penetration	The total number of news in the release chain
Agency penetration	The total number of agencies in the release chain that released the target news after the current agency
Similarity Rate	The similarity percentage between the target news and the starter news. This parameter is also referred to as impact
Change factor	The degree to which a news changes with respect to its first release
Rate of Micro News	Micro News are the news with no significant information because of their short length, which sometimes are just headlines. As these news may be completed after the release time, they are considered as a negative indicator in media evaluation.

Table 2: News characteristics and their estimation methods

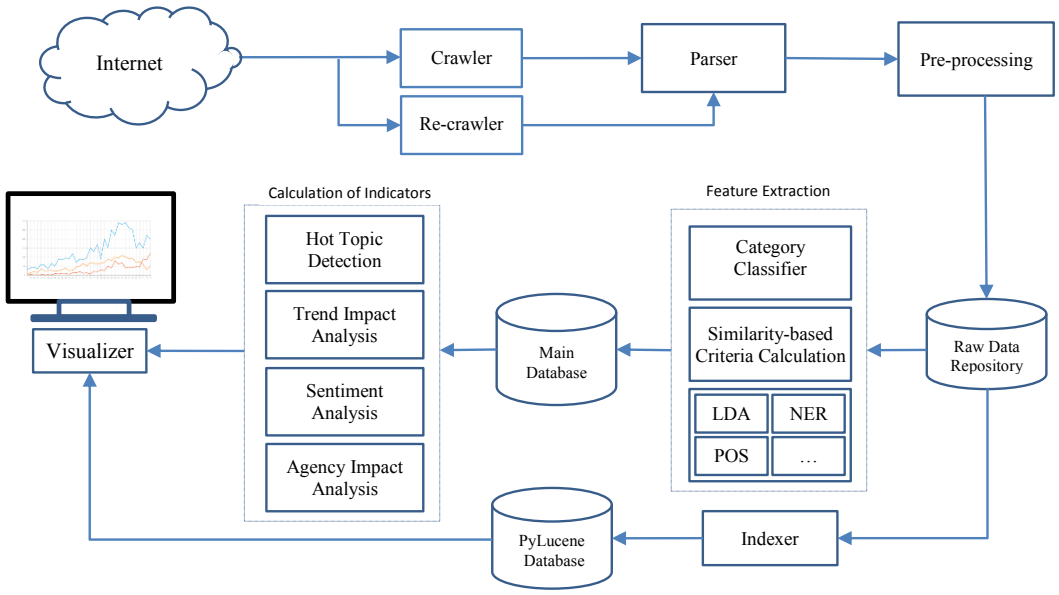

Figure 1: The complete system design architecture

3.6 Visualizer

For visualizing the indicators and parameters extracted from the previous steps, we constructed a web-based platform. Since most of the libraries used in this project were based on python APIs, so we used Django as a robust and powerful web-framework. Moreover, in the client side, several JavaScript frameworks such as JqueryUI[3], Telerik Kendo UI® Core[4], Gephi[5] and other frameworks were used. We divided the visualizations into three categories of News, Agencies, and Reporters.

In the 'News' section, the user can thoroughly search in the crawled news and observe the above-mentioned features and indicators. In the 'Agencies' section, the reports were divided into two categories of time-dependent and non-time-dependent charts. The non-time-dependent charts generally report the number of news published by each agency in a specific time frame. They also report the numbers and figures related to the modified news (which can be derived during the re-crawling phase). These charts show the number, portion, and the change rate of modified news for each agency. These are especially useful for finding news agencies that change and modify their news in an unacceptable manner. The time-dependent charts generally depict the influence and penetration rate of news released by each news agency in each day. These types of charts are also useful for finding the agencies which release the news in first-hand in comparison to the agencies which mostly republish the news of other agencies. Finally, in 'Reporters' section, it's possible to see the total number of news (and the number of news in each news category) published by each agency's reporter. This is especially useful for the evaluation of reporters with respect to their news quality and impact. Figure 1 shows the main segments and the complete data flow of the system.

4. Implementation

In order to reveal the possible barriers and to minimize the risk of the project a two phase approach based on the previous experiments has been implemented. In phase one, a pilot plan of the system was designed and implemented. To begin with, a relational database with related tables was designed. After that, as mentioned before, the major Persian news agencies were hand-picked and crawled (with focus on the recent news) and the text data was stored to the database. Other blocks of the project have also been implemented on the available data. In this phase, many shortcomings of each block were revealed and addressed accordingly. By accumulating the experiences gathered in phase 1, we encountered the problem of computational complexity. So, in the second phase, a two level database (Archive and Live) with required indexes, views and partitions were implemented in order to minimize the queries burden and maximize the speed of visualization. Other tools and methods have been also refined to enjoy the new architecture. Figure 2 illustrates a screenshot of the system.

[3] https://jqueryui.com/
[4] https://www.telerik.com/kendo-ui
[5] https://gephi.org

Figure 2: A screenshot of the system

5. Conclusion and future works

In this paper, we have presented the architecture, methods and approaches used to develop a robust platform for mining and monitoring the Persian news agencies. Despite the completeness of the platform and satisfaction of the end users, there is a long way ahead to implement state of the art techniques into the platform in order to be comparable to the systems in other languages.

As a work for the future, we plan to design an evaluation platform to measure the performance of the various parts of the system. We are also planning to design and implement some NLP related tasks such as hot topic detection, trend impact analysis, sentiment analysis and agency impact analysis.

Acknowledgements

This work has been accomplished in ICT research institute, affiliated to ACECR. We want to thank Salar Mohtaj, Atefe Zafarian, Behnam Roshanfekr, Gloshan Afzali, Sepehr Arvin, and all other team members helped us in the development of the project. We also want to express our sincere gratitude to the members of Iranian Students News Agency (ISNA) for their help to the project. Special credit goes to Dr. Heshaam Faili and Mr. Khashandish for their precious guidance to the project.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting on association for computational linguistics (pp. 363-370). Association for Computational Linguistics.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. *In advances in neural information processing systems* (pp. 856-864).

Martins, F., Magalhães, J. and Callan, J., (2016), February. Barbara made the news: mining the behavior of crowds for time-aware learning to rank. *In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 667-676). ACM.

Mohtaj, Salar, Behnam Roshanfekr, Atefeh Zafarian, Habibollah Asghari, (2018) Parsivar: A Language Processing Toolkit for Persian, *11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, 7-12 May 2018, Miyazaki (Japan)

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L., (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), pp.306-324.

Theobald, M., Siddharth, J., & Paepcke, A. (2008, July). *Spotsigs: robust and efficient near duplicate detection in large web collections.* In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 563-570). ACM.

Volkovich, Z., Granichin, O., Redkin, O. and Bernikova, O., 2016. Modeling and visualization of media in Arabic. *Journal of Informetrics*, 10(2), pp.439-453.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4), 279-292.

Yadav, M. and Goyal, N., 2015. Comparison of Open Source Crawlers-A Review. *International Journal of Scientific and Engineering Research*, 2229, 5518, pp.1544-1551.

Zafarian, A., Rokni, A., Khadivi, S., & Ghiasifard, S. (2015, March). Semi-supervised learning for named entity recognition using weakly labeled training data. *In Artificial Intelligence and Signal Processing* (AISP), 2015 International Symposium on(pp. 129-135). IEEE.

Exploring the Predominant Targets of Xenophobia-motivated Behavior: A Longitudinal Study for Greece

Maria Pontiki, Konstantina Papanikolaou, Haris Papageorgiou
Institute for Language and Speech Processing, Athena Research and Innovation Center
Artemidos 6 & Epidavrou, 15125, Athens, Greece
{mpontiki, konspap, xaris}@ilsp.gr

Abstract

We present a data-driven linguistic approach for exploring the predominant targets of xenophobia-motivated behavior in Greece over time focusing on specific Target Groups of interest. We employ two principal data analytics workflows to produce the corresponding data insights; Event Analysis using news data from 7 different sources for the last two decades (1995-2016) capturing physical attacks along with the involved social actors, and Verbal Aggressiveness using Twitter data, locating xenophobic stances as expressed by Greeks in social media for the time period 2013-2016. The results indicate that examining physical and verbal aggression as indicators of xenophobic attitudes and combining News and Twitter data can provide important insights allowing to measure, monitor and comprehend xenophobia as a violent practice in Greece over time. Hence, our work constitutes a source of valuable information for journalists, political and social scientists, policy makers and all stakeholders interested in this complex social phenomenon, and can also serve as a storytelling and narrative framework.

Keywords: Xenophobia, Event Analysis, Verbal Aggressiveness

1. Introduction

The recent refugee/immigrant crisis in Europe gave burst to xenophobic sentiments, attitudes and practices ranging from individual (re)actions to official state policies. Xenophobia is associated with feelings of dislike implying superiority, or feelings of fear/vulnerability implying the perception of threat (Van der Veer et al., 2013), and is often examined as a violent practice (Harris, 2002). Focusing on the violence aspect, we present a data-driven linguistic approach for exploring the predominant targets of xenophobia-motivated behavior in Greece over time.

In collaboration with political scientists, we defined 10 Target Groups (TGs) of interest (TG1: Pakistani, TG2: Albanians, TG3: Romanians, TG4: Syrians, TG5: Muslims, TG6: Jews, TG7: Germans, TG8: Roma, TG9: Immigrants in general, TG0: Refugees in general). The selection was based on specific criteria such as the population of specific ethnic groups in Greece (e.g. Albanians and Pakistani are the two most populated national groups living in Greece) or the existence of established prejudices and stereotypes in Greece about specific groups (e.g. Jews were selected in order to examine anti-Semitism within the Greek society given that, according to the "ADL Global 100" survey, which elaborated an index of anti-Semitism based on the strength of anti-Semitic stereotypes, Greece was the most anti-Semitic country in Europe scoring 69%).

Then, we employed two principal data analytics workflows; [**A**] **Event Analysis** (Section 4) using news data aiming to capture physical attacks against the predefined TGS of interest. In order to explore whether violent incidents are mostly directed to foreigners or not, Greeks were used as a control group. [**B**] **Sentiment Analysis** (Section 5) in Twitter data aiming to detect Verbal Aggressiveness (VA) targeting the predefined groups of foreigners. An overview of the overall methodology and the data sources is provided in Sections 2 and 3, respectively. The results, presented in Section 6, indicate that a large-scale study combining news data –generated by journalists to report physical attacks against foreigners– and Twitter data –generated by Social Media users in order to verbally attack to foreigners– can provide valuable insights allowing to analyze the complex phenomenon of xenophobia, and to confirm or debunk and disprove the existence of certain stereotypes and prejudices. Moreover, conclusions can be drawn regarding the implications and the various dimensions that are attributed to xenophobia. Hence, our work could serve as a storytelling and narrative framework for journalists interested in this complex social phenomenon. In addition, given the high correlation between verbal and physical aggression (Berkowitz, 1993), the proposed method can provide valuable insights also to political scientists and policy makers.

2. Methodology Overview

The overall methodology consists of 5 steps: [**A**] **Knowledge Representation.** Design of a coding framework covering a wide spectrum of physical and verbal attacks along with their complementary elements. [**B**] **Data Collection**. An important dataset of News and Twitter data was collected, prepared and curated (Section 3). [**C**] **Data Exploration**. Valuable insights were extracted helping to finalize the coding framework, and to create focused data collections. [**D**] **Content Analysis**. The data was modelled according to the information types that our research focuses on (Event Analysis and VA detection). [**E**] **Data Visualization**. The content analysis results, having been revised, were visualized in different ways making them explorable, comprehensible and interpretable.

3. Data Collection

3.1 News Data

A total of **3.638.137** news articles for a time span of more than 20 years, specifically 1995-2016, was collected from 7 news national-wide agencies in Greece (Table 1). All articles are in Greek and metadata (section labels, headlines, names of authors) were gathered for each along with the text itself. Data preparation included tackling normalization problems and transforming the data to a human readable corpus. During the Data Exploration

phase, event-oriented data clusters were created (one collection for each event type) to filter the collected bulk of data. To this end, complex queries were constructed comprising words and phrases in which each event type was expressed and lexicalized, also leveraging Boolean operators.

Source	Articles	Time span
Avgi	792.715	1996-2015
Kathimerini	282.621	2002-2006, 2009-2012
Eleftherotypia	429.364	2002-2006, 2008-2014
Rizospastis	725.108	1995-07/2016
TaNea	330.190	1997-2007
In.gr	428.880	1999-21/09/2016
Naftemporiki	649.259	2000-21/09/2016

Table 1: News data

3.2 Twitter Data

For each TG of interest we retrieved from Twitter relevant Tweets using related queries/keywords (e.g. "ισλάμ" (="islam"), "Πακιστανός" (="Pakistani"), etc.). Given that the search function in the database configuration is stemmed, the queries returned also tweets containing compound words and morphological variations of the selected keywords (e.g. "*ισλαμοποίηση*" for "*ισλάμ*"). The search resulted in 10 collections (1 per TG) containing in total **4.490.572** Tweets covering the time period 2013-2016 (Fig. 1). The peak in the mentions of refugees during the last two years coincides with the refugee crisis, whilst Germans are continuously in the limelight since, along with the IMF and the EU, they have a central role in the Greek crisis. The next most discussed TGs are immigrants and Syrians, who are also related to the refugee crisis. Muslims and Islam follow in the 5th place, with a peak from 2014 onward which coincides with the rise of ISIS.

