Determinants of Politically Charged News in Media: Relationships Between Entity and Sentiment Tonality

MSc Research Project
Data Analytics

Sara Guzman
x13120701

School of Computing
National College of Ireland

Supervisor: Simon Caton

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Sara Guzman
Student ID:	x13120701
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Simon Caton
Submission Due Date:	21/12/2016
Project Title:	Determinants of Politically Charged News in Media: Relationships Between Entity and Sentiment Tonality
Word Count:	4576

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	21st December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Determinants of Politically Charged News in Media: Relationships Between Entity and Sentiment Tonality

Sara Guzman
x13120701
MSc Research Project in Data Analytics

21st December 2016

Abstract

Politically charged news usually have important and powerful information that is expected to reflect the political situation of an event without bias expectations. As text document type of communication, it reflects emotions and sentiment tonality which recently has become a topic of study and research, not only for linguistics but also for data scientists. These emotions reveal psychological insights about the way the author is presenting the news, their perspective and also, meaningful hiding sentiments in the lexicon. Lexically expressed opinions can be found in news texts. This, represents one of the reason to choose the research topic of this paper, mainly based on the identification of the determinants of politically charged news in media. Specifically, the anger sentiment tonality and the relationship that has with the major entities of the article. For this, Text Mining, Natural Language Processing and Machine Learning played the main role in providing the methodologies and tools to implement for the analysis and hypothesis testing. Decision Trees and Association Rules models were implemented to prove the research hypothesis were true.

1 Introduction

The globalization nowadays is lowering the barriers of countries, languages and communications in general. More than ever before, we live in a multicultural environment with people from all over the word moving around freely. As a consequence, social media and news in specic, play a huge role in our society as information tool. In this context, the way we interpret and analyse news and social media is strongly related to the way it is presented to us. Specially, politically charged news, that no surprisingly can contain polarity, bias and in many cases extremist sentiment tonality. Fortunately, text analysis and sentiment analysis are the growing study fields, that identifies and analyse language and sentiment in an automated and objective way.

This paper, is based on the study of sentiment tonality in politically charged news media. The research was formulated to identify relationships between entities and the determinants of negative tonality in a corpus of text data articles. An entity is described in this research as a person, organization, event, product or topic of discussion according to (Barhan and Shakhomirov; 2012). A software platform for the identification of the entities in the news articles was selected, followed by a Linguistic Inquiry and Word Count (LIWC), which is computerized text analysis program that works by reading the text and counting the percentage of words that reflects different emotions, thinking styles and social concerns and part of speech. Stated by (Tausczik and Pennebaker; 2010) provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals verbal and written speech samples, we originally developed a text analysis application called Linguistic Inquiry and Word Count.

1.1 Research Formulation and Hypothesis

In this context, the research formulation was: to identify the relationship between the main entities in the articles. For this, statistical work in text data was applied, as it has been found to bring practical solutions to the NLP methods, (Schuetze; 1997). Subsequently, Machine Learning techniques were used to mine and associate the news entities with Association Rules Mining. Finally, to predict anger tonality in news articles that are politically charged by training a machine model and find the entities that represent the determinants of anger. This was developed with the use of Decision Trees.

With this in mind, the hypotheses of the project were formulated as it follows:

If there is a relationship between sentiment tonality and the main entities in the article, machine learning techniques as association rules mining will associate the entities.

If the selected entities of the news articles present a negative sentiment tonality, then a machine model will predict it.

In order to test the hypothesis, the Methodology section describes the steps that were taken into consideration for the implemented process and therefore the analysis. The paper is structured by the following sections: the next segment situates this work in the field through a discussion of related literature, consequently the methodology and evaluation section describes the approach justifying key decisions and assumptions. The evaluation which is the next section, presents the execution and testing of the hypothesis. Conclusions and future work expose the discussion and summarised findings of the paper.

