University of Sheffield

Sentiment Detection and Tracking in Social Media Streams

Sonia Oyunga

Supervisor: Mark Hepple

A report submitted in fulfilment of the requirements
for the degree of MSc in Advanced Computer Science

in the

Department of Computer Science

May 1, 2019

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Sonia Oyunga

Date: 1st May 2019

Abstract

This report outlines the ideas, methods and techniques of sentiment analysis and describes and evaluates a sentiment analysis task designed to reproduce the results of the 2016 UK Referendum by analysing Twitter data collected on the following topic in the months leading up to the election. Chapter 1 gives a detailed structure of the report focusing on the main ideas discussed in the chapters. Chapter 2 outlines the concept and ideas of sentiment analysis. A theoretical model of the task of sentiment analysis is described. An overview of the statistical processes involved in classification is also outline in this chapter. Chapter 3 describes the planned investigation for the project. it outlines the requirements and resources available for the project. Chapter 4 describes the development of the classification system outlining the main parts of the system that implement the different techniques that are described in the process of sentiment analysis. Chapter 5 describes the experiments carried out to investigate if the analysis of Twitter data can reproduce the results seen in the 2016 UK Referendum. Chapter 6 concludes the project by highlighting the results obtained form the project and the areas that can be explored in future works.

Contents

Chapter 1

INTRODUCTION

Social media influence is growing exponentially and has become a part of our culture with platforms such as Facebook, Instagram and Twitter playing an increasingly important role in our lives. The information shared on these sites can provide insight on political matters, predict market and economic trends and influence the social aspects of our societies. Understanding and being able to process and analyse this information has become significant for governments, world organisations and businesses. In previous years, manual analysis of data was sufficient in the processing of user sentiments and opinions. However, the volume of information available for examination has increased significantly. The number of internet users worldwide in 2018 was 4.021 billion, an increase of 7% from 2017. The number of social media users worldwide in 2018 was 3.196 billion, an increase of 13 percent from 2017. (Kemp, S [2018]) The number of Twitter users has grown more than 10 times between 2010 and 2018. Sentiment analysis is a Natural Language Processing (NLP) task used for text classification. Natural Language Processing attempts to classify texts into subjective or objective classes and then detect the positve or negative opinions in the subjective text Al-Harbi [2019]. Following the development of machine learning techniques the field of sentiment analysis has undergone significant advancements. NLP techniques can now be used to carry out sentiment analysis tasks on the large amount of data and information shared on social media platforms.

1.1 AIMS AND OBJECTIVES

The aim of this project is to investigate whether a software system can perform a sentiment analysis task comparable to human sentiment classification This project will begin with a survey of the current trend in Sentiment Analysis and Opinion Mining. An analysis of the projects in the same field will be carried out, as well as an investigation of the different techniques; Lexicon-Based or Machine based, used to classify text data. Based off this investigation, a method for sentiment analysis will be selected and a sentiment classifier will be designed to carry out a sentiment analysis and opinion mining task. The system performance will be tested and evaluated. The classifier will carry out an analysis on a set of labelled data and the accuracy of the system will be examined. The classifier will then be

used to determine the sentiments expressed on unlabelled Twitter data concerning the 2016 UK Referendum and the results will be compared to opinion poll data.

1.2 OVERVIEW OF REPORT

This report outlines the research and development of a Sentiment Analysis/Opinion Mining task. Chapter 2 discusses Natural Language Processing processes, in particular, Sentiment Analysis and opinion mining. The chapter describes the different techniques of Sentiment Analysis; Lexicon Based techniques, distinguishing between a binary and gradable approach. Corpus Based/Machine Learning techniques are discussed with a description of the machine learning algorithms that are commonly used for sentiment analysis. A review of previous studies in the same area is discussed. A study by O'Connor et al. [2010] is described. They investigated the correlation between Twitter messages and public opinion polls. A description of the trends and position in conventional sentiment analysis research is presented. The considerations one makes when determining how to develop an autonomous sentiment classifier are highlighted as well. Chapter 3 describes the planned investigation for the project. The aim of the project is explicitly stated and the requirements for the completion of the project are highlighted. An analysis of the planned investigation is described, explaining the factors considered when deciding on the implementation chosen for the system. A method of analysing the system is described with equations and metrics described that will be used to investigate the effectiveness of the system. Chapter 4 describes how the system was developed. The chosen language and resources used for the development of the classifier is described. The architecture of the system is described for both the training stage and the testing stage.Chapter 5 describes the experiments carried out on the system and presents the results obtained. The results of the test are also analysed. Chapter 6 discusses and summarises the research project. Key findings are presented and a conclusion is drawn. Future work for is also discussed.