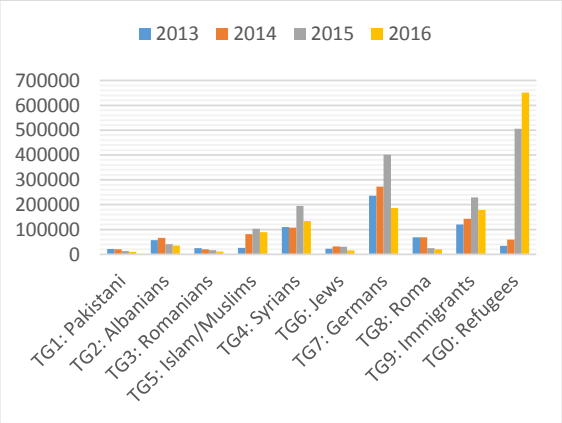

Figure 1: Per-year number of Tweets collected for each TG

4. Event Analysis

4.1 Codebook

A coding schema (q.v. Papanikolaou et al., 2016) covering a wide spectrum of event types related to xenophobia along with their structural components was set up. In this context, the coding unit of the analysis is the event. The proposed event taxonomy includes a major event category, namely Physical Attacks, encompassing various event types like *Violent Attack* and *Sexual Assault*. In the schema, an event comprises a tuple containing five types of information, each of which is also attributed several features as illustrated below:

1. **EVENT**. The word or phrase representing an event type under examination, which is located within the text. Features: *Event type*.
2. **ACTOR**. The entity that performs each event instance. Features: *Summary*, *TG*, *Nationality*, *Age*, *Sex*, *Status*.
3. **TARGET**. The entity to whom the action is addressed. Features: *Summary*, *TG*, *Nationality*, *Age*, *Sex*, *Status*.
4. The **LOCATION** where the event took place. Features: *Category*.
5. The **TIME** at which the event happened. Features: *Day*, *Month*, *Year*.
6. The **CONFIDENCE** element which captures whether in the article there is any indication that an Actor of an assault may not be the actual perpetrator. Features: *Degree*.

4.2 Content Analysis

The overall event extraction framework is data-driven. The adopted approach is to first detect each structural element of an event instance and afterwards to bind the right elements and create the event tuples. The methodology employed is semi-supervised, in the sense that a small fraction of data was labeled and used for the development of the system. Moreover, it is linguistically driven, thus morphosyntactic information from basic NLP tools is used to approach the information types comprising the event tuple as it is defined in the Codebook. The general approach for extracting events is incremental, as every module builds over the annotations produced by previous modules (Stathopoulou et al., 2017). At the first stage, the ILSP NLP tools suite (Papageorgiou et al., 2002; Prokopidis et al., 2011) is used for performing pre-processing over raw text and producing annotations for tokens, lemmas, chunks, Syntactic relations and named entities. In the next phase, the pre-processing output is given as input to the Event Analysis Unit, which performs two subsequent tasks. First, the elements comprising an Event are detected and then linguistic rules based on shallow syntactic relations bind the right elements in an event tuple. The system is implemented as Finite State Transducers (FSTs) using Gate JAPE patterns (Cunningham et al., 2000). These FSTs process annotation streams utilizing regular expressions to create generalized rules. Moreover, they are ordered in a cascade, so that the output of an FST is given as input to the next transducer. Subsequently, the output of the above described workflow, is a set of tuples, each depicting an Event with its structural elements. An illustrative example of the system output is the following:

<**Actor**: *A 24-year-old American*, **Confidence**: *is accused of involvement*, **Event**: *in the arson*, **Target**: *of the Synagogue*, **Location**: *at Chania*, **Time**: *on Thursday*>.

5. Verbal Aggressiveness (VA)

5.1 VA Framework

Based on literature review and explorative analysis findings we developed a linguistically-driven framework where VA messages (VAMs) are classified based on: **A.**

Their focus (i.e. distinguishing between utterances focusing on the target's attributes, and utterances focusing on the attacker's thoughts). **B.** The type of linguistic weapon used for the attack (e.g. formal evaluations, dirty language, humor). **C.** The content of the attack (e.g. threats/calls for physical violence or for deportation). The detailed typology is illustrated in Fig. 2.

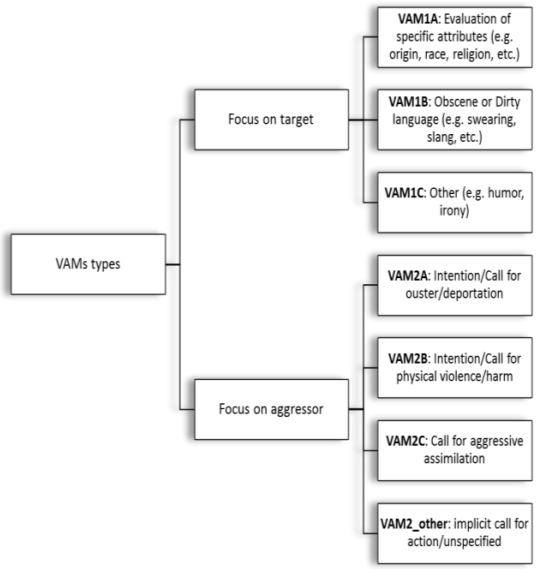

Figure 2 : Typology of VAMs

5.2 VA Analysis

We employed a rule-based method that comprises of a variety of lexical resources and grammars (sets of linguistic patterns). The VA analyzer is an FST cascade implemented as a JAPE grammar in the GATE framework. The input for the analyzer was preprocessed data. In particular, the Twitter collections described in Section 3.2 were tokenized, sentence splitted, part-of-speech tagged and lemmatized using the ILSP suite of NLP tools for the Greek language (Papageorgiou et al., 2002; Prokopidis et al., 2011). Given a preprocessed tweet, the VA analyzer detects candidate VAMs and candidate targets based on the respective lexical resources; if a token is recognized as a lexicon entry then it was annotated with the respective metadata (lexicon labels). In a subsequent step, the grammars determine which candidate VAMs and targets are correct. The grammars implement multi-phase algorithms, where the output of each phase is input for the next one. Each phase comprises several modules that contain a variety of contextual lexico-syntactic patterns. The patterns are templates that generate rules in the context around the candidate VAMs and targets. For each identified VAM, the method returns the type and the linguistic evidence of the attack as well as the id and the linguistic evidence of the object of the attack (TG). For example, for the tweet: "*Muslims should be baptized if they want to find a job in Greece"*, the analyzer returns the following tuple:

<**TG_id**: "*TG5*", **TG_evid**: "*Muslims*", **VAM_type**: "*VAM2C*", **VA_evid**: "*baptized*">.

6. Results - Predominant Targets of Xenophobia in Greece

The targets of the xenophobic attitudes are examined in the context of physical attacks reported in News data and in terms of verbal attacks expressed in Twitter. News data allow to measure and monitor physical attacks as they are reported by journalists in various newspapers for a time span of more than 20 years, and to explore possible correlations with events like the financial crisis in Greece. Due to space restrictions, we only present results for Avgi and Naftemporiki. The particular sources were chosen due to their different political orientation and because they cover the longest time span. Twitter analysis captures users' instant and freely expressed sentiments, thoughts, intentions etc. providing a snapshot of the pulse of the Greek society for the time period 2013-2016.

The quantitative analysis of the physical and verbal attacks indicates that xenophobic behaviors do not seem to be dominant in Greece, since the proportion of physical attacks against foreigners (TGs) and those against Greeks (Control Group) showed that the increase of violent incidents targeting foreigners should be examined in the light of a rise of aggressiveness in general, irrespectively of national identity of the victim. Similarly, the VA rates (VAMs/Tweets) detected in Twitter regarding the specific TGs are low (i.e. the VA rate for the mostly attacked TG is approx. 4%). However, focusing on the research goal of this paper, the identity of the victims can provide valuable insights about the xenophobic behavior of Greeks; it can help to comprehend if this type of behavior is superiority or vulnerability-based as well as if it is driven by deeply rooted stereotypes and prejudices in Greek society or by specific events (e.g. financial crisis, refugee crisis).

6.1 Main Targets of Physical Attacks

The TGs against whom most attacks occur as they were recorded in Avgi and Naftemporiki are presented in Fig. 3 and 4, respectively.

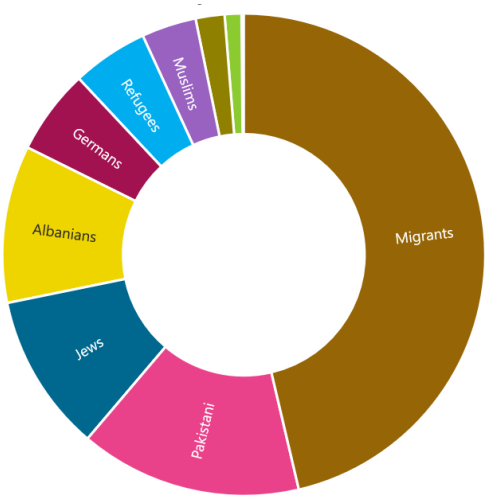

Figure 3 : Mostly Attacked TGs 2000-2015 (Avgi)

Despite some ranking differentiations the five mostly attacked TGs are the same in both sources. The general category Migrants seems to be most frequently referred to

by the newspapers, which is consistent to the regulations about media reports that have been implemented over time. Albanians and Pakistani, two of the mostly attacked TGs, are the two most populated national groups living in Greece. Finally, Jews and Germans complete the top five TGs. What needs to be noted is that there is not a significant number of people from these two ethnicities living in the Greece.

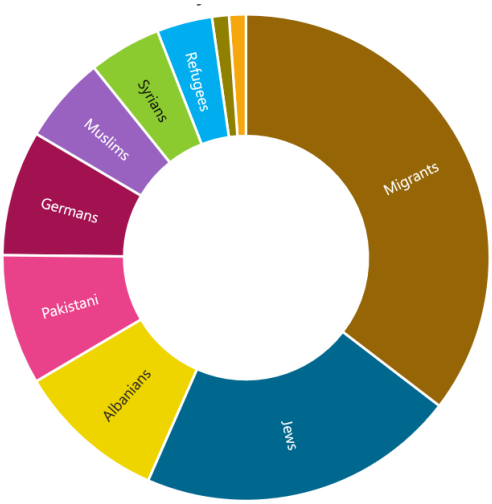

Figure 4 : Mostly Attacked TGs 2000-2016 (Naftemporiki)

We also examined how these TGs evolved over time, having 2009 as a reference point of the financial crisis' beginning as a major event that could affect the physical attacks against them. An increase of violence against foreigners during the financial crisis can be signalled, but should be related to an escalation of violent incidents in general, along with the emergence of far-right extremism in Greece. The results allow for two interesting conclusions. Firstly, there are three TGs that appear to be consistent over time, regardless of the financial crisis, viz Pakistani, Albanians and Jews. Consequently, a continuity is observed against these ethnic groups. On the other hand, new targets seem to emerge depending on the contemporary conditions affecting the country. Thus, Germans appear as TG, with the attacks against them rising as the economic crisis deepens. It is also important that there appears to be no differentiation between the two news agencies. The number of event mentions may differ, though the tendency is quite similar irrespectively of the newspaper's political orientation.

6.2 Main Targets of VA

The results of the VA analysis indicate that the most discussed/mentioned TGs in Twitter are not also the most attacked ones. In fact, refugees are the most discussed but the least attacked TG. The few verbal attacks that were captured are mostly attempts to challenge their identity implying that they are illegal immigrants. This notion of "illegality" or "lawlessness" is also dominant in the case of the generic TG Immigrants, where the most frequent terms used to attack it are the words "λαθρομεταναστες" and "λαθρο" (meaning illegal). According to the results of the analysis (Fig.5) the most attacked TGs are Jews (23%), Albanians (22%), Pakistani (15%), Muslims/Islam (14%), and Immigrants in general (10%). Antisemitism seems to be at the core of xenophobic discourse. Albanians are perhaps the most established group of foreigners in Greek public discourse, given that the image of foreigner as it was constructed in Greece during and after the first wave of migration flow (early 1990s-mid 1990s) was mainly associated with Balkan, and mainly Albanian, nationality.

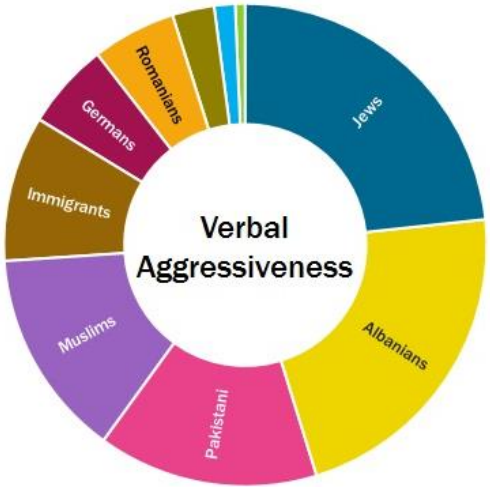

Figure 5: Mostly Attacked TGs 2013-2016 (Twitter)

Focusing on the type of the verbal aggression, attacks that involve calls for physical extinction are far greater for Jews than for any other group. Moreover, aggressive messages related to this specific TG are revealing the emergence of threat perception based on biological and cultural terms, as well as the perception of a particular enmity towards the Greek nation. Threat perception seems to prevail also for Pakistani, Albanians and Immigrants, according to the share of VAM2 attacks and in particular the calls for ouster/deportation for the specific groups.

