Le Pham Minh Duc

Un-polarizing news in social media platform

Master’s thesis of mathematical information technology

January 30, 2019

Author:Le Pham Minh Duc (Lê Phạm Minh Đức)

Contact information: [miduleph@student.jyu.fi](mailto:miduleph@student.jyu.fi), minhduc1993@yahoo.com

Supervisors: Oleksiy Khriyenko (oleksiy.o.khriyenko@jyu.fi)

Title: Un-polarizing news in social media platform

Työn nimi:

Project: Master’s thesis

Study line: Web Intelligence and Service Engineering

Page count: x+y (x = page count without appendices; y = page count of appendices)

Abstract: Abstract…

Keywords: Keywords…

Suomenkielinen tiivistelmä: Abstract in Finnish…

Avainsanat: Keywords in Finnish…

Preface

This is where you can write a preface for your thesis. The preface is optional so you don’t have to write it if you don’t want. Also, the glossary on the next page is optional. List of figures and tables are required if there is at least one figure or table in the thesis.

**The most important guideline** for the use of this template: look from the title list which style has been used to which section of the document and use similar styles also in the future when modifying the document. Use always styles, never do formatting outside styles.

If you notice bugs or mistakes in the template, please inform them to Antti Ekonoja ([antti.j.ekonoja@jyu.fi](mailto:antti.j.ekonoja@jyu.fi)).

Jyväskylä, January 30, 2019

Le Pham Minh Duc (Lê Phạm Minh Đức)

Glossary

NLP Natural language processing

DCOM Distributed Component Object Model  
More explanation…

C++ Shouldn’t need any explanation…

List of Figures

Figure 1. Service architecture 4

Figure 2. Test figure 6

List of Tables

Table 1. Example of the table 4

Contents

[1 INTRODUCTION 1](#_Toc536649045)

[1.1 Problems overview 1](#_Toc536649046)

[1.2 Proposed solution and research questions 1](#_Toc536649047)

[1.3 Research method 2](#_Toc536649048)

[1.4 Thesis structure 3](#_Toc536649049)

[2 HYPOTHESIS AND TECHNOLOGIES 4](#_Toc536649050)

[2.1 Original hypothesis 4](#_Toc536649051)

[2.2 Evaluation criteria 4](#_Toc536649052)

[2.3 Required technologies 5](#_Toc536649053)

[2.3.1 Natural language processing and its sub-domains 5](#_Toc536649054)

[2.3.2 Open Information Extractor / Semantic triple 6](#_Toc536649055)

[2.4 Technologies used in this works 6](#_Toc536649056)

[2.4.1 Stanford CoreNLP 6](#_Toc536649057)

[2.4.2 Node JS 7](#_Toc536649058)

[3 UN-POLARIZING ALGORITHM 8](#_Toc536649059)

[3.1 Overall solution/algorithm architecture 8](#_Toc536649060)

[3.2 Article annotator 8](#_Toc536649061)

[3.2.1 Web content parser 8](#_Toc536649062)

[3.2.2 Stanford Core NLP Annotator 9](#_Toc536649063)

[3.2.3 Sentiment analysis 9](#_Toc536649064)

[3.2.4 Named entity recognizer 11](#_Toc536649065)

[3.2.5 Open information extraction 13](#_Toc536649066)

[3.3 Un-polarizing algorithm 13](#_Toc536649067)

[3.3.1 Articles similarity calculation using only NER 13](#_Toc536649068)

[3.3.2 Triplet pairs, common entities and common statement counts 13](#_Toc536649069)

[3.3.3 Articles relevance processor (1 to 1 comparison to get the relevance score) 13](#_Toc536649070)

[4 RESULTS EVALUATION 14](#_Toc536649071)

[5 FUTURE WORKS 15](#_Toc536649072)

[5.1 Ontologies based entity relevance 15](#_Toc536649073)

[5.2 Word-net verb contradiction 15](#_Toc536649074)

[5.3 Cloud service design 15](#_Toc536649075)

[5.4 User interface and user experience design 16](#_Toc536649076)

[6 CONCLUSION 17](#_Toc536649077)

[Bibliography 18](#_Toc536649078)

[Appendices 19](#_Toc536649079)

[A Title of the first appendix 19](#_Toc536649080)

[B Title of the second appendix 19](#_Toc536649081)

# INTRODUCTION

## Problems overview

(*Should write like a lot more, with more references as well*) Ideological polarization has been a problem in our society for quite a long time. (that reference from 1986). With the rise of social media (citation here or not), it’s estimated that 66% of US citizen use social media as one source of news (citation from that web page), the amplification of ideological polarization has been increasing much faster than previously due to social endorsement, and other social media techniques that is used to keeps its user engaged (Sporh. 2017). This creates the echo-chamber effects that, by the design of social networks that only show what the users want to see, make the user even furthermore polarized into his own belief and makes him see the world wrongly, which may turn the user into some extremists that might be harmful for the society.