The study represents a contribution to Natural Language Processing and Computational Linguistics field. Also to Machine Learning and Text Mining techniques in polarized media tone. Therefore, it also contributes to the Political Sciences research. This research investigation is focused in the English language. It is highly transferable to other fields within the text mining area as other languages thanks to the years of research of LIWC. The media system plays a crucial role, understanding the workings of this complex system is of crucial importance for society business and democracy. (Fortuna et al.; 2009)

2 Related Work

Research in Text Mining Analysis has been lately growing exponentially, which makes it a challenge when it comes to analyse text in different languages. Fortunately, there is work done in this field and involves many languages with a big corpora of text to understand human language expressed in text. LIWC is a Linguistic Inquiry and Word Count text analysis program that has the ability to link the daily word use with a broad array of real word behaviour.(Tausczik and Pennebaker; 2010). This linguistic tool is being implemented in several Natural Language Processing and Social Science studies, as it is of main use in this project. Further research related to the topic is described in this section

2.1 Linguistic Inquiry and Word Count

(Fuller et al.; 2011)implemented LIWC in their research about linguistic cues to deception. Their research tried to find negative emotions in liars. In this context, their study has few similarities with the scope of this research, when analysing negative sentiment tonality in news articles. LIWC has also been implemented in studies that detect (implicit) racist discourse. (Tulkens et al.; 2016). In this context, news in media has shown to present several different sentiments that with or without intention appear in the text communicating a tonality that can be perceived by humans but not for machines. Following this idea, Natural Language Processing, and Text Mining uses several approaches to help to deal with this problem. In order to find meaningful insights in the corpora, there are some statistical NLP approaches used to discover knowledge and mine deep in relationships of entities in text. There is a lot of information in the relationships between words, that is, which words tend to group with each other. (Schuetze; 1997). Text mining techniques uses Machine Learning Algorithms to find associations, hypotheses, or trends that are not explicitly present in the text sources being analysed. (Nisbet, Miner and Elder IV; 2009).

Studies on sentiment and media tonality have a restrictive cope of study automated methods exhibit a strong language bias as they are developed and validated predominantly with textual data in English language according to (Haselmayer and Jenny; n.d.). Fortunately LIWC is an automated text analysis program developed in several languages. Nonetheless there is evidence that there is a need of more research in the topic. Some NLP libraries are not very accurate and some do not really perform a complete task of entity extraction and recognition for example.

2.2 Association Rules and FP-Growth

This research used an online platform for the Entity recognition and traying to find more insight in the data by analysing and correlated the common lexemes. Lexicon is a set of meaningful units in a language according (Eisner; 2001). In in this context, lexicon is used in this study, referring to the specic set of words that contain the tonality information what its required to be found, in order to do so, machine learning algorithms for classification and association were the most used for other scientists and researchers. Specifically, the use of Association Rules Mining which identifies strong association and relationships in data. Recently several researchers have applied traditional rule induction methods to discover relationships from textual data. (Mahgoub and Rösner; 2006) ARM.

In this context it is important to mention that the algorithm used in this project was the FP-Growth. FP-Growth works encodes de dataset using a compact data structure called FP-Tree and extracts frequent items directly from this structure. It works with the platform for machine learning used in this project. Apriori on the other hand it slow according to (Wu, Lu, Pan, Xu and Jiang; 2009) they propose a new model based on the Apriori algorithm for generating association rules. The algorithm not make multiple scanning on the original documents as Apriori algorithm but scan only the generated XML file which contains all keywords that satisfy the threshold weight value and their frequencies in each document. Also Association Rules is atechnique used to discover relationships among a large set of variables in a data set. (Gupta and Lehal; 2009)

2.3 Decision Trees

The other selected method is Decision Trees. According to, (Mairesse and Walker; 2006), an analysis of decision trees confirms previous findings linking language and personality, while revealing many new linguistic markers. Decision trees machine models are evaluated by their precision, recall and accuracy. They classify instances by sorting them based on feature values. (Kotsiantis, Zaharakis and Pintelas; 2007) Decision Trees perform a classification task, which was implemented as a prediction task. In their paper (Mairesse and Walker; 2006) stated, decision trees can be easily understood, and can therefore help to uncover new linguistic markers of personality. Our models replicate previous findings, such as the link between verbosity and extraversion. Classification results are capable of representing the most complex problem given sufficient data. (Nikam; 2015)

3 Methodology

Machine learning techniques have proven to be accurate and user-friendly, (Schrauwen; 2010). There are two main types of machine learning tasks: Supervised and unsupervised Machine Learning (ML). Supervised ML was the implemented in this project because the research question was classification problem. The methodology follows the KDD Process, (Knowledge Discovery in Databases) which is the method of turning data into knowledge. (Fayyad et al.; 1996). Figure1 shows the KDD process.