6.3 Discussion

An important observation, concluding the analysis of the targets of xenophobic attitudes, is that, in general terms, the VA results coincide with the Event Analysis results. In other words, physical and verbal aggression as indicators of xenophobic attitudes in Greece seem to be addressed to the same targets. Four out of five TGs that are mostly attacked both verbally and physically, are the same (Jews, Albanians, Pakistani and Immigrants in general). Germans don't seem to be one of the most prominent TGs of verbal attacks, while the physical attacks against them are more often. However, the qualitative analysis of the verbal attacks against them revealed the correlation to the economic crisis, in line with the physical attacks that are mainly addressed to politicians and diplomats.

The qualitative analysis of the content of the verbal attacks expressed in Twitter confirms the existence of stereotypes and prejudices that are deeply rooted in Greek society. For example, the dominant stereotypes in the construction of the image of Albanians are associated with "crime" and "cultural inferiority" indicating a continuity of the so-called stereotype of the Balkanian criminal. Crime and inferiority stereotypes are dominant also in the verbal attacks against Muslims and Islam, but with rather different aspects. In particular, the attacks are often lexicalized through evaluative and dysphemistic terms of insult or abuse to

debase core Islamic values, principles, practices, etc. indicating irrationalism/inferiority, sexist behavior and fanaticism. The inferiority stereotype is also dominant for Pakistani; most of the verbal attacks against them are lexicalized through derogatory morphological variations of the nationality adjective. In the case of Jews, the verbal attacks entail a perception of a particular enmity towards the Greek nation and blame attribution patterns of the Greek crisis. Common themes in this group are the identification with the negative aspects of the banking system and global capitalism, as well as the frequent appeal to conspiracy theory elements. These observations coincide with surveys that establish a correlation between conspiratorial thinking and ethnocentricism, and elaborate an interpretation of Greek anti-Semitism building on aspects of national identity and by employing the concept of victimhood.

The results illuminate two different dimensions usually correlated to the conceptualization of the phenomenon of xenophobia. On the one hand, attacks against TGs like Germans and Jews, who are considered more powerful, are related to the concept of vulnerability, which implies the perception of threat. On the other hand, dominance is directed against Albanians and Pakistani who are thought of as inferior in socio-economic or cultural perspectives.

7. Conclusions

Focusing on the violence aspect of xenophobia, in this paper we presented a data-driven linguistic approach for exploring the predominant targets of xenophobia-motivated behavior towards specific TGs of interest in Greece over time. The results indicate that examining physical and verbal aggression as indicators of xenophobic attitudes and combining News and Twitter data can provide important insights regarding the nature of this type of behavior and also illuminate the possible reasons behind this complex social phenomenon.

9. Acknowledgements

We acknowledge support of this work by the project "Computational Science and Technologies: Data, Content and Interaction" (MIS 5002437) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). Part of the work reported here was made possible by using the CLARIN infrastructure**.** The authors are grateful to the political scientists Vasiliki Georgiadou, Jenny Laliouti, Anastasia Kafe, and Ioannis Galariotis for their insights in the interpretation of the results and to Dimitris Galanis for his valuable help in the data processing phase.

8. Bibliographical References

Berkowitz, L. (1993). "Aggression: Its causes, consequences, and control*"*. New York, NY: McGraw-Hill.

Cunningham H., Maynard D. and Tablan V. (2000)/. "JAPE: a Java annotation patterns engine". Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield.

Harris, B. (2002). Xenophobia: A New Pathology for a New South Africa? Psychopathology and Social Prejudice. In D. Hook and G. Eagle, *Psychopathology and Social Prejudice*, 169-184. Cape Town: University of Cape Town Press.

Papageorgiou, H., Prokopidis, P., Demiros, I., Giouli, V., Konstantinidis, A., and Piperidis S. (2002). Multi–level XML–based Corpus Annotation. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, pp. 1723–1728.

Stathopoulou, T., Papageorgiou, H., Papanikolaou, K., Kolovou, A. (2017). Exploring the dynamics of protest with automated computational tools. A Greek case study. In Computational Social Science in the Age of Big Data. Concepts, Methodologies, Tools, and Applications. German Society for Online Research.

Papanikolaou, K., Papageorgiou, H., Papasarantopoulos, N., Stathopoulou, T., Papastefanatos, G. (2016). "Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter. In International AAAI Conference on Web and Social Media. North America.

Prokopidis, P., Georgantopoulos, B., and Papageorgiou, H. (2011). A suite of NLP tools for Greek. In: Proceedings of the 10th International Conference of Greek Linguistics, Komotini, Greece, pp. 373–383.

Van der Veer, K., Ommundsen, R., Yakushko, O., Higlers, L., Woelders, S., and Hagen K.A. (2013). "Psychometrically and qualitatively validating a cross-national cumulative measure of fear-based xenophobia", Quality & Quantity 47.3 (2013): 1429-1444.

Extending the *Loughran and McDonald Financial Sentiment Words List* from 10-K Corporate Fillings using Social Media Texts

Marcelo Sardelich, Dimitar Kazakov
University of York
Heslington, York YO10 5GH, UK
{marcelo.sardelich, dimitar.kazakov}@york.ac.uk

Abstract

This article describes a novel text corpora and sentiment lexicon for financial text mining. The text corpora comprises social media messages, specifically, comments on stocks by *Yahoo Message Board* service users. The messages contains the user opinion and is labelled by the user with an overall sentiment label. This novel dataset with 74,641 messages covering 492 stocks over a period of two years is made publicly available. State-of-the-art methods are used to extract terms that convey positive and negative connotation from each message of the corpora. Then, each message is represented as a vector of these terms and sentiment classifiers are trained. The best combination of text representation weights and classifier model achieves 91.4% accuracy in the test set. We then use this sentiment classifier to build a sentiment lexicon, which contains words associated with positive and negative sentiments. We show that this lexicon is useful to extend previously proposed words lists, which were manually crafted from 10-K or 10-Q financial documents, and is able to capture the sentiment of terms from the formal and informal language of financial stock markets. Our novel financial domain text corpora and sentiment lexicon constitute valuable language resources to help advance the work on financial narrative processing.

Keywords: financial texts corpora, domain-specific sentiment lexicon, supervised learning, neural networks

1. Introduction

The sentiment classification of texts is a Natural Language Processing task that has increasingly attracted attention of the research community in recent years. Broadly speaking, we can group the sentiment classification into two approaches: On the one hand, those employing supervised (Pang et al., 2002; Pang and Lee, 2008; Glorot et al., 2011; Socher et al., 2013) or semi-supervised machine learning methods (Dasgupta and Ng, 2009; Zhou et al., 2013; Ponomareva and Thelwall, 2013), and on the other hand, those using unsupervised learning (Turney, 2002).

Lexicon-based sentiment classification is performed by retrieving information from *sentiment word lists* or *sentiment lexicon*, i.e. a database of words with positive and negative annotations. The main challenge of this approach is to compile the word list while avoiding any time-consuming human intervention. In other words, the goal would be to learn the sentiment words lists rather than compiling it manually. The techniques developed to build a sentiment word lists can be arranged in three broad categories: *Dictionary-based*, *Corpus-based* and *Emoticon-based*. The former method starts with a seed of initial words that contains at least one positive and one negative word. Then, the seed is bootstrapped, e.g. using WordNet (Miller, 1995) *synsets* as in Hu and Liu (2004);Hassan and Radev (2010);Rao and Ravichandran (2009). The *Corpus-based* technique is similar to the *Dictionary-based* one, however, it attempts to bootstrap the seed using a domain specific corpus. This method largely exploits *grammatical coherences*[1] of a given language (see, for example, the early studies in Hatzivassiloglou and McKeown (1997) and posterior advancements in Kanayama and Nasukawa (2006)). One of the main drawbacks of this method is the limited occurrence of linguistic conventions in a given corpus. Finally, the *Emoticon-based* methods are grounded on the fact that Emotion icons (*Emoticons*), such as `-)`, `:)` and `:-(` have an advantage of summarizing feelings. Therefore, they are useful to automatically assign a sentiment label to a given text. This method is employed in Go et al. (2009) and in Davies and Ghahramani (2011).

As a matter of fact, many publicly available language resources for sentiment classification, e.g. *Sentiment140* (Go et al., 2009), *Bing Liu Sentiment Lexicon* (Liu, 2012), *MPQA Sentiment Lexicon* (Wilson et al., 2005), *Harvard Dictionary* (the *General Inquirer*) (Stone et al., 1966) and *VADER* (Hutto and Gilbert, 2014), are built based on three fundamental methods discussed above, named *Dictionary-based*, *Corpus-based* or *Emoticon-based*.

Although these resources are effective for sentiment classification in the general contexts of customer reviews, they are of limited use for the financial domain corpora, such as US 10-K/10-Q corporate fillings, conference press releases or social media content related to stock markets. For instance, as stressed in Loughran and Mcdonald (2011): "Almost three-fourths of the words identified as negative by the widely used *Harvard Dictionary* are words typically not considered negative in financial contexts."

This work focuses on building a sentiment lexicon specific for texts from the financial domain. Three main contributions are made to the existing literature. First, we propose a novel sentiment lexicon for words in financial contexts. This sentiment lexicon is learnt from user posts of the *Yahoo Message Board* applying a supervised learning approach. In this regard, our work is helpful to extend

[1] Grammatical coherence can be understood as linguistic conventions on connectives such as *and*, *or* and *neither nor*. To illustrate, from the text "*The service is good and staff is friendly*" we could infer that "friendly" and "good" have the same sentiment connotation even without knowing *a priori* the sentiment of each word individually.

the manually annotated *Loughran and McDonald Financial Sentiment Dictionary* of Loughran and Mcdonald (2011). Second, the method we propose to build a sentiment lexicon from a text sentiment classifier can be utilized as a general method to similar problems, regardless the corpus domain. Third, we make the sentiment annotated dataset used to build the sentiment lexicon publicly available as an additional language resource.

2. Financial Domain Dataset

2.1. Description and Characteristics

Until *Yahoo*'s recent acquisition by Verizon, the company provided a financial message board service covering a broad range of individual stocks. When discussing a given stock, users could annotate their posts with one of the following fixed five sentiment labels: *Buy*, *Strong Buy*, *Sell*, *Strong Sell* and *Hold*.

Aiming to make use of this sentiment annotation, we collected raw HTML content from each stock message board. Then, we parsed this content extracting tags that contain relevant information. Finally, we converted the parsed HTML content into open JSON (JavaScript Object Notation) format. This step converted the unstructured message board content (HTML) into structured data (JSON).

In total, we collected 4.9GB of Python serialized JSON objects by sending web requests through 8 parallel processes during two consecutive weeks[2]. Messages published in 2014 and 2015 were collected for a list of 492 stocks[3]. Below, we show two samples from the JSON dataset for *IBM* and *Exxon Mobil* stocks (the field `message_sentiment` describes the label):

```
{'is_reply': True,
 'message_sentiment': 'Buy',
 'message_title': 'IBM profit machine
    slows; layoffs planned',
 'timestamp': 1366340436.652},

{'is_reply': False,
 'message_sentiment': 'Strong Sell',
 'message_title': "Bloomberg: Crude
    Oil Erases Advance on OPEC's
    Reduced Demand Forecast",
 'timestamp': 1421371595.252},
```

We aggregate the messages of each stock into three classes. In the **POS** (positive) class, we group all messages originally labelled as *Buy*, *Strong Buy*. The **NEG** (negative) class receives all *Sell*, *Strong Sell* messages. Finally, all residual messages are assigned to the class **NEUTRAL**.

Our strategy to collapse the messages into the coarse-grained classes **POS** and **NEG**, regardless whether it is a *Strong* message or not, is grounded on the fact that without the aggregation each class would be underrepresented and few labels would be left to discriminate, for example, between *Sell* and *Strong Sell*. The same follows to the *Hold* messages, i.e. few samples would be left to discriminate this specific class.

One interesting characteristic of the users' behaviour is their general optimism regrading the stock market. Based on the distribution of **POS**, **NEG** and **NEUTRAL** labels of Table 1, we see that our data has a strong bias towards messages with positive tone. We perceive this bias as a behavioural manifestation of the overconfidence and excessive optimism investors see in the stock markets, as described in Shiller (2000).

Label	Number of Samples	Percentage
POS	46,981	63
NEG	20,610	28
NEUTRAL	7,050	9
total	74,641	100

Table 1: Dataset sentiment labels distribution.

Finally, a closer look at some random samples reveals a certain degree of noise in the annotated labels. For example, the text: "*All the cards on the table today!*" is labelled by the user as **POS**. However, without the label most annotators would probably consider the message neutral. We presume that this "labelling mismatch" happens because some message board users tend to mistaken the message true connotation for their own judgment about the future performance of the company. That being said, potentially, the user that posted this message was betting the market would go up and not exactly the fact that, from a linguistic viewpoint, "*all cards on the table*" is an utterance with neutral sentiment.

2.2. Pre-Processing and Wrangling

Our pre-processing phase starts by filtering out all messages with the following characteristics: duplicate title, without any sentiment annotation and reply messages.

After this phase, we end up with $74,641$ messages that are dumped separately to a JSON file for each stock ticker.