Chapter 2

LITERATURE SURVEY

2.1 SENTIMENT ANALYSIS AND OPINION MINING

Natural Language Processing has had a primary focus on facilitating information access (Information Retrieval) rather than analysing information (Information Extraction). Information retrieval involves providing the right information to the right person when it is required. It deals with developing systems for retrieving relevant documents from text collections. The task involves deciding which documents are relevant based on a query of two or three words. Information Extraction finds specific information in documents, to be handled in further automated processes. Sentiment analysis and opinion mining goes beyond information retrieval with the aim of helping the user analyse the data for reasons such as facilitating decision making, discover patterns, trends and outliers etc. Gaizauskas, R. [2018]

2.2 APPROACHES TO SENTIMENT ANALYSIS AND OPINION MINING

2.2.1 LEXICON BASED APPROACH

A Lexicon Based approach uses a dictionary of words (lexica) with associated values for different opinion/emotion words like: good, bad, nice, terrible etc Gaizauskas, R. [2018]. For subjective sentences the number of positive and negative words is counted and used to determine the polarity of the whole sentence. This analysis can be done at various levels i.e. on a sentence or an entire document or in relation to a particular feature of the subject of the statement. Liu and Zhang [2012]

BINARY LEXICON

In a simple Binary Lexicon based approach positive and negative words are compared such that, more negative words or phrases would mean the statements is negative, while more positive words or phrases would mean the statement is positive and if there are equal the

text is classied as neutral. Based on this a piece of text is given a score of either -1 for negative statements, 0 for neutral statements or +1 for positive statements.

GRADABLE LEXICON

In a gradable lexicon approach, each descriptive word and its intensiers or diminishers are assigned a numerical value with negative words being assigned negative integers and positive words being assigned positive integers. These numbers are then added up to determine the nal emotional weight of a statements. Intensifiers are words that increase the weights of the adjective in the meaning of a sentence. For example, in the sentence I am feeling very good the word very intensifies the effect of the word good in the sentence. For a positive adjective the intensifiers adds to the weights of the final score while if the adjective is negative the weight of the intensifier is subtracted from the final score. Additional rules are applied to this approach: The negation rule checks for the presence of a negation word close to the adjective, in which case, the sign of the integer value assigned to the adjective is inverted and some addition arithmetic may be applied to increase or decrease the emotional weight of the text. For example, I am not good today, the value of good will be decreased. The capitalization rule increases a positive weight or decreases a negative weight. For example, I am GOOD today would have a higher positive score than I am good today. The reverse works for negative statements. Gaizauskas, R. [2018]

Other rules include exclamation rules and emoticons rules which also add to the nal weight of the statement. Liu and Zhang [2012] Some of the advantages of a lexicon based approach is that it works eectively with a wide variety of texts. It is language independent because lexicons can be built in any language. The approach does not require extensive pre-processing steps like training processes. It is also possible to extend the lexica to include new words as the use of language evolves which often happens especially on social media sites. Gaizauskas, R. [2018]. A draw back of this approach is that the process of building an extensive, up-to date, accurate lexicon is tedious. Also, this method would be unable to deal with user inaccuracy like spelling mistakes.

2.2.2 MACHINE LEARNING

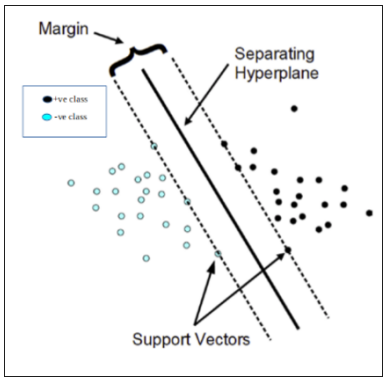
Machine learning is the science of how computers learn without being explicitly programmed. Machine learning developed by incorporating computational skills and Math and Statistics knowledge. A machine learning approach to sentiment analysis can be divided into supervised, unsupervised and semi-supervised. Supervised methods use a large amount of labelled training data. Unsupervised learning is carried out without the use of labelled training data. Semi-supervised learning uses a small amount of labelled data and a large amount of unlabelled data for training. Dorothy and Rajini [2016] Corpus-based analysis is a supervised machine learning approach. Corpus-based analysis uses a collection of text segments, a corpora, of examples containing human annotated emotional indicators, to train the classier, using a learning algorithm. A corpus-based approach allows for a higher volumes of data

that to be analysed automatically once the training stage has been completed. The results of this method are also more accurate and reliable. A major drawback to this approach is that there are pre-processing steps required before the analysis can be carried out which are often extensive and tedious for instance, preparing the data into program readable form. Gaizauskas, R. [2018]. Classication algorithms are intended to identify which categories an object belongs to. There are several machine learning algorithms used to carry out classication. In text classication, a Naive Bayes classier is often used.