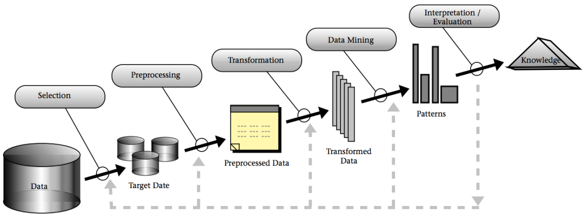


Figure 1: KDD Methodology

The high level architecture of KDD describes the methodology from the selection of the data selection until the knowledge discovery:

3.1 Data Selection and Collection

According to the type of research, the required text data for the project was an up to date database body of news articles. As the project identifies negative sentiment tonality and the relationship of the entities, the news articles were required to be in English language and within the time of Brexit news. The topic Brexit we selected also for having controversial news that were expected to involve strong sentiment tonality and extremism at some point. The selection of the data was a strategic part of the project. This, due to a diverse variety of newspapers sources from all over the word, would minimize bias in the selection of the news. Information retrieval and machine learning tools were implemented for extracting the data.

3.2 Data Pre-processing

As the collected text data was unstructured, in this stage the data was cleaned and stored in a NoSQL database. Unstructured data makes the cleaning process of text challenging. Nonetheless, the gathered data was consistently of high quality. As part of the pre-processing, the extraction of the entities in the news articles was required. In order to test the hypothesis, the selection of the main entities in the articles allowed to the identification of the specific Person/Organization/Location. The selection of a narrow scope of entities made the study more specific. Also, a Linguistic Inquiry and Word Count was added to the selected entities.

3.3 Data Transformation

Subsequently, the next step was to transform the corpus of articles to a new dataset by joining the entities and word count. This created a new numerical dataset for a statistical analysis. Next, by applying a rule based approach data was transformed to categorical for a ML approach.

3.4 Data Mining

Supervised Machine Learning Techniques for Text Mining were implemented in this step. One the research hypothesis required the identification of relationships between entities and sentiment tonality, therefore, one of the models that was selected for this requirement was Association Rules Mining (ARM). This model was mined in order to identify strong interesting rules from the association between entities in the corpus of data. This technique allows analyst and researches to uncover hidden patterns in datasets. (Nisbet et al.; 2009) Association rules has been lately used in text and linguistic researches. (Mahgoub and Rösner; 2006) presented a system for extract association rules from collection of textual documents based on word feature. Association rules appears to perform well in text mining tasks and entity extraction.

The second implemented model was Decision Trees, which belongs to the classifier algorithms but was used in this research as a predictor. The second hypothesis of the study stated to predicted negative sentiment tonality in articles. In this task, decision trees as a classifier have been used, as it performs well according to the dataset with a typical level of 20. (Witschel; 2005) used decision trees for extending taxonomies. It is important to mention that for the Entity Extraction in news, several NLP Libraries were tested in order to find the one that full fill the needed requirements for the project. Some of those are: Python NLTK, Stanford, Google Analytics and OntoText. Data and API was generated but due to text challenges was not possible to implement any of the above mentioned tools.

3.5 Interpretation and Evaluation

The evaluation of the models was based on the review and interpretation of the machine learning applied models including the statistical analysis. For the statistical analysis, SPSS output was the first analysis to be carried out. For the selected models Precision, Recall and Accuracy was identified for Decision Trees and Support and Confidence for ARM

4 Implementation

Following the methodology part, this section explains in detail how the described techniques were implemented. Therefore, a classification modelling approach was used to address the research. The implementation section is also supported by previous research on the field of text mining and news in media described in the Literature Review and Related Work part. The following set of steps describe the projects implementation diagram showed in Figure2