Before training the model described in Sec. 3. we carry out the following additional pre-processing steps:

1. A simple lexical normalization to convert Out-Of-Vocabulary (OOV) words to its canonical form. The normalization treats the following cases: Repeated words (e.g. convert from "going up up up" to "going up"). Repeated symbols (e.g. convert from "AMAZING!!!!!" to "AMAZING!"). This task is pipelined and executed before the Part-Of-Speech (POS) tagging task.

2. Spell checking using *GNU Aspell*. The words that are still not recognized are filtered out.

3. We ignore terms that appears in less than three message titles.

[2]We relied on data parallelization techniques where each process/thread took care of one stock independently.

[3]The list of stocks was compiled based on all constituents of the Standard & Poor's 500 Index (S&P500) as in 2017. Subsequently, stocks with no messages were disregarded, hence, reducing the initial universe of 500 stocks.

3. Methodology

3.1. Document Representation

We use a sparse vector space model to represent each message of our dataset. However, we expand each message in a base of Semantic Orientation[4] (SO) occurring in it, rather than the standard "bag of words" model, according to which the message is represented by the set of words or *n-grams* occurrences (weighted or not). Our sentiment classifier is thus trained on what might be called "bag of Semantic Orientation (SO)".

The motivation to use the "bag of Semantic Orientation (SO)" representation resides in two facts:

1. The SO keywords are Part-of-Speech (POS) tag patterns that work as good indicators of explicit opinions. For instance, the tag pattern `JJ` (adjective) + `NNS` (noun) + `<any tag>` extracts "economic concerns" from the message "*Stocks tank on global economic concerns*". Since we represent each text of our corpora in this SO base, we can, at inference time, predict the sentiment of each SO keyword separately in order to build our sentiment lexicon.

2. The SO tag patterns are handy to disregard messages that do not convey any connotation and appropriate in our context of binary sentiment classification, i.e. only **POS** or **NEG** classes. In other words, the SO tag patterns constitute a simple algorithmic way to filter texts without explicit polarity out.

3.2. Semantic Orientation Tag Patterns Extraction

To build the SO base representing each text described in the subsection above, we extract the same Part-of-Speech (POS) tag patterns proposed in Turney (2002). Table 2 replicates these tag patterns, and the respective *TGrep2* (Rohde, 2001) expressions we used in our code.

The Part-of-Speech (POS) tagging is performed using the tagger proposed in Toutanova et al. (2003).

3.3. Sentiment Lexicon Learning and Compilation

Up to this point, we have not made use of any sentiment annotation of our dataset. That said, we could learn the polarity of each tag pattern using a totally unsupervised approach. One such approach is the *SO-PMI* method proposed by Turney (2002) and, for example, extensively discussed in Taboada et al. (2011). This approach uses search engines hit counts to calculate the Pointwise Mutual Information (PMI) between a given "keyword" and two fixed strong opinion words, such as "good" and "bad", which are expected to have opposite sentiment polarity. The *SO-PMI* is the difference between the two PMI measures[5]. Sticking to our example, we would expect that the PMI between the word "bad" and "economic concerns" would be higher than the PMI between the word "good" and "economic concerns", what would make the text "*Stocks tank on global economic concerns*" more biased to a negative sentiment label than the positive one. Nonetheless, as pointed out in Taboada et al. (2006), search engines are living organisms, subjected to a constant updating process, making the *SO-PMI* measure highly unstable over time. Additionally, the goal of this study is to learn a domain-specific lexicon but, typically, search engines do not segregate queries to texts from a specific domain, which is in our case the financial markets domain.

Tgrep 2 expression	POS Tag pattern
`(JJ.(NN|NNS))`	JJ + NN or NS
`(RB.(JJ!.(NN|NNS)))`	RB + JJ + not NN, not NNS
`(RBR.(JJ!.(NN|NNS)))`	RBR + JJ + not NN, not NNS
`(RBS.(JJ!.(NN|NNS)))`	RBS + JJ + not NN, not NNS
`(JJ.(JJ!.(NN|NNS)))`	JJ + JJ + not NN, not NNS
`(NN.(JJ!.(NN|NNS)))`	NN + JJ + not NN, not NNS
`(NS.(JJ!.(NN|NNS)))`	NS + JJ + not NN, not NNS
`(RB.(VB|VBD|VBN|VBG))`	RB + VB or VBD or VBN or VBG
`(RBR.(VB|VBD|VBN|VBG))`	RBR + VB or VBD or VBN or VBG
`(RBS.(VB|VBD|VBN|VBG))`	RBS + VB or VBD or VBN or VBG

Table 2: Extracted POS tag patterns using *TGrep2* expressions.

As an alternative to unsupervised learning, we leverage the sentiment annotation of our dataset and train three supervised binary sentiment classifiers: Logistic Regression, Linear Support Vector Machine and Neural Network. All classifiers are trained to predict the probability of the positive sentiment label and except for the Neural Network classifier were implemented using the Scikit-learn library (Pedregosa et al., 2011). The Neural Network uses the Keras library (Chollet, 2015) and is trained using an architecture with one hidden dense layer and one final dense layer with only one neuron.

Below, we provide a detailed explanation of all steps leading from the dataset messages to our proposed Sentiment Lexicon compilation:

1. *SO Tag Pattern Extraction*: After performing the preprocessing steps described in Sec. 2., for each message we extract all possible tag patterns (terms) described in Table 2. We assign the set of all different tag patterns extracted from our dataset as the vocabulary set V. When performing this step we ended up with a vocabulary with $1,185$ entries, which constitutes the dimension of our sparse vector space model.

[4] Semantic Orientation is a measure of subjectivity and opinion of a given text: see the early works of Osgood (1952) and a more recent review in (Taboada et al., 2011).

[5] Note that the measure will be negative if the PMI ("distance") between a given keyword and "bad" is higher than between the keyword "good". In simple terms, negative(positive) measures are associated with negative(positive) sentiments.

2. *Instance Representation*: We represent each message in the base of terms V using three different weight schemes: Term-Frequency (TF), Term Frequency-Inverse Document Frequency[6] (TF-IDF) and One-hot, where the term representation is assigned to one if the term appears in the text and zero otherwise.

3. *Hyperparameter Selection*: We randomly split 85% of the data for hyperparameters optimization (training) with the remaining 15% "left out" as test set. The hyperparameter selection is performed in the training data using 10-fold cross-validation. The cross-validation is implemented using greedy-search, which to sweep all possible hyperparameters of the search space for each of the three classifiers and selects the best model for a given metric. Table 3 describes the hyperparameters space for each classifier. In total this step outputs 9 models, i.e. 3 (classifiers) times 3 text representations per classifier.

4. *Sentiment Lexicon Compilation*: At this stage, our models can be consumed to classify the binary sentiment of any text. However, in order to compile a sentiment lexicon as a handy language resource, we perform the following tasks:

   - First, at inference time, we predict the **POS** label probability $\{p_i\}_{i=1}^{1,185}$ for all the entries of our vocabulary V using the One-hot models. Technically, this step is implemented passing though our classifiers $1,185$ vectors. Each of these vectors have zero elements for all dimensions expect for the ith dimension corresponding to the lexicon V_i which has entry one.
   - Second, we introduce a cut-off probability for the decision boundary, i.e. the cut-off probability decides if a given term should be grouped in the positive or negative word lists. Thus, all terms V_i with probability p_i greater (less) than 0.60 (0.40) are classified as positive (negative). The remaining terms are filtered out $(0.40 \leq p_i \leq 0.60)$.

Classifier	Hyperparameters
Logistic Regression	`regularization_type` = {l1, l2} , `C_regularization` = {0.10, 1, 10, 100}
Linear SVM	`C_regularization` = {0.10, 1, 10, 100}
Neural Network	`hidden_layer_n_neurons` = {64, 128}

Table 3: Hyperparameters space.

[6]The standard IDF weight scheme is employed in our work. This weight will lower the total TF weight of any term $V_i \in V$ that appears frequently in all instances of the dataset. For example, a term that appears in all instances (documents) will have a final TF-IDF weight equal to zero.

Our proposed "bag of Sentiment Orientations" representation address two main challenges. First, it filters out factual texts (neutral opinion). Second, it is our proposed solution to build a sentiment lexicon straight from a binary sentiment classifier. We make available the positive and negative word lists (sentiment lexicon) as a language resource, which can be found in the files `stocksenti-word-list-pos` and `stocksenti-word-list-neg` for the positive and negative sentiments, respectively.

4. Results

Table 4 shows the Sentiment classifiers test set accuracy for the 15% of our dataset samples that were "left out". The reported values are the accuracy for the best model selected during the cross-validation phase (training set) for each classifier.

Classifier	TF	TF-IDF	One-Hot
Logistic Regression	82.9	83.7	82.4
Linear SVM	82.8	83.0	82.7
Neural Network	91.3	90.8	91.4

Table 4: Test set accuracy for different classifiers and instance representations.

Our best classifier is the Neural Network and, for this classifier, the performance remains even when different text representations are used. For the other classifiers we can see that the TF-IDF representation performs the best. Based on these results, we selected the Neural Network model with One-hot representation to compile the Sentiment Lexicon using the approach described in the previous section. The final confusion matrix of the compiled Sentiment Lexicon can be found in Table 5.

Predicted/Actual	POS	NEG
POS	10,888	902
NEG	572	4,824

Table 5: Test set confusion matrix for the best classifier consumed to build the Sentiment Lexicon.

To evaluate our learnt sentiment lexicon, here named *StockSenti*, we perform two different strategies. To begin with, we assess how far our word lists are able to capture the sentiment of terms commonly used in the financial parlance, taking into account formal and informal language variations. In addition, we evaluate the effectiveness of our sentiment lexicon as a potential tool to extend the manually compiled *Loughran and McDonald Financial Sentiment Dictionary* of Loughran and Mcdonald (2011), which was built using exclusively corporate disclosures.
We provide many examples where our Sentiment Lexicon thrives in learning terms related positive and negative polarity for stock market texts. To exemplify, the term *next resistance* and *strong support* have high positive probability (0.82). On the contrary, *next support* is highly negative (with probability equal to 0.21). Further, the keywords

short squeeze, *short covering* and *too cheap* are positive as *shorting opportunity*, *great short*, *too high*, *high price* and *buy puts* are negative[7].

Interestingly, our dictionary is able to capture some relationships between the economic environment and stock markets. For instance, *cheap oil* is classified as positive, in agreement with the average negative correlation between inflation and stock prices. Even phrases like *bad weather* (0.2 probability) that are less obvious to grasp[8] were correctly classified.

Particularity, all the examples cited above are misclassified by all publicly available dictionaries built using general corpora (*Sentiment140* (Go et al., 2009), *Bing Liu Sentiment Lexicon* (Liu, 2012), *MPQA Sentiment Lexicon* (Wilson et al., 2005), *Harvard Dictionary* (the *General Inquirer*) (Stone et al., 1966) and *VADER* (Hutto and Gilbert, 2014)).

Regarding the effectiveness in extending the *Loughran and McDonald Financial Sentiment Dictionary* of Loughran and Mcdonald (2011) we grouped in Table 6 a few examples of words that are not present in this financial domain dictionary and, thus, are potential candidates to extend the same.

Positive	Negative
solid quarter, extremely undervalued, green day, buyback program, strong cash, outperform recommendation, solid company, major upgrade, legislative inaction	strong sell-off, insider trading, unprofitable allocation, expensive debt, litigious fraud, profitless resources

Table 6: Examples of terms part of our sentiment lexicon.

5. Conclusion

This work makes available to the research community a novel text corpora and sentiment lexicon for financial text mining. Indisputably, both language resources are valuable to the studies of corporate disclosures, e.g. corporate press releases, annual reports and so forth.

The sentiment classifier built on top of the sentiment labelled *Yahoo Message Board* service covers a broad range of stocks and is effective in classifying the sentiment of terms common to the stock markets parlance. We extensively assessed different classifiers and text representations and the best combination of text representation weights and classifier model achieves 91.4% accuracy in the test set.

Additionally, we propose a method to build a sentiment lexicon from a sentiment classifier by representing each dataset instance (message) in a base of terms with high polarity, what we named "bag of Semantic Orientation".

We assessed the potential of our learnt sentiment lexicon to be utilized to extend manually annotated sentiment lexicons (crafted using 10-K or 10-Q financial documents). Not only our sentiment lexicon is effective to extend financial sentiment dictionaries, but also it is able to capture the sentiment of terms from the formal and informal language of financial stock markets.

[7]The reader not familiar with the words "long", "short", "support", "resistance", "covering" and "put"/"call" derivatives instruments is encourage to consult introductory capital markets books to gain specific domain knowledge.

[8]A closer look at the dataset reveals that the *bad weather* phrase was extracted from oil companies. In this case, the negative hint is a consequence of damages caused by hurricane seasons.

6. Bibliographical References

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

Dasgupta, S. and Ng, V. (2009). Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification. In *ACL '09 Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709. Association for Computational Linguistics, aug.

Davies, A. and Ghahramani, Z. (2011). Language-independent Bayesian sentiment mining of Twitter. In *5th SNA-KDD Workshop 11 (SNA-KDD 11)*.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In Lise Getoor et al., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 513–520, New York, NY, USA, jun. ACM.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford University (CS224N Project Report).

Hassan, A. and Radev, D. (2010). Identifying text polarity using random walks. In *Proceeding ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403. Association for Computational Linguistics, jul.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, page 168, New York, New York, USA, aug. ACM Press.

Hutto, C. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Procceddings of 8th International AAAI Conference on Weblogs and Social Media*, pages 216–225.

Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, jul.

Liu, B. (2012). *Sentiment analysis and opinion mining*.

Loughran, T. and Mcdonald, B. (2011). When is a Liability

not a Liability? Textual Analysis , Dictionaries , and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, nov.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197–237.

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, jan.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, volume 10, pages 79–86, Morristown, NJ, USA, jul. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ponomareva, N. and Thelwall, M. (2013). Semi-supervised vs. Cross-domain Graphs for Sentiment Analysis. In Galia Angelova, et al., editors, *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'13)*, pages 571–578, Hissar, Bulgaria.

Rao, D. and Ravichandran, D. (2009). Semi-supervised polarity lexicon induction. In *EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics, mar.

Rohde, D. L. T. (2001). TGrep2 User Manual.

Shiller, R. J. (2000). Measuring Bubble Expectations and Investor Confidence. *Journal of Psychology and Financial Markets*, 1(1):49–60, mar.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., and Christopher D. Manning Andrew Y. Ng, C. P. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Taboada, M., Anthony, C., and Voll, K. (2006). Methods for Creating Semantic Orientation Databases. In *Proceeding of LREC-06, the 5th International Conference on Language Resources and Evaluation*, pages 427–432.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, jun.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, volume 1, pages 173–180, Morristown, NJ, USA, may. Association for Computational Linguistics.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 417–424, Stroudsburg, PA, USA, jul. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 347–354, Morristown, NJ, USA, oct. Association for Computational Linguistics.

Zhou, S., Chen, Q., and Wang, X. (2013). Active deep learning method for semi-supervised sentiment classification. *Neurocomputing*, 120:536–546, nov.

IREvent2Story: A Novel Mediation Ontology and Narrative Generation

VenuMadhav Kattagoni and Navjyoti Singh
Center for Exact Humanities,
International Institute of Information Technology, Hyderabad, India
venumadhav.kattagoni@gmail.com, singh.navjyoti@gmail.com

Abstract

Event detection is a key aspect of story development which is itself composed of multiple narrative layers. Most of the narratives are template-based and follow a narration theory. In this paper, we demonstrate a narrative from events detected in the international relations domain using our novel mediation ontology. We also introduce a novel method of classifying events through the mediation ontology using Beth Levin Verbs Classification, word2vec and Universal Dependencies. The selected feature space is a result of mapping language entities to ontological entities where we obtain substantially good results. Our methodology involves action classification based on the verb categorization of Beth Levin and its arguments determined by universal dependencies. The narration also presents interactions of international actors over various topics and other visualizations which would help journalists and researchers in the international relations domain.

Keywords: international relations, international politics, international news, ontology, event analysis, narrative generation, machine learning

1. Introduction

International relations are mostly framed by the pronouncements, engagements, responses, comments or force postures made by the actors. Actors (Kan, 2009) in international relations include individuals, groups (including ephemeral groups like crowds), organizations(including corporate entities, both public and private) and all generally recognized countries (including states and related territories). Classification of the events detected is important so as to analyze the group as a whole rather than each event discreetly. Thus, we propose a new Mediation Ontology for international relations. This new mediation ontology also provides a correlation between language entities and ontological entities which is used for classification of events into the proposed categories. We use its result in the narrative generation. Journalists often dig deep and push back against conventional wisdom, take time and resources from media companies - many of which are time-taking. We also intend to reduce this process of going through the history of similar articles from most of the media companies.

We present a brief background on the event ontologies and event coding in conjunction with media in Section 2. We present our new Mediation Ontology in Section 3. We then present our dataset in Section 4 and methodology and results for classification of events to identify the event type and also to describe the features used along with the machine learning techniques for classification in Section 5. In subsequent sections, narratives and visualizations demonstrated in section 6. We end the paper in section 7 with proposals on the future work sparked by this study.

2. Related Work

The last few decades have witnessed a considerable escalation in studies which are directed at event coding ontologies in the political domain. This kind of research began during the 1970s with the purpose of forecasting International Conflict under the sponsorship of the U.S. Department of Defense Advanced Research Projects Agency (DARPA) (Choucri and Robinson, 1978), (Andriole and Hopple, 1988). The kind of research that has been focused mainly on:

1. the political event data coding ontologies.
2. the generation of the political event data.
3. forecasting of international conflict.

Our focus in this paper is restricted to international relation event coding ontology i.e., Ontology for international relations events or mediation types. Such ontologies include WEIS (Goldstein, 1992), COPDAB (Azar, 1980), CAMEO (Gerner et al., 2002), IDEA (Bond et al., 2003) etc. The WEIS Ontology is made up of 22 top-level categories that encompass actions such as Request or Grant. Each of these 22 top-level categories contains single level children which are more fine-grained. For example, the code 07 is the top-level code for Reward with the sub-code 072 representing extended military assistance. The CAMEO ontology is an upgraded version of WEIS with mediation event types added to it. It is more fine-grained with 20 top-level categories that encompass actions such as Make Statement or Protest. Each of these 20 top-level categories contains finer-grained categories in a hierarchical manner. For example, the code 14 is the top-level code for Protest with the sub-code 142 representing a general demonstration or rally. Under the code 141 is code 1411 which codes demonstrate or rally for leadership change. Thus, as one moves down the hierarchy of CAMEO, it becomes more fine-grained. Based on one's need, CAMEO or any event data coding schemes can be evolved using a mix-and-match framework whereby a researcher could adopt most of his or her coding categories from a standard set, and then elaborate on a smaller number of newer categories.
Our work presented in this paper carves a similar problem by computing event types and narrative generation of the international events.

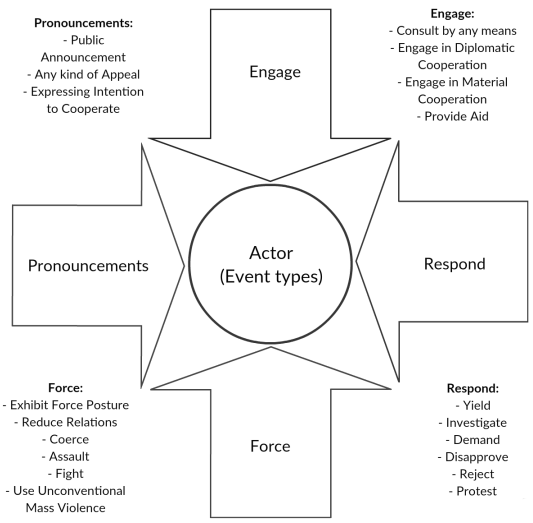

Figure 1: Mediation Ontology

Code	Class Name
01	Make Public System
02	Appeal
03	Express intent to cooperate
04	Consult
05	Engage in Diplomatic Cooperation
06	Engage in Material Cooperation
07	Provide Aid
08	Yield
09	Investigate
10	Demand
11	Disapprove
12	Reject
13	Threaten
14	Protest
15	Exhibit Force Posture
16	Reduce Relation
17	Coerce
18	Assault
19	Fight
20	Use unconventional mass violence

Table 1: CAMEO's top-level classification

3. Mediation Ontology

Bercovitch (Bercovitch, 1997) defines mediation as "a process of conflict management, related to but distinct from the parties' own negotiations, where those in conflict seek the assistance of, or accept an offer of help from, an outsider (whether an individual, an organization, a group, or a state) to change their perceptions or behavior, and do so without resorting to physical force or invoking the authority of law." He also mentions, "Mediation may well be the closest thing we have to an effective technique for dealing with conflicts in the twenty-first century". The main goal of this research is narration generation so as to help journalists and researchers identify interactions among actors during international conflicts. Since there are 250+ classes, there is overlap in their mappings from verbs to classes. Verb classification is an extremely context-sensitive exercise. Hence, we map language entities with ontological entities while proposing a new statistical model of event classification system which meets all our requirements. We classified an event type into broadly four classes rather than 20 broad classes that CAMEO (Gerner et al., 2002) consists of (with nearly 250+ sub-classes).

This idea was inferred from the concept of a person. A person could be said to make some pronouncements, have certain engagements with other persons, respond to an another person's opinions and make use of force in unhealthy relations. In a similar manner, an actor in international relations interacts with other actors through pronouncements, engagements, responses and force mechanisms. The responses and force mechanisms of an actor determine the pronouncements and engagements made by peers. This is because pronouncements and engagements happen only when some kind of base event has occurred. Hence, force mechanisms and responses are ground event types whereas pronouncements and engagements are lateral event types. Therefore, multiple actors coming together would determine international relations. Our mediation ontology is described in figure 1. We mapped the CAMEO (Gerner et al., 2002) categories as following in order to come-up with the current definitions of event types.

1. Pronouncements
   - declining to comment, making pessimistic and optimistic comment, claiming, denying, empathetic, accord, symbolic act, policy option.
   - appeal for material or diplomatic cooperation, aid, political reform, negotiation, settling disputes, accepting mediation.
   - Expressing intent to cooperate, material or diplomatic cooperation, providing aid, political reform, yield, negotiating, settle dispute, mediation.
   - CAMEO Classes - 01, 02, 03.
2. Engage
   - Consult, discuss, meet, negotiate, mediate.
   - Engaging in diplomatic, material, economic, military, judicial, intelligence cooperation, endorse, defend verbally, support, recognize, apologize, forgive, formal agreement.
   - CAMEO Classes - 04, 05, 06, 07.
3. Respond
   - Any type of response in the form of yield, investigate, demand, disapprove, reject, threaten, protest.
   - CAMEO Classes - 08, 09, 10, 11, 12, 13, 14
4. Force

Class	Training Data	Testing Data	Total
Pronouncements	38130	4237	42367
Engage	23136	2571	25707
Respond	16275	1809	18084
Force	15510	1724	17234
Total	93051	10341	103392

Table 2: Dataset Description

Class	Beth Levin Verb Classes
Pronouncements	Characterize Verbs , Appeal Verbs , Long Verbs , Verbs of Transfer of a Message , Tell , Verbs of Manner of Speaking , Say Verbs , Complain Verbs , Reflexive Verbs of Appearance
Engage	Pit Verbs , Drive Verbs , Contribute Verbs , Verbs of Future Having , Verbs of Exchange , Build Verbs , Grow Verbs , Create Verbs , Performance Verbs , Dub Verbs , Conjecture Verbs , Admire Verbs , Judgment Verbs , Correspond Verbs , Meet Verbs , Talk Verbs , Chitchat Verbs , Dine Verbs , Gorge Verbs , Verbs of Spatial Configuration , Verbs of Contiguous Location , Verbs of Inherently Directed Motion , Roll Verbs , Verbs That Are Not Vehicle Names , Accompany Verbs
Respond	Banish Verbs , Banish Verbs , Manner Subclass , Verbs of Possessional Deprivation: Cheat Verbs , Get Verbs , Hold Verbs , Verbs of Concealment , Separate Verbs , Split Verbs , Disassemble Verbs , Amuse Verbs , Verbs of Assessment , Search Verbs , Investigate Verbs , Advise Verbs , Break Verbs , Bend Verbs , Other Alternating Verbs of Change of State , Verbs of Lingering
Force	Throw Verbs , Hit Verbs , Swat Verbs , Sight Verbs , Murder Verbs

Table 3: Mapping between Mediation categories and Beth Levin Classes.

- Any type of force posture, reducing relations, coerce, assault, fight, use unconventional mass violence.
- CAMEO Classes - 15, 16, 17, 18, 19, 20

All the 20 broad CAMEO classes are described in table 1.

4. Dataset

Our system listens to 248 media feeds[1] for news articles daily. We used our previous work (Kattagoni and Singh, 2018) to extract events from the news articles dated between August 15, 2017 and September 30, 2017. Since we mapped our categorical information with CAMEO, we used the Petrarch system (Clayton Norris, 2017) based on CAMEO (Gerner et al., 2002) to generate data and map to our categories. The same data is fed to Petrarch (Clayton Norris, 2017) system. We generated a total of 103,392 events distributed across all the four class. Detailed description regarding dataset is described in table 2

5. Methodology and Results

The methodology is described in figure 2. The sentence in which the event is detected is sent to a Dependency parser to find verb and its dependencies. This passes through 3 different modules which finally unite to form our feature space.

1. In the first module, we identify the class of the verb with respect to Beth Levin Verb Classes (Levin, 1993) considering verb and its alternations. We chose 59 classes which are relevant to our classification. Refer to Beth Levin (Levin, 1993) Verb Classes in table 3

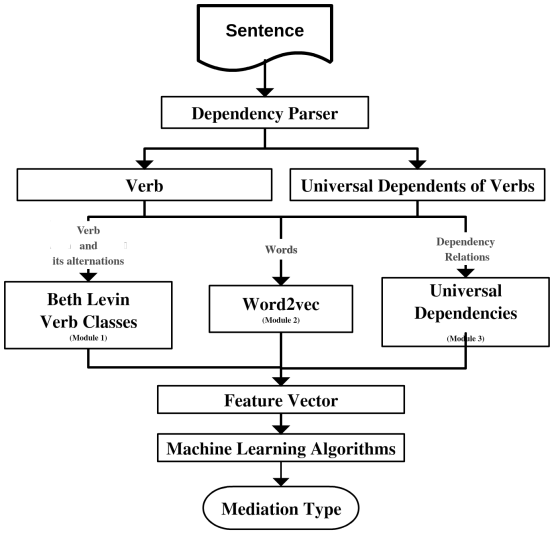

Figure 2: Methodology.