Various classification methods like Support Vector Machines (SVM), Nave Bayes (NB) and maximum entropy (ME) are used for sentiment classification. Other classication algorithms include: Random Forest this method handles categorical features. It captures non-linearity and feature extraction.

SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is a non-linear model Linear Regression. Support Vector Machines are used to classify given data linearly where the data is linearly separable. During training, the system tries to determine the most suitable hyperplane that divides the labelled data into the respective spaces. The Figure below is a diagramatic representation of a support vector classification.Fletcher [2009]



Where there are multiple hyperplanes that can be used to classify the data, the maximum margin hyperplane is used. The margin is the perpendicular distance from the hyperplane to the closest point on either side of the of the plane. The best hyperplane is that with the smallest margin.

NAIVE BAYES

The Nave Bayes Classifier is used in many text classification tasks due to its high performance and speed. Nave Bayes determines the probability of an occurrence belonging to a particular class. In a piece of text, the algorithm assumes that each piece of text contributes to the overall probability independently. The Naive Bayes Classier is based on Bayes theorem. Bayes Theorem gives the conditional probability of an event A given another event B has occurred.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where: P(A/B) = Conditional Probability of A given B P(B/A) = Conditional Probability of B given A P(A) = Probability of event A P(B) = Probability of event B

2.2.3 HYBRID APPROACH

A hybrid approach combining the machine learning and the dictionary-based approaches may be used for sentiment analysis. It employs the lexicon-based approach for sentiment scoring followed by training a classifier to assign polarity to the entities in the newly found reviews. Hybrid approach is generally used since it achieves the best of both techniques, has high accuracy from a powerful supervised learning algorithm and stability from lexicon-based approach.

2.3 REVIEW OF PREVIOUS STUDIES

2.3.1 SENTIMENT ANALYSIS OF TWITTER DATA USING A LEXICON

O'Connor et al. [2010] used a simple sentiment detector based on Twitter data to replicate consumer confidence and presidential job approval polls during the 2008 to 2009 election period in the US. They connected measures of public opinion measured from polls with sentiment measured from text. They analysed several surveys on consumer condence and presidential job approval polls in the U.S and found that they correlate to sentiment word frequencies in contemporary Twitter messages. In their study, they used a lexicon-based approach. The task was divided into two: Message Retrieval and Opinion Estimation. In Message Retrieval they used topic keywords which were manually specied for each poll. For consumer condence they use the words economy, job and jobs. For presidential approval they used obama For the US elections they used obama and mccain. While collecting the data they found that the data produced would increase based on the news of the day, and in general increased towards the end of 2008, as the elections were approaching. In this task, the method used for opinion estimation was a simple binary lexicon based approach. They counted the number of positive and negative words in a piece of text to determine if the message was positive or negative. If a positive word was found the message was said to be positive, and if a negative word was found the message was said to be negative. This allowed for messages to be both positive and negative. The results were similar to comparing the number of positive and negative words because the tweets were short. They used a lexicon, from OpinionFinder, to determine the polarity or the words. Upon evaluation, they found that the system had many falsely detected sentiments. One of the main reasons

for this was the nature of natural language use on Twitter, where people often break the grammatical rules leading to inaccuracies when working with a lexicon. In conclusion, they found that a relatively simple sentiment detector based on Twitter data replicates consumer confidence and presidential job approval polls. While the results had some inaccuracies, it showed that expensive and time-intensive polling can be supplemented or supplanted with the simple-to-gather text data that is generated from online social networking. There results showed that more advanced NLP techniques that can be used to improve opinion estimation may be very useful. They also concluded that there are a number of considerations for future such as more well-suited lexicons and considerations for the modes of communication.

2.3.2 SENTIMENT ANALYSIS OF REVIEWS USING A LEXICON

Hardeniya and Borikar [2016] proposed an approach to sentiment analysis using dictionaries. They classified reviews from Amazon incorporating SentiWordNet and WordNet to find the proper word from the dictionary and assign sentiment polarity. The reviews were classified into positive, negative and neutral and they applied fuzzy logic to handle negations.SentiWordNet is a dictionary used to determine the polarity of a word. It assigns three scores to the word, positive neutral and negative. The score is calculated by a semi-supervised method. Hardeniya and Borikar [2016]

2.3.3 SENTIMENT ANALYSIS USING A HYBRID APPROACH

Cao and Zukerman describe a probabilistic approach that combines information obtained from a lexicon with information obtained from a Nave Bayes classifier for multiway analysis. They evaluated the performance of the combined system. They examined the performance of three methods based on supervised machine learning applied to multi-way analysis. The used seven datasets of different sizes, review lengths and writing styles. They concluded that the hybrid of the lexicon and corpus based system performs at least as well as the state-of-the-art systems.