Scholars have been researching about this problem and solutions are proposed (many citations needed) but these solutions are either too impractical (e.g: needing the giant media companies to change their entire business models) or just way out of reach of the scholar scope (e.g: needing of the government’s intervention on the issue or people to stop using the services).

## Proposed solution and research questions

(*The question seems nice*) The main goal of this thesis is to find the way to break the people’s echo chamber that is mostly caused by the effect of social medias only show the user what he/she wants to see. To combat this, we need to show the user the news from the other side of view. If he/she reads about the opening of a new coal mines help creating a few hundreds of new jobs for the area, he should also know that the new coal mines will cause a great damage to the environment and might cause some local wild-life to disappear.

On top of that, the service must be accessible and easy to use, as the reason of many people using social media as their main source of news as it’s so convenience to have one place to go to and can see both your friend’s status as well as news.

With that goal in mind, the main research question of the thesis is:

* **How to find articles with alternative (different) points of view to a given article?**

We will only attempt to find the news that is relevant to the article but also provide oversight from different point of view that the first article misses. We will not check if the news is credible (but we will try to only provide news from credible sources) or if it is true, we simply provide the user different articles from many points of views so that he/she can choose to interpret it whatever he/she wants to.

Afterwards, with the first question answered, we will address two additional support questions on deploying the news un-polarizing service for the mass to use:

* **What is the most convenience way to deliver the service for the user?**

If the service is too complicated to use, or requires too many unnecessary steps, the user will rarely use the service at all, which will defeat the whole purpose of the system.

* **How to engineer the service so that it is autonomous, up-to-date and scalable?**

As a news service, it must always catch up with the latest news to be relevant to use.

## Research method

(*Just copy from the book*) Abc, test text. Please fill in later.

## Thesis structure

(*Fill this later*) This thesis will be divided into four main parts, including this introduction. The second chapter, named “Un-polarizing algorithm” will attempt to answer the main research question on how to find articles with alternative points of view to a given article.

As this chapter attempts to answer the main thesis question, including the theoretical approach as well as solution implementation from the ground, it thus can be furthermore divided into 4 smaller sub-chapters in which, the first sub-chapter starts with the fundamental technologies used for the works, the author original hypothesis and the reasons for choosing the services used in the thesis implementation part. The second sub-chapter focus on processing and storing the articles/news/documents retrieved from the internet using the technologies presented previously. With the articles processed and stored, the third sub-chapter provides the un-polarizing algorithm that answer the main thesis questions. Finally, the second chapter ends with evaluation of the un-polarizing algorithm as well as comparison to other existing solution.

The third chapter tackles the two subsequence researches question of providing the un-polarizing algorithm in the most convenience way for the general mass to use and how to develop the algorithm to be a cloud service that can support a large amount data if needed.

Finally, conclusions for the thesis as well as possible future work and extension are given in the last chapter.

# HYPOTHESIS AND TECHNOLOGIES

## Original hypothesis

(*Write this to be longer, reference???*) Our original hypothesis is that: given two articles talking about one similar topic (*for example: The US’s involvement in Iraq*), if one article has a positive views on the situation (*ie: Saddam Hussein is a terrible man and the people living under his reign are suffering badly*) and the other has a negative views regarding the same situation (*ie: it furthermore destabilizes the region and the main intention of the war was because of oil*). With articles telling about the same story but with different sentiment value, it could be interesting for the reader to see from different kind of attitudes about the same topic, thus, bring him to different point of views about a problem.

This hypothesis, however, after some implementation and evaluation, was proven to be not good enough (**more in-depth in chapter 2.3.2 Sentiment analysis**). The final solution of this thesis utilizes more complexed calculations and processing techniques that was not originally planned from the start. However, the initial hypothesis did create a solid technology base to start working on: Named entity recognizer (*to understand the article topic*) and sentiment analysis (*to understand the positivity/negativity of the article*).

## Evaluation criteria

(*Longer, maybe no reference needed*) TDD – Test driven development -To test the rigidity of our algorithm, we gathered a small dataset of 78 articles (+ 8 non-readable by the web content parser) – *could be more, should update* – with 3 main themes: Muslim in Europe (24 articles), Muslim in Asia (39 articles) and Asians in Europe (17 articles) (*somehow, the number of articles in each categories and the total number doesn’t match. Something must be wrong with the* listing*, we need to re-check these things*).