2. In the second module, the verb and its arguments which are found using the universal dependencies are converted to vectors using Google word2vec pretrained model (Tomas Mikolov and Sutskever,). All the argument vectors are added with the verb vector.
3. In the third module, all the universal dependency relations (Nivre et al., 2016) of the verb with its arguments are taken into account.

The results of a few Machine Learning algorithms on the feature space obtained from the above methodology are described in the table 4. All the metrics (precision, recall and

[1] http://ceh.iiit.ac.in/international_relations/source.txt

Method (One Vs Rest Classifier)	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.79	0.80	0.79	0.78
Random Forest	0.75	0.77	0.76	0.73
Ensemble (Logistic Regression + Random Forest)	0.78	0.80	0.79	0.77
Multi-layer Perceptron	0.80	0.80	0.80	0.80

Table 4: Results

accuracy) are the average of the corresponding class metrics. The optimum result was obtained using Multi-layer Perceptron (Hagan and Menhaj, 1994) with precision, recall and accuracy of 80%. The results are favourable using MLP because of the backpropagation training algorithm. It is worth to note that all the other Machine Learning algorithms produce nearly same results which gives a strong base for the choice of our feature space.

6. Narrative Generation

We inferred event model from our previous work (Kattagoni and Singh, 2018) with attributes date-time, location, actors, media-source, event-title, source-url, sentence. Extending this model with action (verb) and action-type (event-type), our mediation ontology adds two new attributes - action and action types. These attributes capture individual actions when events are grouped with topics helping in capturing subtler details of the generated narrative. Our system visualizes the event actor interaction using graphical, topical, geographical and temporal features. The graphical visualization represents the interaction wherein nodes are the actors and its connected entities and the edges are topics. This visualization helps place an actor level context to the conflict. The topical visualization helps situate the gravity of the topics spoken of and thus giving a subjective view of the conflict. The geographical visualization helps corner the narrative about actor's stakes in the conflict and the geopolitical persona to the event. The temporal visualization helps bring a coherency to the event-actor duo and place the interaction over a span of the dialogue until its closure. A live prototype of the system is available here: `http://ceh.iiit.ac.in/international_politics/`

7. Conclusion and Future Work

Our paper described a novel ontology for categorization of the news corpus and help in event detection in the international fora. We built a system, IREvent2Story that helps identify the various narrative features behind the events. Our ontology is a step towards framing a further attuned vocabulary for discussion of any international exchange thus setting the base for more theory work on ideating a framework for mapping not only political entities but also include non-political entities in a similar framework. Our ontology helps to not only drive meaning through the vast news data corpus but also acts as a step towards conceptualizing self-hydrating and sustaining systems of data journalism that will usher with the Web 2.0.

8. References

Andriole, S. J. and Hopple, G. W. (1988). *Defense Applications of Artificial Intelligence*. Lexington Books.

Azar, E. E. (1980). The conflict and peace data bank (copdab) project. *Journal of Conflict Resolution*, 24(1):143–152.

Bercovitch, J. (1997). Mediation in international conflict: An overview of theory, a review of practice. *Peacemaking in international conflict: Methods and techniques*, pages 125–154.

Bond, D., Bond, J., Oh, C., Jenkins, J. C., and Taylor, C. L. (2003). Integrated data for events analysis (idea): An event typology for automated events data development. *Journal of Peace Research*, 40(6):733–745.

Choucri, N. and Robinson, T. W. (1978). *Forecasting in international relations: Theory, methods, problems, prospects*. Freeman.

Clayton Norris, Philip Schrodt, J. B. (2017). Petrarch2: Another event coding program. *Journal of Open Source Software*.

Gerner, D. J., Abu-Jabr, R., Schrodt, P. A., and ÃmÃ$\frac{1}{4}$r Yilmaz. (2002). Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. In *of Foreign Policy Interactions.â Paper presented at the International Studies Association*.

Goldstein, J. S. (1992). A conflict-cooperation scale for weis events data. *Journal of Conflict Resolution*, 36(2):369–385.

Hagan, M. T. and Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE transactions on Neural Networks*, 5(6):989–993.

Kan, H., (2009). *In Government and Politics*, volume II, chapter Actors in World Politics. UNESCO-EOLSS. edited by Masashi Sekiguchi, Tokyo Metropolitan University, Japan.

Kattagoni, V. and Singh, N. (2018). Towards an unsupervised learning method to generate international political event data with spatiotemporal annotations. In *the second edition Workshop on Corpus-Based Research in the Humanities (CRH)*.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Tomas Mikolov, Quoc V. Le and Ilya Sutskever.). *word2vec*.

Linking written News and TV Broadcast News topic segments with semantic textual similarity

Delphine Charlet, Géraldine Damnati
Orange Labs, Lannion, France
firstname.lastname@orange.com

Abstract

This article explores the task of linking written and audiovisual News, based on the use of semantic textual similarity metrics. It presents a comprehensive study of different linking approaches with various configurations of inter-media or intra-media association. The influence of document length and request length is also explored. It is shown that textual similarity metrics that have proved to perform very well in the context of community question answering can provide efficient News linking metrics, whatever the media association configuration.

Keywords: textual semantic similarity, multimedia linking

1. Introduction

Linking multimedia information has become an important subject from several perspectives, ranging from helping professional journalists in order to perform news analytics to helping end-users to compare various sources of information. Linking can be seen as an input for the design of news exploration tools (e.g. (Bois et al., 2017b) or as an input for the design of efficient search engines in order to be able to retrieve related information. We propose to address the latter use-case and we compare several similarity metrics in various information retrieval configurations.

Multi-modal linking is a domain where semantic similarities are searched among multi-modal documents and/or across modalities. Video hyperlinking is a task in multimedia evaluation campains such as MediaEval (Eskevich et al., 2014) or TRECVid (see *e.g.* (Bois et al., 2017c)). The objective here is to be able to link an anchor (a piece of a BBC program which has been selected by experts as a segment of interest) to other segments defined as targets, that can be extracted from 2,700 hours of programs. This task, similarly to textual semantic similarity tasks, refers to homogeneous data: the objective is to link a fragment to another fragment from the same source. Some other works attempt to link heterogeneous sources but from an alignment perspective (*e.g.* books and movies (Zhu et al., 2015) or video lectures and scientific papers (Mougard et al., 2015)).

In the News domain there has been several studies about linking press articles with other information sources. (Aker et al., 2015) explore linking press articles and comments on the AT corpus (Das et al., 2014) which has been built from article of The Guardian. Linking press articles and Tweets have also been studied (Guo et al., 2013). (Bois et al., 2017a) attempt to build graph representations for News browsing. The authors have collected over a 3 week period a corpus of documents in French including press articles, videos and radio podcasts. Recently, the FrNewsLink corpus (Camelin et al., 2018) has been released allowing several multi-modal linking tasks to be addressed, with heterogeneous data from various sources and of various length. We propose to study the impact of several semantic similarity metrics on the task of retrieving related pieces of information on the basis of a request provided by a source piece of information.

Semantic similarity between texts have been the subject of several recent research challenges, e.g. (Cer et al., 2017) for computing similarity between sentences, or (Nakov et al., 2017) for the ranking of similar questions from a community forum, for a given question as a request. In these challenges, many sophisticated systems have been developed, mainly in a supervised way, when training data was available to learn how similar the paired texts were. Nevertheless, most of the proposed solutions were a supervised combination of unsupervised similarity metrics. Among these unsupervised similarity metrics, one appeared to perform noticeably well, in the case of Community Question Answering: the `soft-cosine` measure, which take advantage of word-embeddings based word relations, in a framework that generalizes the cosine similarity between bag-of-words (Charlet and Damnati, 2017).

In this work, we propose to explore the potential of this metric for linking heterogeneous data, namely written News and topic fragments of audiovisual News, with experiments on the FrNewsLink corpus. Section 2. presents this corpus and the various linking tasks. Section 3. presents the similarity metrics which are evaluated in section 4..

2. Corpus and linking tasks

2.1. Multimedia Corpus

We use *FrNewsLink*, a corpus which is publicly available (Camelin et al., 2018). This corpus contains automatic transcriptions of TV Broadcast News (TVBN) shows, as well as texts extracted from on-line press articles of the same period. The *FrNewsLink* corpus is based on 112 (TVBN) shows from 8 different French channels recorded during two periods in 2014 and 2015. Manual annotation for topic segmentation is provided for TVBN, thus the corpus contains mono-thematic segments of automatic transcriptions of News. For this work, we use the set of TVBN shows that have been collected during the 7^{th} week of 2014, containing 86 news shows from 8 different channels, and yielding an amount of 992 mono-thematic segments.

A set of 24,7k press articles published at the same period

has been gathered. Additionally, manual annotation is provided, that links press articles and TVBN segments. A press article is linked to a TVBN segment if they are both from the same day and if the title of the press article can be considered as an acceptable title for the TVBN segment.

2.2. Inter-media linking

Thanks to this multi-media annotated corpus, we can evaluate different inter-media linking tasks. One is to consider a speech segment as a "request" with the purpose of retrieving all the press articles of the same day that are linked to the segment. Conversely, we can consider a press article as a "request" and the task is to retrieve all the speech segments of the same day linked to the press article. Table 1 gives figures describing the corpus W07_14 (corresponding to 7^{th} week of 2014) in the inter-modal perspective.

# TVBN segments	992
# TVBN segments with at least one linked press article	707
average number of linked press article per segment with at least one linked article	11.1
# press article	5024
# press article with at least one linked TVBN news	1784
average number of linked TVBN segments per press article with at least one segment	4.4
# of inter-media linked pairs (TVBN segment with linked press article of the same day)	7830
# of potential pairs (TVBN segments $\times$ press article of the same day)	734 113
percentage of linked pairs among potential pairs	1.1%

Table 1: W07_14 statistics for inter-media linking

2.3. Intra-media linking

From the above mentioned manually annotated corpus, we can also build 2 intra-media linking tasks, through indirect supervision.

2.3.1. Linking TV Broadcast News segments

We consider that 2 TVBN segments are linked if there exists at least one common press article linked to both segments. If 2 TVBN segments are linked to press articles but without any article in common, we consider that these 2 segments are not linked. We cannot conclude about the existence or the absence of a link between 2 TVBN segments which are not linked to any press article (they could be linked, based on a topic which is not present in the press article corpus). Thus, in order to explore linking between TVBN segments, we restrict the corpus to the set of TVBN segments having at least one linked press article. Table 2 presents the statistics related to this task.

It is worth noticing that among the 707 TVBN segments which have at least one linked press article, 85% of them (604) also have a link with another TVBN segment. Only 15% of the TVBN segments linked with press articles has no other linked segments. It means that a topic from a TVBN show which is also present in written press, is very likely to be addressed in other TVBN shows during the day.

# TVBN segments (*with at least one linked press article*)	707
# TVBN segments linked with at least one TVBN segment	604
average number of linked segments per segment with at least one linked segment	11.3
# of intra-media TVBN segments linked pairs of the same day	6844
# of potential pairs of TVBN segments of the same day	76444
percentage of linked pair among potential pairs	9.0%

Table 2: W07_14 statistics for TVBN segments linking

2.3.2. Linking press articles

Conversely, we can apply the same approach to build a corpus of linked press articles. We consider that 2 press articles are linked if there exists at least one common TVBN segment linked to both articles. If 2 press articles are linked to TVBN segments but without any one in common, we consider that these 2 press articles are not linked. We cannot conclude about the existence or the absence of a link between 2 press articles which are not linked to any TVBN segments (they could be linked, based on a topic which is not addressed in TVBN shows). Thus, in order to explore linking between press articles, we restrict the corpus to the set of press articles having at least one linked TVBN segment. Table 3 shows some statistics if we consider the task of news retrieval for a given press article as a request. We can notice that among the 1784 press articles which have at least one linked TVBN segment, 97% of them (1734) have also a link with another press article. It means that once a topic from on line press is also present on TV, it is highly probable that other press articles deal with the same topic. It emphasizes the phenomenon noticed in previous section about TVBN. Thus, the existence, for a specific topic, of a cross-media link between TV and on line press implies that this topic has a high probability of being treated multiple times within each media.

# press article (*with at least one linked TVBN news*)	1784
# press articles linked with at least one press article	1734
average number of linked articles for an article with at least one linked article	20.8
# of linked pairs of articles of the same day	36126
# of potential pairs of articles of the same day	482132
percentage of linked pairs among potential pairs	7.5%

Table 3: W07_14 statistics for press articles linking

3. Similarity Metrics

3.1. Preprocessing and baseline

The texts are lemmatized and only the lemmas of adjectives, verbs and nouns are selected. Okapi TF-IDF$_{BM25}$ weights are estimated from the aggregated news corpus of the whole week. Hence, text representation consists of a weighted bag of lemmas. As a baseline similarity metrics, a `cosine` similarity is computed between vectors X and Y respectively representing texts T_X and T_Y:

$$cos(X,Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{with } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad (1)$$

3.2. `soft-cosine` similarity

When there are no words in common between texts T_X and T_Y (*i.e.* no index i for which both x_i and y_i are not equal to zero), `cosine` similarity is null. However, even with no words in common, texts can be semantically related when the words are themselves semantically related. This is why some authors (Sidorov et al., 2014) (Charlet and Damnati, 2017) have proposed to take into account word-level relations by introducing in the `cosine` similarity formula a relation matrix M, as suggested in equation 2.

$$cos_M(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \quad (2)$$

$$X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad (3)$$

where M is a matrix whose element $m_{i,j}$ expresses some relation between word i and word j. With such a metric, the similarity between two texts is non zero as soon as the texts share related words, even if they have no word in common. Introducing the relation matrix in the denominator normalization factors ensures that the reflexive similarity is 1.
Here, M reflects the similarity between word embeddings. This metric proved in SemEval2017 to be very efficient to measure similarity between questions in social data in order to address the Community Question Answering task(Charlet and Damnati, 2017). As proposed in the paper, the matrix element $m_{i,j}$ is computed as:

$$m_{i,j} = max(0, cos(v_i, v_j))^2 \quad (4)$$

where v_i and v_j are the embeddings for words i and j. They are estimated with word2vec tool (Mikolov et al., 2013) on the whole corpus of press articles of the given week.