2.3.4 SENTIMENT ANALYSIS USING PARTS OF SPEECH

Taboada et al. [2011] presented a word-based method for extracting sentiment from texts by extending the Semantic Orientation CALculator (SO-CAL) to other parts of speech including adverbs verbs and nouns in addition to the conventional adjectives that had previously been used. They made use of intensifiers for pairs of words and developed a more extensive method for negation of statements. The results represent a statistically significant improvement over previous instantiations of the SO-CAL system. They showed, as well, that constructing a dictionary manually provides an essential foundation to maximize the performance of a system like SO-CAL.

2.4 CLASSIFICATION MODELS

2.4.1 FEATURE EXTRACTION

Feature extraction is important for all clustering algorithms. Features are extracted from the opinionated text which will be used to train and test the classifier. Aggarwal and Zhai [2012] For example, if the system was to be used to determine the gender of a specific species on animal, the length, height and weight of the animal could be used as features to analyse the characteristics of the different classes of the animals. With textual data, it is difficult to determine which features will be used to determine the class to which the text belongs to. The biggest determinant of the sentiment expressed in a piece of text is the words used. Analysing these would be just as efficient as if a Lexicon was used to carry out the analysis. Furthermore, extracting the features from text, and in particular tweets, would be infeasible because there is no standard structure followed by twitter users. For example, a statement such as Regarding the EU referendum, I definitely think if would be best if the UK remained in the EU contains most of the features described by Bing Lui when discussing his model for feature extraction as well as the different parts of speech that could be used in speech tagging. However, a statement such as EU, no is much less descriptive. While it is relatively easy for a human to understand the sentiment expressed in the text, it is far more difficult for a system to arrive at a conclusion.

BING LUI'S MODEL FOR FEATURE EXTRACTION

Bing Liu describes an opinion in terms of a quintuple set with values distinguishing: the target object of the text, the feature of the object, the sentiment value of the opinion (this can be about the target object as a whole or about a feature of the target object) the opinion holder the time when he opinion about the feature on the object was expressed. (Liu and Zhang [2012])

Liu and Zhang [2012]give an illustrative example with a review segment on iPhone.

(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop ...

The text has been broken down into its component sentences. The rst observation is that there are several opinions expressed in the review. Sentences (2), (3), and (4) express some positive opinions while sentences (5) and (6) express negative opinions or emotions. The opinions are expressed about dierent target objects. Sentence (2) refers to the iPhone as a whole, sentence (3) refers to the touch screen, sentence (4) refers to the voice quality, sentence (6) refers to the price. In Sentence (5), an opinion is expressed but it neither refers to the iPhone or any of its features, it refers to the author of the review. This review has two opinion holders. The opinions expressed in sentences (2), (3), and (4) are by the author of the review while those expressed in sentences (5) and (6) are by the authors mother. Using

this model, as seen in the example, the task becomes to discover the quintuple set of values in a piece of text.

2.4.2 BAG OF WORDS

This is a widely used word based representation for text mining. The text is represented as a bag of words which accounts for the frequency of the words but the order of the words is lost. The model results in a vector representation that can be analysed and processed using machine learning, for example dimension reduction.(Aggarwal and Zhai [2012])

2.4.3 PARTS OF SPEECH

This is the marking of words in a text corpus. Part-of-Speech is used to disambiguate sense which in turn is used to guide feature selection. In Part-Of-Speech tagging each term in sentences will be assigned a label, which represents its position in the semantics of the sentence. In part-of-speech tagging, words like adjectives and adverbs are identified. (Vohra and Teraiya [2013])

2.4.4 LEXICON FOR SENTIMENT ANALYSIS

A comprehensive and high quality sentiment lexicon is essential to most sentiment analysis applications. They are necessary for the sentiment analysis when no training data is available, (making supervised learning unfeasible) and also for improving the effectiveness of any machine learning approach being used for sentiment analysis by providing high quality sentiment features.(Lu et al. [2011]). Following the observation that different words carry different meaning in different domains Lu et al. [2011] carried out a study to construct a domain specific lexicon. This was able to handle the complexity of different semantic meaning of word depending on what the adjective was addressing.