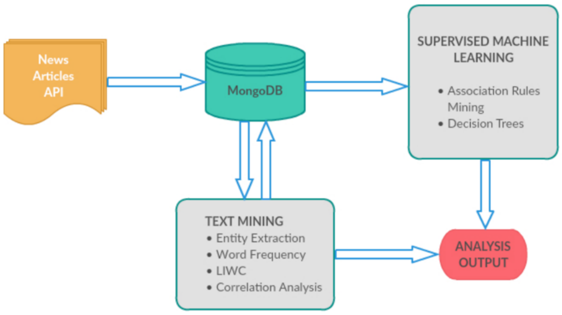


Figure 2: Implementation Diagram

News Articles The required data for the project was a selection of an up to date database body of news articles that provided title, body and URL. For this, there were several news articles databases named. Nonetheless, the requirement of the dataset to be collected was, a content of mixed newspapers of English speaking countries, with a specific topic in common as it was Brexit, in a determined time frame.

In order to avoid biases in the selection of the newspapers, and due to target dataset was expected to have highly political content, the chosen database was Aylien Text Analytics. It provides an Application Programming Interface (API) that query the news from different sources all over the world.

Aylien provides a NLP package of information retrieval and machine learning tools for extracting meaning and insights from textual and visual content data. Also a source of news in real time which was the database used for the collection of the unstructured data through the Aylien API by:

- Topic title: Brexit
- Language: English
- Sorted by: Hotness
- Since: 11 months until now
- Entities: Immigration, Trump, Islam, Racism, Immigrant

Then, the formation was queried from the Python Software Development Key (SDK) which is the Python client for Aylien, allowing to get the data through Python. By automating the process in Python, was possible to retrieve the articles with body, title, URL and sentiment positive, negative and neutral. Then, the output was stored in the chosen NoSQL Database MongoDB. As mentioned, the collected data was unstructured. At this stage the data was cleaned and stored in a NoSQL database. For being a text data driven approach, the articles were stored in MongoDB, which is a NoSQL document oriented database available to use with JSON like documents, easy to get, use and store. The data was stored by: id, URL, sentiment, title, body and date of each article. The use of MongoDB Compass allows to visualise the data in a structured way.

The main part of the data cleaning took place in Python. After that, there was a corpus of text data of 581 articles with the characteristics and requirements needed for the analysis.

Text Mining This phase presents a way for finding information in a collection of indexed documents. (Mahgoub and Rösner; 2006) As required for the research problem, the extraction of the entities in the news articles was done by MonkeyLearn. It is a Machine Learning Software Platform on the cloud for Text Analysis that provides an API. The Entity Extractor offers the service of Named Entity Recognition (NER) for data in English and Spanish, labelling three classes: PERSON, ORGANIZATION and LOCATION. It is important to mention that the output of MonkeyLearn also includes the word count frequency by article. The software platform was implemented using R programming language.

The Selection of Entities for the analysis was done in the following steps:

a) Selection of the top 50 more relevant entities in the corpus of news articles through a visualization tool using word-cloud, that works with n-grams (1- gram and 2- gram) by count of frequency of the entities. For this, the software Orange was implemented.

b) The following step was to get NER, which was done using MonkeyLearn to the 50 more relevant entities selected from the word-cloud. The output of NER in R, also included a tag indicating when the entity was a person, an organisation or a location with the respective word frequency count.

An important role of the text mining was carried out by gathering the Linguistic Inquiry and Word Count (LIWC). The output of the program is compromised by 82 variables, which includes raw word count, words per sentence, analytical thinking and emotional tone between many others. The body of the articles was analysed by LIWC and output was a large dataset of variables (analysed words) in the provided body including the word count. For this research, the selection of the relevant variables was compromised of 13 variables, which are: I, we, you, past, present, future, positive emotion, negative emotion, anxiety, anger, work, money and religion. The output from LIWC, (Tausczik and Pennebaker; 2010) is a transparent text analysis program that counts words in psychologically meaningful categories. As a result, LIWC data was used for two different text analysis as it will be described.

The 13 selected variables from LIWC were joined to the output of the 50 entities and word count from MonkeyLearn. Subsequently, at this stage there was a new dataset with a narrowed and more specific scope. This was made in order to get data for a different type of analysis. A relevance analysis was conducted to the selected entities joined from the word frequency and LIWC numerical dataset. In order to discover whether there were entities in the articles with high significant relationship between them. There are clear and compelling scientific reasons to be interested in the frequency with which linguistic forms are used, in other words, statistics, as one approaches the study of language. According to (Schuetze; 1997). The analysis was conducted using the IBM Statistical Package for Social Science software (SPSS). In order to understand the strength and direction between the selected entities a Pearson Correlation Analysis was carried out. The strength of the relationship ranked from -1 to +1. After this, there was a significant reduction in the variables by selection, than afterwards were used for association rules.