With these hand-picked data, we can look through each article, to judge for ourselves which we think is the most relevant, and which is not, and then, compare our result to the result returned from the algorithm and judge the result for ourselves. With these three different categories, we can make sure that there we will know if the algorithm returns the relevance information or not, how close is the suggestion and in some case, if the return result can even surpass our hand-picked solution (because human is flawed and cannot read through all these data).

## Required technologies

### Natural language processing and its sub-domains

(*Consider rename the title*) (*Lots of references in this, this is like a wildly researched fields*) Named entity recognizer and sentiment analysis are two of the many sub-tasks covered by Natural language processing (NLP) technology, *which is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, how to program computers to process and analyze large amounts of natural language data. (Source from wiki- more information needed).*

For just named entity recognizer only and sentiment analysis only, there exists multiple tools that do not require NLP. For example, a simple named entity recognizer (*citation needed)* consists of a dictionary of many individual nouns and a lookup function to match the result from the document and the dictionary. Similarly, a basic sentiment analysis can be made by assigning a sentiment value to each of the adjectives in the sentence/paragraph and then calculate the sum of these values as the sentiment value for the sentence/paragraph. However, these methods are quite barebone and quite usually not really correct since their rules are quite flawed and there are many ways for them to misinterpret the true sentiment of the sentence. Sophisticated NLP tools with “sentiment analysis” supported have a more complex ways of defining the sentiment value

Named entity recognizer and Sentiment analysis can be easily created without NLP using a rule-based system. However, these basic tools are not good and sophisticated NLP tools are much better because they have a more complex system that take many other things into consideration.

### Open Information Extractor / Semantic triple

(*What is semantic triple, why it is important and a critical part of this thesis.*) – After our first implementation of the algorithm, we realized that just Sentiment analysis and NER was not enough. To make the result more interesting and more useful to the user, we have to use Open Information Extraction, which, unfortunately, was not as widely developed as the two previous features since they are much more complicated.

*https://nlp.stanford.edu/software/openie.html*

*https://nlp.stanford.edu/pubs/2015angeli-openie.pdf*

[*https://getd.libs.uga.edu/pdfs/hooge\_david\_c\_200705\_ms.pdf*](https://getd.libs.uga.edu/pdfs/hooge_david_c_200705_ms.pdf)

[*https://en.wikipedia.org/wiki/Semantic\_triple*](https://en.wikipedia.org/wiki/Semantic_triple)

*https://github.com/dair-iitd/OpenIE-standalone*

## Technologies used in this works

### Stanford CoreNLP

(*This could have a lot of comparison and references, or not, depends on my mood*) Many big companies offered NLP services like Microsoft, IBM, Google, each with their own technology and strong/weakness. These tools, however are 1. Expensive to run, 2. Close source and are subjective to change in any moment (like IBM) which make them not attractive to use.

Stanford CoreNLP is an NLP tool created by Stanford university, free and open source and quite easy to setup/use and very powerful as it supports many features.

After some testing some of these tools, we decided to go with Stanford CoreNLP because Microsoft is quite weak and in-accurate,

IBM is expensive and is a pain to setup and they change the name/domain quite often so not reliable, even though their named entity recognizer and sentiment analysis are really good.

Google is ok, looking and feeling very similar to Stanford CoreNLP (like they took the source code of NLP and improved it upon). They strip a lot of features from CoreNLP but the things they keep (like NER and Sentiment) are much better than the stock version in CoreNLP.

Still, I would like to have everything in one place, free and will guarantee to work no matter what, so, I went with Stanford CoreNLP. The important part of this thesis is the algorithm behind the news unpolarizer, not from how can I choose a better service to use.

Furthermore, if we want to include OpenIE in our tool, Google or IBM or Microsoft does not offer such feature, so Stanford NLP or that standalone Openie tool is our only option. Fortunately, both of them are free and opensouce.

### Node JS

In this work, we choose NodeJS to do our coding, as it’s one of the easier and faster programming languages to code in, as well as the author expertise in it would make implementing the algorithm faster than other language.

# UN-POLARIZING ALGORITHM

## Overall solution/algorithm architecture

(*Read this and fix this carefully – also, add a nice graph*) – Two main parts of the algorithm: annotating and article matching.

## Article annotator

### Web content parser

Web content parser. The article on the internet are presented inside a web page, with just not only the article itself, but with tons of other unnecessary things like html tag for formatting, images and captions, links to other things on their website and advertisement text.