2.5 PRE-PROCESSING

For sentiment analysis, data is converted from unstructured to a structured form before sentiment polarity calculation. The process involves selecting the words from the sentence which will be used to calculate the polarity of the sentence. In general, the process involves: - Extracting the words in front of the product name (adjectives) from the text. - The stop words from these extracted words are used. Stop words are the most common words in the English language but are of no use in the polarity calculation. They are, for example words like an a and the. - Stemming is the process of reducing derived forms of a word to its root. An additional lexicon can be used for this process. For example, rain rained and raining all have the same root

Additional pre-processing tasks include: - Parts of speech tagging where the words are assigned parts of speech such as noun, verb, adjective and adverb. - Correction of spelling

errors and converting the whole text to either lowercase or uppercase depending on the form used in the lexicon.(Hardeniya and Borikar [2016])

2.5.1 PRE-PROCESSING OF SOCIAL MEDIA DATA

(Farzindar, Atefeh and Inkpen, Diana [2015])Social media data is the collection of Open Source information that can be publicly obtained. Among the key properties of this data, the properties that need to be considered when carrying out Linguistic Pre-processing are: - The data is non structured text in many formats - The data is written by any people in many languages - The data is written in everyday (informal) language - The authors are not professional writers Further pre-processing tasks need to be conducted on social media data in addition to the standard pre-processing tasks before sentiment analysis can be carried out. - Re-training Natural Language Processing tools for social media text - Language Identification

2.6 SUBJECTIVITY

Statements expressed in Natural Language can either be subjective or objective. Objective statements refer to factual events for example I watched the new lm today Subjective statements refer to opinions about a particular thing, person or event for example I enjoyed the lm I watched today. Sentiment analysis focuses on subjective statements therefore subjectivity classication is often the rst step in sentiment analysis. A simple Binary Lexicon based approach using a lexicon of emotion words can be implemented to determine the subjectivity of a statement. The lexicon would be used to determine the subjectivity of a statement by counting the number of emotion words used in the statement. A reference number, n, is selected such that if the total number of emotion words is above n the statement would be classied as subjective and below n would be classied as objective. The task of Sentiment Analysis is further complicated by the fact that subjective sentences do not always express positive or negative opinions for example I think it will rain today. The use of the word think would normally indicate that a person is expressing their opinion however in this sentence the word think is used to show uncertainty about a fact. In addition, objective statements can express opinion indirectly for example, I cannot believe it has been raining all day This statement is referring to a fact however there is some sentiment expressed in it. (Gaizauskas, R. [2018])

2.7 CONSIDERATIONS

2.7.1 CHARACTERISTICS OF NATURAL LANGUAGE

In Sentiment Analysis, text is usually expressed in Natural Language and so is subject to various dynamic characteristics observed in Natural Language, for example use of Co-referencing and (synonym) resolution. For example, in the statement I tried the new signature meal at the new restaurant on Parkway yesterday. The food was so good. It was also very lling,

the food and it all refer to the signature meal. This process is something human beings do intuitively, however computationally the task is dicult. Some other, more subtle challenges for Sentiment Analysis is the use of exaggeration and irony. In informal settings, such as on social media, people often use these stylistic devices in their posts and reviews.(Gaizauskas, R. [2018])

The main issues in sentiment analysis are negation handling and domain dependency. Negation words are the words which reverses the polarity of sentence if occur in a sentence. Domain dependency is there because the word has positive orientation in one domain and the same word has negative orientation in different domain. It is most important to handle this issue for correct classification of reviews.

2.7.2 CONTEXT AND DOMAIN DEPENDANCY

Sentiment classification has to deal with the problem of context dependant orientations, for example where words like small could be a positive in terms of power consumption of a phone however could be negative when referring to the size of the screen on a device. A classifier trained to perform classification on a specific topic may not perform as accurately when applied to a different domain. For Lexicon Based system, the lexicon may be too general to provide an accurate result for more specific topics.

Another consideration the technique needs to consider is negation of negative words which would result in a positive opinion but using the counting method of analysing a statement would result in the opposite conclusion. [2012].

2.8 SUMMARY

This chapter introduces the field of Sentiment Analysis. It describes the methods and techniques that are used to determine the opinions expressed in a statement. Lexicon-Based and Corpus Based techniques are discussed.Some conventional practices like feature extraction and pre-processing stages are highlighted and described and related to the process of sentiment analysis.