Also LIWC data was transformed to categorical data by a rule based approach: Median and Mean of values. According to(Schuetze; 1997), the mean is simply the average offset. If the value was greater than the value of mean/median, then variable was True, if not, then False. For the MonkeyLearn variables, if the word was present in the article was True if not, then, False. This was done in order to get another dataset for the Machine Learning section.

Supervised Machine Learning In order to test the hypothesis, supervised machine learning models were implemented:

Association Rules Mining: The corpus of variables was loaded from MongoDB to RapidMiner. At this stage the loaded data was categorical as mentioned in the methodology section. Once data was loaded to RapidMiner, the selected algorithm to work with was FP-Growth which shows to perform better and faster than Apriory Algorithm by reading the files twice and not several times as Apriory does, according to the literature review section. The evaluation of this model is done by support and confidence.

Decision Trees: in accordance with the dataset, the variables selected to predict the negative sentiment tonality in the articles showed a highly correlation with the sentiment anger in previous analysis as correlation analysis and association rules mining. Also, anger was one of the most prominent of the negative emotions, gathered from the LIWC output dataset, therefore, it was chosen Anger as the negative tonality of prediction. In this context, the evaluation of the model was based on the support, confidence and recall of the model.

5 Evaluation

In order to test the hypothesis, there were two main analysis conducted in the project. This section presents the findings of the implemented methodology, and further, a discussion on the results.

5.1 Association Rules Model

ARM gave the output of relationships between entities and the LIWC analysis. As rule based model, the performance can be measured by the support and confidence value. Figure3shows associations between the entities with higher occurrence in the model. Based on that, the formed rules are presented in the Figure3

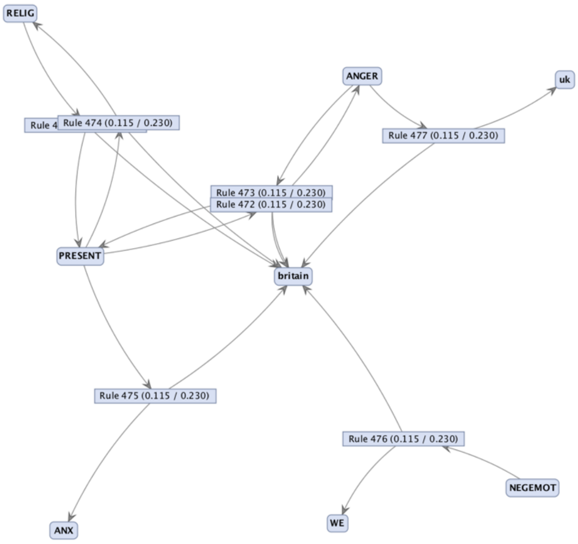
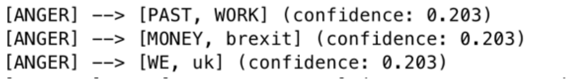


Figure 3: ARM Britain

The figure shows the rules that the model created with the minimum confidence at 0,1. The confidence of the rule in the output measures the reliability of the rule, in other words, how frequent the entity appears in the transaction. Therefore, the selected rules are the ones with a higher confidence. As the graph shows, Britain is located in the middle of the graph surrounded by negative sentiment tonality such as Anxiety, Anger and Negative Emotion. It is also important to consider that WE and PRESENT are in between the rules and relationship with Britain. Also, Even though Brexit plays the main role in the articles, it does not appear in the graph with any direct string rule and relationship.



```
[ANGER] --> [PAST, WORK] (confidence: 0.203)
[ANGER] --> [MONEY, brexit] (confidence: 0.203)
[ANGER] --> [WE, uk] (confidence: 0.203)
```

Figure 4: Rules - Anger

Following the output of the rules and confidence, Anger is related to money, Brexit and UK as Figure4shows.