3.3. Text Embeddings

A very simple yet efficient representation for texts consists in simply averaging the embeddings of the words of the text. It was used by many participants of the last SemEval challenge on Community Questions Answering (Nakov et al., 2017). Weighting the contribution of each word in the average embeddings appears to give significant improvement and to be competitive compared with other methods of text embeddings (Arora et al., 2017). Thus, if x_i is the weight of word i and $v_{i,k}$ is the k^{th} component of word i in the embeddings space, the k^{th} component of vector $\tilde{X}$, which represents text T_X is:

$$\tilde{X}_k = \frac{1}{\sum_i x_i} \sum_i x_i v_{i,k}$$

The similarity measure `wavg-w2v`(X,Y) between T_X and T_Y is then computed as the cosine between $\tilde{X}$ et $\tilde{Y}$.

4. Experiments

4.1. Protocol and evaluation metrics

We adopt a general protocol, whatever the type of texts to be linked (TVBN segments or press article). The task of retrieving, for a given request, the linked texts, is evaluated with Mean Average Precision (MAP). For a given request, similarities between the request and all the potential texts of the same day are computed, and the texts are ranked according to decreasing similarity. MAP@10 is used to evaluate the pertinence of the ranking of the 10 most similar texts. MAP measures how well the texts that should be linked (based on the ground-through annotations), are ranked before the non-linked texts, when the ranking is based on textual similarity metric.

As a complement to Information Retrieval evaluation, we can also consider the task of detecting linked pairs, among all potential pairs. Similarities are computed between all potential pairs of texts, and those whose similarity is above a certain threshold are considered as linked. For this set of detected pairs, precision and recall rates can be computed, as well as their harmonic mean, the F-measure. MAP and F-measure reflect different points of view: MAP translates the ability of the similarity metrics to rank the linked texts before the non-linked texts, for a given text request, without any notion of decision threshold. The F-measure additionally measures the ability to set a threshold on the similarity value in order to decide if a pair is linked or not, this threshold being common to all requests. Beyond the ability to rank, a good F-measure reflects the fact that similarity metrics between different text pairs are comparable. In the tables presented in the next section, the threshold is set *a posteriori*, so as to get the maximal F-measure.

Press articles are composed of a title (the first line of the extracted text) and a body (the rest of the text). Contrastive experiments are systematically done, considering either the title or the full article, to compute the similarity metrics. In fact, text length of the elements to be linked is expected to have a major influence on performances. Text length is given in terms of different selected lemmas, which is the size of the bag-of-words vector we keep. The average length of TVBN segments is 42.1 words. For press article, the average length of titles is 6.6 words, whereas the average length of the full articles is 120.8 words.

4.2. Results

The first set of experiments reflect the condition where the request is a TVBN segment. Table 2 presents the results for TVBN segments linking. The task is to retrieve the TVBN segments that address the same topic as the request TVBN segment. `soft-cosine` and `wavg-w2v`, which are the

request: TVBN segment		MAP@10	Fmax
target: TVBN segments	cosine	0.896	61.5
	soft-cosine	**0.932**	66.8
	wavg-w2v	**0.930**	**68.0**

Table 4: TVBN segments intra-linking

request: TVBN segment		MAP@10	Fmax
target: press article title	cosine	0.680	53.7
	soft-cosine	**0.750**	**66.1**
	wavg-w2v	0.743	65.1
target: full press article	cosine	**0.834**	73.5
	soft-cosine	0.820	**74.5**
	wavg-w2v	0.807	72.1

Table 5: linking press articles to TVBN segment

metrics which exploit word embeddings, perform equivalently (with an advantage to wavg-w2v for Fmax) and significantly better than the cosine metric. One can notice that, if the MAP@10 gives pretty good performance, the F-measure, which also involves a common decision threshold on the metrics, is not so good.

Then, table 5 presents the results for linking TVBN segments to press articles, with 2 variants: first, the press article is only represented by its title, second, the entire press article is considered. When it comes to link TVBN segments towards very short texts (titles), performances are worse than when the full article is considered. Interestingly, while the metrics which use word-embeddings perform much better than the bag-of-words cosine for linking towards titles, it is not the case for linking towards longer texts. Indeed, the bag-of-words cosine obtains the best MAP in this case. It is consistent with the fact that the advantage of word-embeddings based metrics is to measure a similarity even between texts without any word in common. The shorter the texts, the more likely it is that they do not share any word. When documents are long, it is very probable that they have common words if they are related. In this case, introducing semantic relations between words in the metrics can yield some noise by inducing too many relations and the metrics loose in their ability to rank target articles. However, when considering F-max, soft-cosine performs the best whatever the length of the target. This suggests that this metric remains better in its ability to set a threshold for detecting similar pairs, as the similarity values are more consistent.

Table 6 presents results obtained for the symmetric task: the request is now the press article, which has to be linked to TVBN segments. The variants considering only the title or the entire press article as request are evaluated. Here again, we can observe that performances obtained when using the full article are better than the ones obtained with the title only. It is also the cosine metric which is the most sensitive to increasing request length: MAP raises from 0.761 with the title to 0.917 with the full article. wavg-w2v does not benefit a lot from the increasing of text available in the request: its MAP starts at 0.887 with the title, and ends at 0.896 with the full article. soft-cosine performs best, whatever the length of the request.

request: press article title		MAP@10	Fmax
target: TVBN segments	cosine	0.761	53.7
	soft-cosine	0.**889**	**66.1**
	wavg-w2v	**0.887**	65.1
request: full press article		MAP@10	Fmax
target: TVBN segments	cosine	0.917	73.5
	soft-cosine	**0.923**	**74.5**
	wavg-w2v	0.896	72.1

Table 6: linking TVBN segments to press articles

request: press article title		MAP@10	Fmax
target: press article title	cosine	0.807	56.2
	soft-cosine	0.865	58.6
	wavg-w2v	**0.910**	**75.4**
target: full press article	cosine	0.907	67.7
	soft-cosine	**0.941**	**80.3**
	wavg-w2v	0.933	0.79.3
request: full press article		MAP@10	Fmax
target: press article title	cosine	0.918	67.7
	soft-cosine	**0.933**	**80.3**
	wavg-w2v	0.911	79.3
target: full press article	cosine	**0.940**	78.9
	soft-cosine	**0.939**	**83.7**
	wavg-w2v	0.927	82.3

Table 7: Press articles intra-linking

Finally, table 7 presents the results for intra-media press article linking, with all the possible variants, whether the request or the target are built with the title only or the full article. When it comes to link short texts together (title/title), wavg-w2v performs significantly better than the other metrics. soft-cosine, which also used word-embeddings, performs better than cosine, but not as good as wavg-w2v. When it comes to link long texts together (full/full), cosine and soft-cosine obtain the best MAP. It is worth noticing that for an equivalent MAP, cosine get a far worse F-max than the alternative metrics. For instance, in the case of (full/title), cosine and wavg-w2v get a MAP around 0.91 but a F-max respectively of 67.7 and 79.3. Likewise, in the case of (full/full), cosine and soft-cosine get a MAP around 0.94 but a F-max respectively of 78.9 and 83.7. It means that cosine provides scores which are good for ranking, but not as good as the other metrics for a global decision threshold. When it comes to link content of very different lengths (full/title or title/full), the soft-cosine performs the best.

5. Conclusion

We have shown in this article that similarity metrics that have proved to perform well for social media linking can be also very efficient for multi-media News linking. We have proposed a comprehensive study which presents the pros and cons of three different metrics in various inter-

media and intra-media linking configuration. We can draw some specific conclusions about the metrics. The baseline bag-of-words `cosine` is the most sensitive to the length of texts to be linked. It performs the worst for very short texts, and the best, or close to the best, for long texts. `wavg-w2v` is the best metric by far when it comes to linking very short texts together (title/title), but as soon as there is a mismatch in the size of texts to be linked, `soft-cosine` is better. In our experiments, the obtained MAP are pretty high, but there is a lot of room for improvement for F-max. It means that further work is necessary to make the metrics more robust to global decision threshold.

6. Bibliographical References

Aker, A., Kurtic, E., Hepple, M., Gaizauskas, R., and Di Fabbrizio, G. (2015). Comment-to-article linking in the online news domain. In *Proceedings of the SIGDIAL 2015 Conference*, pages 245–249. ACL.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*, Toulon, France, April.

Bois, R., Gravier, G., Jamet, É., Morin, E., Robert, M., and Sébillot, P. (2017a). Linking multimedia content for efficient news browsing. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, pages 301–307.

Bois, R., Gravier, G., Jamet, E., Morin, E., Sébillot, P., and Robert, M. (2017b). Language-based construction of explorable news graphs for journalists. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 31–36, Copenhagen, Denmark, September. Association for Computational Linguistics.

Bois, R., Vukotić, V., Simon, A.-R., Sicre, R., Raymond, C., Sébillot, P., and Gravier, G., (2017c). *Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity*, pages 185–197. Springer International Publishing, Cham.

Camelin, N., Damnati, G., Bouchekif, A., Landeau, A., Charlet, D., and Estève, Y. (2018). Frnewslink : a corpus linking tv broadcast news segments and press articles. In *Proceedings of LREC 2018*, Miyazaki, Japan, May.

Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 1–14.

Charlet, D. and Damnati, G. (2017). Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 315–319.

Das, M. K., Bansal, T., and Bhattacharyya, C. (2014). Going beyond corr-lda for detecting specific comments on news & blogs. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 483–492. ACM.

Eskevich, M., Aly, R., Racca, D., Ordelman, R., Chen, S., and Jones, G. J. (2014). The search and hyperlinking task at mediaeval 2014.

Guo, W., Li, H., Ji, H., and Diab, M. T. (2013). Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Mougard, H., Riou, M., de la Higuera, C., Quiniou, S., and Aubert, O. (2015). The paper or the video: Why choose? In *Proceedings of the 24th International Conference on World Wide Web*, pages 1019–1022. ACM.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August. Association for Computational Linguistics.

Sidorov, G., Gelbukh, A. F., Gómez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Tacit Knowledge - Weak Signal Detection

Alina Irimia, Paul Punguta, Radu Gheorghiu
firstname.lastname@uefiscdi.ro

Abstract

Before a certain topic becomes a very searched subject on news platform, there are some weak signals that, if correctly recognized and handled, may anticipated the popularity of that topic. One big problem with detecting such weak signals is that their recognition relies to a large extent on human tacit knowledge. Human tacit knowledge is a type of information having as main characteristics the fact that there is not a direct formal definition of it, and there is not a direct label in the text which explicitly marks it. In this paper we report on building an annotated news corpus for detection of weak signals. We also report on experiments using a supervised machine learning technique.

1. Introduction

In a diachronically ordered news corpus we can discover that a particular breakthrough event may have been predicted by corroborating small pieces of evidence existing in previously published pieces of news. That is, there are no pieces of text that directly mention or describe the breakthrough event, but there are scattered paragraphs, each one containing a faint and indirect indication to a certain possibility that further on becomes a breakthrough-event. Not being definable, this type of information cannot be identified by a precise set of rules written in a guideline for annotators. It is part of human tacit knowledge to identify the causes and consequences of certain events. In this paper we focus on weak signals, that is, on the information that a human reader is able to extract from a piece of news which is no more than a hint that a certain event is going to happen. The task we address is the classification of pieces of news into two categories: containing weak signals vs. non-containing weak signals. We have compiled a large corpus of news, made of some 40,000 scientific articles published in the last 50 years. A team of annotators were asked to annotate each piece of news as a whole according to whether the news contained or not weak signals. The annotation was carried out individually and conflicting opinions were discussed without any pressure to eventually reach a total agreement, via a process that is presented in details in Section 3. We selected a subset of roughly 20,000 documents on which the inter agreement was almost perfect (more than 99%) regarding the existence or non-existence of weak signals. We devised a set of machine learning experiments using this corpus. In section 4 we present the learning methods The fundamental result we report after these experiment is that machine learning methods can be used efficiently for tasks where the human tacit knowledge plays an important role. Few research directions which will investigate other aspects related to prediction and tacit knowledge are presented in the Conclusion and Further Research section.