To strip away all the unnecessary function, I have a web content parser module which utilize a similar technique to reader mode on Firefox or Safari which will automatically strip away all the non-article part in the web content. However, this is not enough since it only strips the advertisement and related news, the core article and the html tag around it still persists, which, to solve, I wrote a smaller module to automatically remove all the html tag as well as the image and the caption, which is not really accurate since each website have different layouts and ways to present content, make parsing out the content really not effective, since, for example, some website, when they end the sentences or the caption of the image, the don’t add the “.” (dot) or sometimes they make multiple dots, which make finding out the article contents with proper sentences are quite a problem in most of the article, which make the overall annotating result worse.

One effective way of ensuring that the sentences forwarded into the CoreNLP annotator are correct is to use a tools call SMMRY, an article summarization tools, which will read through the article and gives out the sentences that it thinks contains the most important information of the article. This tool is quite effective for our case as it first, strip away all of the garbage contents like html, tag and about this website … which make the article annotation work correctly. One downside of this is that it will not return the whole article, only part of it (the more important part) so there might be possible information that will be lost during the stripping of the content. Still, I think it’s better to use this SMMRY tool instead of my own web content parser because it’s not very good and accurate.

When all the data-collecting and parsing is completed, we push the parsed data into our annotation pipeline.

### Stanford Core NLP Annotator

(*What is the role of the Stanford Core NLP in our program and why we need the local db*) - After extracting the text part from the web article, we push the text into the Stanford Core NLP engine. Stanford Core NLP have many annotators, like NER, Sentiment, lemma … (*find more*), we can specify which annotator we want to run through to save processing power. After receiving the annotated result from the Core NLP, we will run the result through our special filter for sentiment, ner and openie to retain the only relevance information for our un-polarize algorithm.

The filtered annotation result (let’s call them core feature) will be saved into the local database for future comparison of the un-polarizing algorithm. The use of the local database for the core-results is necessary because the processing time of each article is quite long, around 30 seconds each, so, to compare one article to other hundred to find the most relevance one, we must process them beforehand so that our results are readily available for future accessing.

### Sentiment analysis

(*How do we use the sentiment analysis, how it’s not working and how we decided to just not use it*) – This is our very first idea upon figuring out a solution to this problem. Our initial hypothesis was simple, if two articles talking about one problem, for example: The U.S President, Donald Trump, if one article talk about him with a positive sentiment, and the other with a negative sentiment, then they are from the different point of view and we should suggest the other article to our user.

Sentiment analysis services are offered by many big companies like IBM, Google, Microsoft, to some other smaller startups and a lot of open-source library exists as well. However, this method proves to be completely useless for our use-case (at least using the Stanford’s NLP lib). Because practically, sentiment analysis is only good for single sentences only, without taking into account the whole paragraph or the article. It is mostly used for analyzing customer reviews for products or customer feedbacks.

Furthermore, the Stanford NLP sentiment analysis result are quite limited, only 5 values, from very negative to normal to very positive, and work on the context of a single sentence only. Google’s service proves to be a bit more useful with sentiment scale from 0 to 100, but still work on a single sentence context. IBM’s service is the fanciest with scale of 100 for sentiment but also some other adjectives like “Anger”, “Nice”, “Happiness”, … which seems to be the most suitable for our case if we decided to go with this.

However, as this is a public research with source code fully available for people to use on their own, we don’t want to tie ourselves to private services that might either be unavailable someday, or restrict our access because of reasons, we decided to go with Stanford NLP, the open-sourced solution from other academia.

We calculate the sentiment value for article using Stanford NLP using the weighted average sentiment value of each sentence in the article. In which, we calculate the average value for the whole thing, but longer sentence (by characters count) will have a bigger weight than short sentence.

* **V** is the sentiment value of the article.
* is the sentiment value of sentence *i*
* is the length in number characters of sentence *i*

The sentiment analysis hypothesis later on proves to be almost useless, as the sentiment value of the sentence/paragraph have very little correlation to the content in the articles. Which mean, even if the paragraph talks about the killing of Yemen people, it might still have a normal or positive sentiment (*TODO: find example*), because journalism is usually supposed to give provide information in the most neutral way, so it’s hard to find any correlation between them. Furthermore, it’s kind of easy to fool the system, using word like nice, good, or similar to that, to make it have a higher sentiment value.

In short, sentiment analysis proves to be almost useless.