The next Figure6 shows the relationship in the main entities. In this example the confidence level of the model was modified to 0,5 in order to show a small scope of the strongest correlated rules. As a result, again Brexit does not appears included in the graph, however EU and Britain show strong relationship with anger.

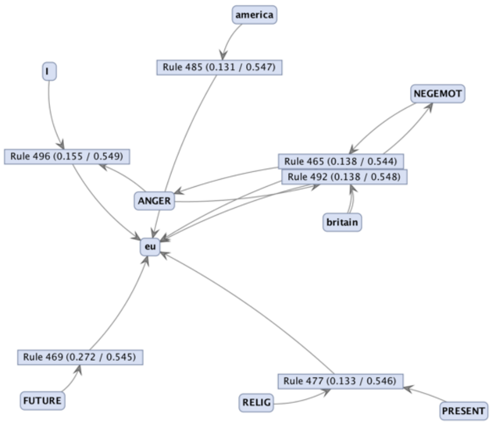
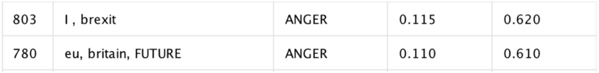


Figure 5: Anger - Strong Correlation

This rules with a modified confidence level shows the relationship between Brexit highly related to Anger with a confidence of 0.620 and a support of 0.115. Figure5



803	I , brexit	ANGER	0.115	0.620
780	eu, britain, FUTURE	ANGER	0.110	0.610

Figure 6: Rule - Strong Correlation

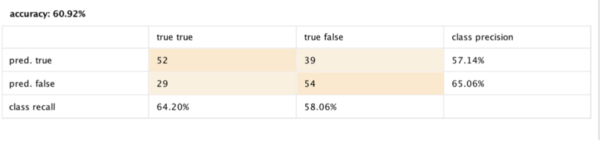
5.2 Decision Trees model

The second model implemented from the machine learning techniques was decision trees. This, due in order to predict negative sentiment tonality in the articles as the hypothesis two states. The chosen variable to predict the negative sentiment tonality was Anger.

The selection was made because anger showed to have predominance over the other negative sentiments.

After the implementation of decision trees, the following Confusion Matrix was collected (Alvarez; 2002). precision, is the fraction of the items retrieved by the system that are interesting to the user, and recall, the fraction of the items of interest to the user that are retrieved by the system.

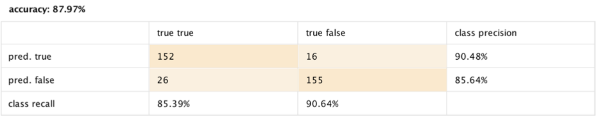
As Figure7 shoes, the evaluation of the model with a dataset of 581 news articles gives an accuracy of 60.92. In contrast, Figure8, shows an accuracy of 87.87 percent. The difference in the accuracy is due to the dataset. The model was tested several times in order to find the best accuracy and it was found out a technique that is implemented in Text Mining which doubles the data. By doing this, the model performs with higher accuracy, precision and recall.



accuracy: 60.92%

	true true	true false	class precision
pred. true	52	39	57.14%
pred. false	29	54	65.06%
class recall	64.20%	58.06%	

Figure 7: Low Accuracy



accuracy: 87.97%

	true true	true false	class precision
pred. true	152	16	90.48%
pred. false	26	155	85.64%
class recall	85.39%	90.64%	

Figure 8: High Accuracy

Even though for Text Mining the amount of data was considerable enough for the training set, in terms of ML models, as is the case of decision trees, data needed to be duplicated to increase the accuracy of the model.

5.3 Discussion

The extracted entities from the dataset have three categories: Person, Organization and Location. According to the output that ARM shows, there is a strong relationship between the entities that indicated just one type of entity which was Location. Specifically, Britain. The set of rules que ARM gives is a narrowed scope but strongly related to each other. This are: Anger, Future, America.

As the topic of research was politically charged, and the event was Brexit. It was expected to be in the strong rules and highly correlate relationships. But surprisingly, the entities Person and Organization do not played a major role in the negative tonality. Instead, the Locations are strongly correlated with anger, negative emotion and anxiety.

Arguably it could be said that the negative tonality is been put on the places rather than the people. Also that the conjugation of words most used was the negative emotion with timeframe.