Chapter 3

PLANNED INVESTIGATION

The aim of this research project is to design, develop, implement and test a sentiment classifier to analyse Twitter data regarding the 2016 UK Referendum. An investigation will be carried out to determine whether the results obtained from the system are comparable to the manual polls collected regarding the election in the months leading up to the vote and if the results from the system displayed are a reflection of the final result of the referendum.

3.1 REQUIREMENTS AND RESOURCES

3.1.1 LABELLED DATA

The training data used for this project was obtained from (Go et al. [2009]) who carried out an investigation to automatically classifying the sentiment of Twitter messages into using tweets with emoticons for distant supervised learning. The data is manually collected, using the Twitter web application. A set of 177 negative tweets and 182 positive tweets were manually marked. They searched the Twitter API using queries to retrieve tweets on different subjects. The tweets consisted of consumer products, companies and public figures. They looked at all the collected tweets and marked them as positive and negative independent of whether the tweets had emoticons. This data set was selected for the purpose of training the system because it was manually marked and so could be used to investigate how well the system performs. The data was stored as a CVS file with the emoticons removed. The data was arranged in six fields. The polarity of the tweet; 0 for negative tweets, 2 for neutral tweets and 4 for positive. The ID of the tweet. The date the tweet was posted. The query used to find the tweet from the Twitter database. The user that posted the tweet. The tweet.

3.1.2 UNLABELLED DATA

The test data set used in this investigation is twitter data about the UK Referendum that took place in 2016 on whether the United Kingdom should leave or remain in the EU. The data was stored in JSON format which is similar to a python dictionary with name and value fields for the different parts of information collected. It contains; tweet, hashtags, links to

articles (where present), information about the user such as their twitter id, real name and screen name. There are several challenges faced when carrying out sentiment analysis using tweets. Neutral tweets are more common than positive and negative ones. Most domains concerned with sentiment analysis do not deal with this problem for example product/movie reviews. Another challenge is that tweets are often short and often show limited sentiment cues There are linguistic representational challenges like feature engineering.(Da Silva et al. [2014])

3.1.3 LEXICON

The text is analysed using Vader for classification. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a simple rule-based model for general sentiment analysis. The VADER sentiment lexicon is gold-standard quality and has been validated by humans. To determine the classification of a piece of text VADER works similarly to OpinionFinder used by O'Connor et al. [2010]. Vader assigns a normalised positive or negative value between -1 and 1, with 0 indicating a neutral score. In the development of Vader, Hutto and Gilbert [2014] used a combination of qualitative and quantitative methods to produce VADER. They empirically validated the lexicon. They combined lexical features with a focus on five general rules that embody grammatical and syntactical conventions that are used to express intensity. VADER is bigger than traditional sentiment lexicons like LIWC but is just as simple to inspect, just as easy to understand and can be quickly implemented without the need for extensive training. In addition, VADER is more sensitive to sentiment expressions in social media domains while performing effectively in other domains. VADER can also be easily extended to include more words for increased and domain-specific effectiveness. (Hutto and Gilbert [2014])

3.2 ANALYSIS

As discussed in chapter 2, there are two methods of analysing the sentiments expressed in a piece of text. Using a lexicon, a numerical value is assigned to the different words in a piece text. Following some computation, a final score is calculated and the polarity of the text is obtained. A machine Learning approach would analyse a training dataset and determine the features that are associated with the text that has been labelled. It is then able to recognise the patterns that lead to different statements being classified as positive, negative or neutral. The system can then be used to perform the analysis on any given dataset.

While the machine learning technique is more robust there are requirements for training the system that are unobtainable for this project. To train an analysis system a labelled dataset is required. Because sentiment analysis is domain specif, a domain specific labelled data set is required. In addition, in order to analyse the patterns being displayed by the dataset, the data has to have standard features that can be extracted and analysed; feature extraction.

The method chosen for this project is a Lexicon-Based system. The tweets are subjective so in order to obtain actual opinions on the subject-matter as opposed to just the polarity of the text a Lexicon-Based system is more suitable.

3.3 EVALUATION

Investigating the performance of the system will define how effective the system is and allow tuning of the system to achieve the highest possible effectiveness. This will also aid in estimation of how well the system will perform when a different dataset is used.(Lewis et al. [1995].)

3.3.1 EVALUATION METRICS

Binary classification tasks, where a system decides whether an item belongs to a single class or not, are the simplest and most common classification task. Using a labelled dataset, where a classification has been done by a human expert who can determine the true classification, the relationship between the labels given by the system and those assigned by the expert can be investigated. (Lewis et al. [1995])

RECALL

Recall is the proportion of items that the system assigns to a particular class.