### Named entity recognizer

(*How do we use the NER, the initial solution/algorithm use NER. Briefly the result of the one using only NER*) – Named entity recognizer (NER) is also presented in our original idea. With the sentiment value used to judge the \**difference in point of view*\* between two articles, we use named entity recognizer (NER) to find the relevance articles to the one the user is reading and then, compare the sentiment result to suggest it to him.

The default NER of Stanford NLP contains a lot of unnecessary information that bloat the return result for our algorithm so much. Many information (named entities) detected are quite un-relevance to our case, for example, he/she, year, number and some proposition text like Mister, Miss … We have to write a filter for the NER results to get the relevance information only. For the user convenience, we also split the result into two categories: abstract entities and discrete entities.

The two groups contain: (*maybe write this in a table?*)

Discrete entities: PERSON, LOCATION, ORGANIZATION, MISC, CITY, STATE\_OR\_PROVINCE, COUNTRY

Abstract entities: RELIGION, NATIONALITY, TITLE (Job title), IDEOLOGY, CAUSE\_OF\_DEATH (violence, shooting ....)

Stanford Core NLP work on a single sentence basis only, which mean the meaning of the next or previous sentence in the paragraph doesn’t affect the annotation of one sentence in anyway, which is a bad thing for us since we want to work on a rather macro way, instead of micro based way like this. Thus, after filtering out the un-needed result, we perform the entities counting, in which we count the number of each time an entity has appeared in the article, as well as calculate the sentiment value or the average sentiment value of that entity, and finally, save all the result to the local database.

|  |  |
| --- | --- |
| (*Remove the table border later*) – Initially, the annotated value stored for one article in our database look like the part on the right:  All article will be stored as a JavaScript object, in one big single database.json file.  Each annotation object contains:   * Meta data about the articles: url, title * Annotated title, which contains sentiment value, length, and entities appearance * Annotated contents of the article, which contains information on all of the entities, their appearance and their average sentiment value. | Figure n: Example of an annotated article stored in our database |

### Open information extraction

(*How openie improve our result, how do we use it and store it information*) - Abc

## Un-polarizing algorithm

### Articles similarity calculation using only NER

(*What a crap result we got with NER*) – Some mathematics and equation here

### Triplet pairs, common entities and common statement counts

(*What are these 3 above stuffs and our sorting equation*) – Some mathematics and equation here

### Articles relevance processor (1 to 1 comparison to get the relevance score)

(*List of top 3, what is presented to the user*) – Some mathematics and equation here

# RESULTS EVALUATION

(*Yeah, I don’t know, should have some smaller sub part here, Maybe*) – A lot of bullshit here

How the result is presented to the user, each relevant information and stuffs like that.

# FUTURE WORKS

## Ontologies based entity relevance

(*How using ontologies, can help finding similar words/entities, possibly even link Chinh’s thesis*) - Abc

## Word-net verb contradiction

(*With the triplet pairs implemented, we could find contradiction between the triplets. Using word-net to find verb that have similar meaning/or opposite meaning*) - Abc

## Cloud service design

Local

database

Web

content

processor

Article

annotator

Requested URL

News gatherer

News

suggestions

module

Suggested

articles

Figure 1. Test figure

With the algorithm ready, we need to automatically get our data somehow. Fortunately, there are a lot of news APIs available, for example, Google News or many other things, just one simple APIs and it can give you all the thing you want. It’s also good for evaluation later, because these APIs allow you to search for query by word, date and time, which will be useful to compare the results between our stuffs and theirs.

## User interface and user experience design

# CONCLUSION

Hope you enjoyed the text...

In the bibliography the recommendable style is Chicago. You can also use other styles: the main thing is that the styles of the bibliography and referring technique are **consistent** in the whole thesis.

|  |  |  |  |
| --- | --- | --- | --- |
| Word | Year | Magnitude | Example |
| example | [1700,2000] | [1,10] | example |
| example | [1950,2000] | [1,106] | example |
| example | [1995,2000] | [10–6,106] | example 1, example 2 |

Table 1. Example of the table

Bibliography

Spohr, Dominic. "Fake News and Ideological Polarization: Filter Bubbles and Selective Exposure On Social Media." *Business Information Review* 34.3 (2017): 150-160.

Meikäläinen, Matti. *Tietotekniikan perusteet.* Helsinki: WSOY, 2000.

Santanen, Jukka-Pekka. "Opinnäytteiden kirjoittaminen, lyhyt oppimäärä." 2000. http://users.jyu.fi/~santanen/info/kirjoittamisesta.html (accessed 5.10.2012).

Writer, John. "Name of the article." *Computer*, no. 2/2000 (2000): 31-32.

Appendices

1. Title of the first appendix
2. Title of the second appendix