It is important to point out that the words that describe timeframe such as future and past are highly related to negative emotions to anxiety and anger. Possibly and arguably because the news is related to events that have happened or are going to happen. We and I are the pronouns that show highly correlation with the negative tonality.

6 Conclusion and Future Work

This paper presented a research on Politically Charged News in Media. The hypothesis testing found relationships between negative sentiment tonality specifically anger and the main entities in the articles. This relationship is carried out specifically by the type location, which is strongly linked and positively correlated by the rules of ARM with high confidence level.

The second hypothesis was also found to be true. It was possible to detect the sentiment tonality anger in the articles by an accuracy in the machine model of 87,97 percent by using Decision Trees.

As the research is highly transferable to other domains within text mining, it can be extended in detail to analyse a broad variety of sentiments in news articles and the meaning whiting the context, also to find the psychological impact of biased news not only in politically news but also in different languages so the scope can increase and therefore the datasets.

Philosophically, this brings us close to the position adopted in the later writings of Wittgenstein (that is, Wittgenstein 1968), where the meaning of a word is defined by the circumstances of its use. Under this conception, much of Statistical NLP research directly tackles questions of meaning.

Acknowledgements

I would like to thank my Supervisor Dr. Simon Caton for his collaboration and support during the whole year. My very good friend Mayank Gulaty for his patience, help and dedication in everything (without him, would have been impossible). Of course, to my beloved family, for being my support in life, lastly all the friends that heard my complains and gave me a voice of support. Gracias totales!!

References

Alvarez, S. A. (2002). An exact analytical relation among recall, precision, and classification accuracy in information retrieval, *Boston College, Boston, Technical Report BCCS-02-01* pp. 1–22.

Barhan, A. and Shakhomirov, A. (2012). Methods for sentiment analysis of twitter messages, *12th Conference of FRUCT Association*.

Eisner, J. M. (2001). *Smoothing a probabilistic lexicon via syntactic transformations*, PhD thesis, University of Pennsylvania.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI magazine* **17**(3): 37.

Fortuna, B., Galleguillos, C. and Cristianini, N. (2009). Detection of bias in media outlets with statistical learning methods, *Text Mining* p. 27.

Fuller, C. M., Biros, D. P. and Delen, D. (2011). An investigation of data and text mining methods for real world deception detection, *Expert Systems with Applications* **38**(7): 8392–8398.

Gupta, V. and Lehal, G. S. (2009). A survey of text mining techniques and applications, *Journal of emerging technologies in web intelligence* **1**(1): 60–76.

Haselmayer, M. and Jenny, M. (n.d.). Sentiment analysis of political communication: combining a dictionary approach with crowdcoding, *Quality & Quantity* pp. 1–24.

Kotsiantis, S. B., Zaharakis, I. and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Mahgoub, H. and Rösner, D. (2006). Mining association rules from unstructured documents, *Proc. 3rd Int. Conf. on Knowledge Mining, ICKM, Prague, Czech Republic*, pp. 167–172.

Mairesse, F. and Walker, M. (2006). Words mark the nerds: Computational models of personality recognition through language, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 543–548.

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms, *Oriental Journal of Computer Science & Technology* **8**(1): 13–19.

Nisbet, R., Miner, G. and Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*, Academic Press.

Schrauwen, S. (2010). Machine learning approaches to sentiment analysis using the dutch netlog corpus, *Computational Linguistics and Psycholinguistics Research Center* .

Schuetze, H. (1997). Document information retrieval using global word co-occurrence patterns. US Patent 5,675,819.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods, *Journal of language and social psychology* **29**(1): 24–54.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B. and Daelemans, W. (2016). A dictionary-based approach to racism detection in dutch social media, *arXiv preprint arXiv:1608.08738* .

Witschel, H. F. (2005). Using decision trees and text mining techniques for extending taxonomies, *Proceedings of the Workshop on Learning and Extending Lexical Ontologies by Using Machine Learning Methods*, Citeseer.

Wu, H., Lu, Z., Pan, L., Xu, R. and Jiang, W. (2009). An improved apriori-based algorithm for association rules mining, *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, Vol. 2, IEEE, pp. 51–55.