RECALL = $\frac{TCM}{TCM+FCTM}$

Where TCM is a correctly classified member of the classified and FCTM is an item that belongs to the class according to the label the expert assigned, but has been placed in a different class by the system

PRECISION

Precision is the number of items that have been correctly assigned to a particular class; items that have been classified in the same way by both the system and the expert.

PRECISION = $\frac{TCM}{TCM+FCM}$

Where TCM is a correctly classified member of the classified and FCM is an item that has been classified as a member of the class but following the label assigned by the expert is not really a member of the class.

ACCURACY

Accuracy measures the total number of correct classifications from the total dataset used for classification. A value of one would indicate that the system classified all the data points accurately and the system is perfect

ACCURACY = $\frac{CorrectlyClassifiedItems}{TotalNumberofItems}$

3.3.2 CONTEMPORARY OPINION POLLS

To evaluate the performance of the classifier on the unlabelled UK Referendum the results will be compared to political opinion polls. These will be used as a gold standard for analysing the performance of the system. The trusted opinion polls and surveys that track the EU referendum is YouGov, is an international Internet based market research and data analytics firm. It is an government run organisation and so is taken as a reliable source for comparison of the results obtained. The site also provides an analysis of their polls following the conclusion of the referendum and so explains any inaccuracies in their polls in the months leading up to the election. This could be used to explain the results of the system.

Chapter 4

SYSTEM DEVELOPMENT

4.1 CLASSIFIER DEVELOPMENT

For this project, the system was developed using Python programming Language. Python is an open source general-purpose language. Python comes with in-built modules. Modules are functions and variables defined in separate files. Python also has libraries that useful for natural language processing such as the Natural Language Processing Toolkit. Python has an easy to use interface making it easy to read and write code.(Huenerfauth et al. [2009]) Some of the modules used in this project are numpy library, gzip library, json library, re library and nltk.sentiment.vader

4.1.1 SYSTEM DESIGN

The first stage in the system involves training the classifier. The labelled training data is loaded into the system. The training data was provided by (Hutto and Gilbert [2014]), who developed the VADER model. The data has been structured for the purpose of this sentiment classification by VADER therefore pre-processing is not required. The relevant fields (features) are extracted. The features are then passed through the VADER classifier and a score between -1 and 1 is obtained. The results of the classifier are compared with the labels that had been manually assigned to the training data. Computation is then done to determine the effectiveness of the classifier.

The second stage of the system involves testing the unlabelled dataset. The unlabelled dataset was collected in semi-structured form from twitter and was stored in Json format. The data provided for this project was very large, the entire dataset contained over 11 million tweets. Processing large amounts of data was expensive and subject to computational limitations. Therefore, while there was plenty of data available for the analysis, the system was only able to process a small subset of it for use in the project. The system used a randomly generated array for the selection of a random subset of the information. The range used to generate the random array ensured that the entire data set was included in selection of the subset. This subset was stored in separate files which were then read into the classifier

for the classification task. Feature extraction was carried out to eliminate the unnecessary fields present in the data such as username, mentions and twitterID. Twitter data contains hashtags. These are words that highlight the topic being referred to in the tweet. Hashtags are a common feature used to collect data from Twitter as they are often used and easily queried. Hashtags were used as a main feature in the classifier. As discussed in chapter 2, social media data is unstructured therefore requires pre-processing. Elements such as hyperlinks and line breaks were removed from the data and special characters are encoded to standard UTF-8 format.

Chapter 5

EXPERIMENTATION

5.1 Experiments

5.1.1 EXPERIMENT 1: INVESTIGATE THE ACCURACY OF THE VADER CLASSIFIER USING A LABELLED TWITTER DATASET

Experiment 1 was conducted with the aim of investigating the accuracy of the VADER classifier. Hutto and Gilbert [2014] developed the VADER classifier and tested it against several state-of-the-art systems. Using a labelled dataset, the VADER classifier is tested to determine the accuracy and effectiveness of the system.

RESULTS

The system was tested with the labelled data and the results showed an accuracy of 73%. Of the positive tweets classified, the system achieved a precision of 0.67& and a recall of 0.87%. Of the negative tweets classified the system achieved a precision of 0.83% and a recall of 0.61%.

5.1.2 EXPERIMENT 2: SENTIMENT ANALYSIS OF TWEETS USING FEATURE EXTRACTION AND CLASSIFICATION USING THE VADER LEXICON

Experiment 2 was conducted to investigate the classification of unlabelled data.