2. Related Work

The literature on weak signals is not very large, as this field is about to emerge. A ground breaking paper (Brynielsson et al., 2013) was looking mainly at weak signals for detecting deviational behavior in order to efficiently provide preemptive counter measures. However, the probabilistic model presented is very close to the one used in language modeling, being an estimation of posterior probability of certain class via chain formula. In (Wang et al., 2012) an automatic detection of crime using tweets is presented. They use LDA to predict classes of similar words for topics that are related to violence. While we can gain a valuable insight from these papers, their scope is limited because there is a direct connection between the overt information existing in text and the intention of the speaker. However, in scientific prediction this relationship is much more blurred, if it exists at all. We believe a new technology must be used. The diachronicity, that is the evolution of certain topics in mass media over time, is linked to detection of weak signals. Diachronicity is also an emerging field. We found useful two statistical tests presented for epoch detection in (Popescu and Strapparava, 2014), or temporal dynamics in (Wang and McCallum, 2006; Gerrish and Blei, 2010). In (Abu-Mostafa et al., 2012) we found very useful insights from dealing with discriminative analysis and support vector machine respectively. In order to improve our results we had to be able to deal with the masking effect and to understand how we could restrict further the objective function. The work of (Popescu and Strapparava, 2013; Popescu and Strapparava, 2014) is focused on diachronic analysis of text, in particular on trends. Their work centers on finding non-random changes in distribution of topics. However, their work is not concerned with prediction on the further evolution of the analyzed topics. In (Rocktäschel et al., 2015) the principle of an attentive neural network is presented. We used these principles to implement the network presented in Section 4.4. The literature on neural networks has become rich recently and there are more than a few papers reporting on their performances on semantic tasks, such as textual entailment, semantic text similarity, short text clustering (Mueller and Thyagarajan, 2016; Palangi et al., 2016; Xu et al., 2015). However, these approaches rely on the existence of a word or sentence level annotation, and an approach based on sequence to sequence alignment is doable. In this sense, our study extends these findings, by showing that it is possible to achieve good performance for tasks where there are no direct sequences of words that are aligned.

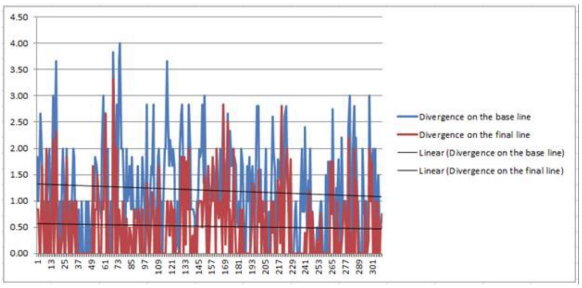

Figure 1: Towards reaching a stable shared tacit knowledge.

3. A weak Signal Corpus

Our assumption is that weak signals represent a form of tacit knowledge. As such, it may be counterproductive to define a formal set of guidelines aiming to precisely identify the weak signal. Rather, we let the annotator the liberty to mark a whole document as containing weak signals or not. In a second round of annotations we wanted to restrict the scope to paragraph rather than the whole document. Most of the annotated paragraphs contained 100 to 250 words. Therefore, we obtained two annotated corpora, which, for convenience, we refer to as short and long respectively. The long corpora, LC, refers to full documents as training/test corpora. The short corpora, SC, refers to paragraphs. There is no perfect overlap between these two corpora; approximatively 15% of paragraphs come from different documents than the ones considered on LC corpora. The annotation is binary, *yes* or *no*, signaling the existence or lack of weak signals, respectively. In case of SC all the paragraphs that were not explicitly classified as *yes*from the analyzed documents are considered *no*. However, we double checked the SC *no* for some of these paragraphs in order to make sure that there are as little as possible mis-classification. Eventually we have the following distribution in SC, LC corpora, see Table 1:

	Weak Signal	No Weak Signal
LC	4,100	14,020
SC	3,700	14,500

Table 1: Weak Signal Corpus

We wanted to have a similar ratio of weak vs. non weak in both corpora for easing a fair comparison of the performances for these two corpora. For these documents there was a large agreement regarding their category, over 99% agreement. The exact process of annotation is described below.

3.1. Annotation via tacit knowledge

We had a team of 18 undergrad volunteers. The main question was whether the periphery of tacit knowledge will become stable after practicing several hundred annotations or whether there would be a large area subject to dis-agreement between annotators. On a given set of 300 documents the annotators were encouraged to discuss their doubts and to defend their position in case of disagreement. In Fig.1 we plot the evolution of the average number of documents on which there was a strong disagreement, for samples of 10 documents out of the chosen 300. The average disagreement lowered from 1.4 to 1.1 and the divergence also decreased from .55 to .38.

It seems that 1.1 is a hard threshold for this task. When we repeated the experiment after we had 1,200 of documents annotated as carriers of weak signals, the average of disagreement for samples of ten documents, was still 1.1. However, the average time for making a decision decreased for time between these two experiments, see Fig.2. It can be considered that these results suggest that this task, in spite of being driven by tacit knowledge, is learnable by algorithmic probabilistic hypothesis space search. The annotators developed patterns, they seem to filter out a lot of the content, otherwise the time to reach a decision would not have decreased that dramatically, and there is a grey zone where experience does not help. This behaviour tends to help an automatic classifier, as it does not have to be very precise in order to obtain a human like performance. After a preliminary round of trial annotation of several hundreds of documents, we decided to create a taxonomy that sprung naturally from this experiment. This flat taxonomy has the following components:, technology, innovation in services, trend shift, behavioral change, major actor move, breakthrough discovery, top research, wild card.

The intention in using these labels was to try to capture the intuition of annotator on why a certain document/paragraph is considered as carrier of weak signals. As people usually tend to overweigh the famous research centers, famous names etc, this taxonomy helps us to see if there are indeed any subjective differences that may affect the learning process. The indication here was that wild card, which is

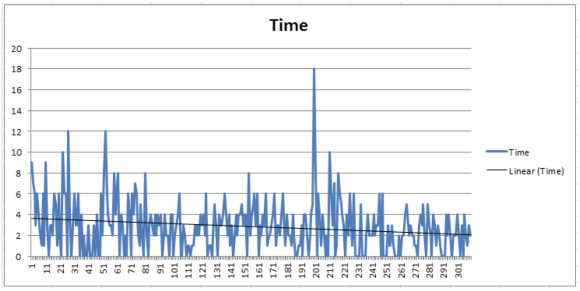

Figure 2: Average time for making a decision.

Categories	Technology	Others									
		Innovation in Services	Trend shift	Behavioral Change	Major actor move	Discovery	Studies	Wildcard	NS	total votes	Total votes/ total events
Votes	606	126	176	60	184	104	132	13	401	1802	1.24
Unique classification	367	26	53	26	70	46	100	11	401	1096	

Figure 3: Weak Signal taxonomy Distribution

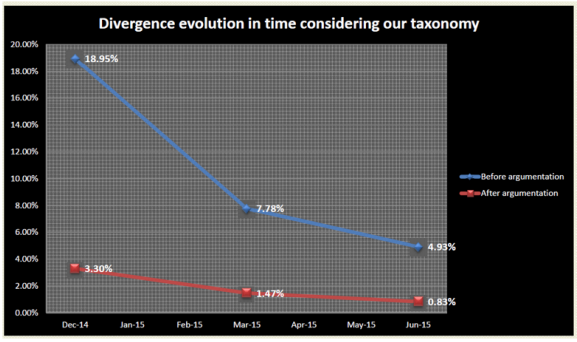

Fig 4. Reaching consensus over taxonomies

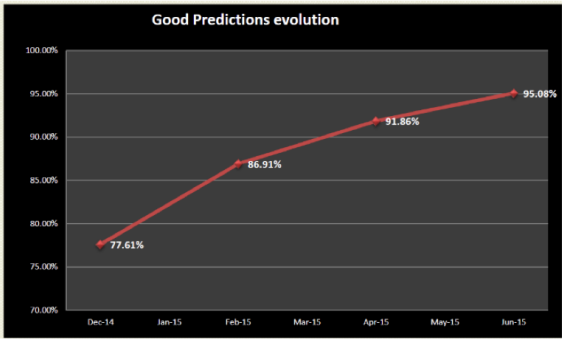

Fig 5. Control Group Judgement

$$P(y = k|X) = \frac{P(X|y = k)\,P(y=k)}{P(X)} - \frac{P(X|y = k)\,P(y=k)}{\sum_l P(X|y = l)\cdot P(y=l)} \quad (1)$$

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \pi(f(X_i) \neq Y_i \quad (2)$$

equivalent to none of the above, is always a valid option.

It came as a surprise that the annotators did not want to use often the wild card taxonomy. It can be seen that the number of documents that received just one category is relatively high and quasi constant (50%). The number of documents that received more than three categories is non-significant, less than 3%. In Fig 4. we draw the dynamics of reaching consensus among annotators. We wanted to check whether this consensus was reached due to an increasingly strong and commonly shared tacit knowledge, that is, due to acquiring an expertise, or due to accepting a dominant view.

A control group checked the validity of the agreement and what we found is that the results strongly suggest the first alternative, that is acquiring an expertise, see Fig. 5.

In conclusion, all these experiments strongly suggest that we have a tacit knowledge about weak signals that is shared at least 80

4. Learning Weak Signals

In this section we present a series of learning approaches which we tried step by step.

In a supervised approach, finding the pieces of news containing weak signals is a binary classification task. A first approach is to use tfidf weights to compute the similarity between a document and the documents in one of the two classes. This provides us with a weak baseline. However, it is an informative one. It tells how much of the weak signals are judged to be expressed via some special words or patterns. Anticipating, it turns out that this is not the case at all. This baseline has negligible accuracy, far distanced from the best results we obtained eventually. This preliminary finding confirmed that the task is not trivial at all and that many clues on the basis of which a human judges the correct answer are not necessarily expressed by clearly defined overt phrases. As such, we can use a couple of offthe-shelf approaches that will provide a set of baselines for this task. We looked at two libraries which implement quadratic discriminative analysis, QDA from scikit library, and support vector machine, linear SVM form Weka library, respectively. See also the equations 1 and 2.

The reasons behind our choice have to do with the type of data we employ here. The fact that the tf-idf obtained a very low score does not immediately imply that maximizing the prior probability P(word—weak signal) is inefficient. In fact, we will see in the next section that the gradient descent is an effective technique for this task. At this point, we have to understand whether the projection of the data into a bi-dimensional space will lead to con-like structures, that is, that the data can be separated by a quadratic function. On the other hand, if the difference between the SVM and QDA is large enough this will show that QDA suffers from the masking effect. We run both QDA and SVM in a cross-validation setting, 10 folds 1/8 ratio for train/test and 1/8 ratio for development/train. That is we used a tenth of the corpus for test and development respectively. For test we used 500 weak-signals and 500 no-signals. In Table 2 we present the results for QDA and SVM for SC, and in Table 3 the results for LC for cross validation. The tf-idf scored 0.18 for SC and 0.12 for LC respectively. As we can see both QDA and SVM scored significantly better than that. And indeed there is a non-random difference between QDA and SVM results.

To understand better the nature of this difference we ran a series of experiments alternating the ratio of weak signals in the training corpus. We found no significant differences from Table 2 and Table 3. This shows that probably we cannot improve these results by adding more training. Given that SVM is a constraint over a large boundary for ——*Ein-Eout*—— and that the differences from QDA are large, eq. 1, it follows that it is possible to search for a better model even further. That is, particularly for this task, we could find a better estimation, as the worst case scenario seems not to characterize this corpus. Because we cannot directly compute the number of dichotomies, and therefore, the exact VC dimension is unknown, on the basis of the Tables 2, 3 it is intuitively tempting to consider that the VC bound is indeed too loose for this task. As such, we can do better in estimating the posterior probability. The right question is whether we have enough data to train a more detailed clas-

	Weak Signal	No Signal		Weak Signal	No Signal
QDAcr	0.412	0.877	QDAcr	0.38	0.901
SVMcr	0.663	0.913	SVMcr	0.472	0.946
QDAts	0.403	0.865	QDAts	0.365	0.890
SVMts	0.610	0.905	SVMts	0.455	0.930

Table 2: Supervised Learnig of Weak Signals

sifier. We may guess that deep learning methods may be up to the task.

5. Conclusion

In this paper we presented an experiment on prediction. Rather than final, we consider these results as a very promising beginning for research into this field. The possibility of trend prediction on the basis of weak signals is very exciting and it has a lot of applications. Our study shows that even when we do not know what the weak signals are, we are still able to use them in predicting future trends via supervised learning. This is an excellent result, showing that it is viable to talk about predictions. In many applications, it becomes critical to have an accurate prediction. In science, making predictions almost equates to having a bright idea on how apparently disparate small achievements may converge to a breakthrough discovery. In our digital era, accessing billions of documents is easy but selecting the ones carrying relevant information which is not yet fully developed is difficult. A starting point is to understand better how we could narrow down the search for weak signals. The results suggest that we can have a major improvement of several points if we could pin point a paragraph instead of a document as source of weak signals. So our next effort is to narrow down the search for the pre boom period at the paragraph level, rather than document level.

6. References

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:.

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., and Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, 2(1):11.

Gerrish, S. and Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *ICML*, volume 10, pages 375–382. Citeseer.

Mueller, J. and Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., and Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(4):694–707.

Popescu, O. and Strapparava, C. (2013). Behind the times: Detecting epoch changes using large corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355.

Popescu, O. and Strapparava, C. (2014). Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Wang, X. and McCallum, A. (2006). 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*, pages 424–433. ACM.

Wang, X., Gerber, M. S., and Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 231–238. Springer.

Xu, J., Wang, P., Tian, G., Xu, B., Zhao, J., Wang, F., and Hao, H. (2015). Short text clustering via convolutional neural networks. In *VS@ HLT-NAACL*, pages 62–69.