On twitter, users sometimes post their tweets with an accompanying hashtag. The hashtag will either be before or after the tweet itself. It can indicate the topic the person is referring to in their tweet or the position they have on an issue. The way in which hashtags are used makes them an excellent feature to use when investigating the sentiment expressed in a tweet. For this dataset, tweets that supported the Leave side of the Referendum used the hashtags #voteleave and #leaveeu. Those that were in favour of Remain used #remain and #strongerin. The system checked the data for any of these sets of hashtags and classified them as either leave or remain

RESULTS

Using the hashtags to determine the sentiment expressed in the tweet was very effective. The results of the investigation indicated a 69&
31& split in favour of Leave. This was significantly close to the numbers obtained by YouGov when they investigated the split in support using online polls which indicated a 63% lead for Leave, with 23% supporting remain and 10% being undecided.

5.1.3 EXPERIMENT 3: OPINION MINING USING THE VADER CLASSIFICATION OF THE TWEETS

Experiment 3 was conducted to extract information about the users positions on the Referendum and investigate whether the results of the actual vote carried out in 2016 could be reproduced by the sentiment analysis system. Using the twitter dataset, the polarity of the tweets was investigated using the VADER lexicon. Each tweet was assigned a polarity regardless of the associated hashtag.

RESULTS

The system was able to determine the polarity of the tweets and found that there were positive, negative and neutral tweets regardless of the associated hashtag which was used as a main indicator for the position of the user on the Referendum. Of the tweets investigated in support of Remain 29% were negative, 40& were positive and 31% were neutral. Of the tweets in support of Leave, 28% were negative, 38% were positive and 36% were neutral.

5.1.4 EXPERIMENT 4: INVESTIGATE THE COMPARABILITY OF THE RESULTS OF THE SENTIMENT ANALYSIS TASK AND OPINION POLLS

Experiment 4 was conducted to compare the similarities and differences obtained from the surveys and opinion polls carried out in the months leading up to the vote and the results obtained from the classifier. The percentage split between leave and remain votes was obtained for the Twitter data set used and this was compared with the poll data obtained from YouGov leading up to the vote. The data obtained was between 16th May 2016 and 22nd June 2016. The data obtained from YouGov were not daily polls between this period however using interpolation the result should still show correlation where the data is similar. Four equations were used to process the data and display the data as graphs for comparison.

EQUATION 1 = $\frac{LEAVE}{LEAVE+REMAIN}$

EQUATION 2 = $\frac{LEAVE}{LEAVE+REMAIN+UNDETERMINED}$

EQUATION 3 = $\frac{LEAVE-REMAIN}{LEAVE+REMAIN}$

EQUATION 4 = $\frac{LEAVE-REMAIN}{LEAVE+REMAIN+UNDETERMINED}$

The equations above compute the graph of Leave votes. The equations were used interchangeably to compute the graphs of Remain.

RESULTS

The result showed the graphs were not comparable over time.

The graphs below are a representation of the results obtained from the classification done by the system and the data obtained from YouGov polling

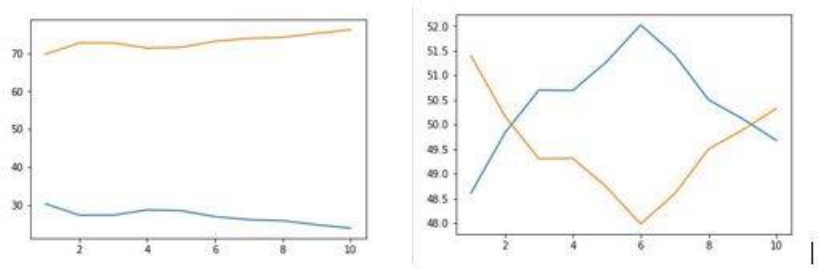


Figure 5.1: graph on left is a representation of the system analysis over time, graph on the right is a representation of YouGov data over time. Computed using Equation 1 described in section 5.1.4

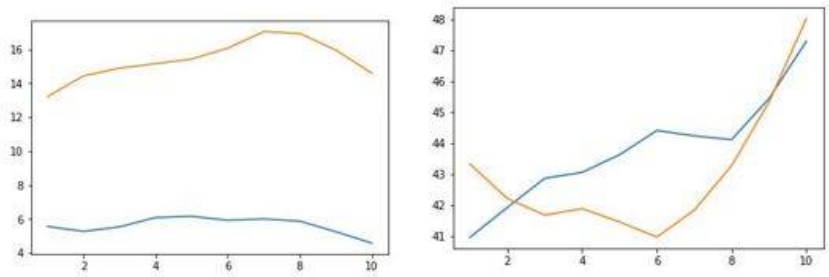


Figure 5.2: graph on left is a representation of the system analysis over time, graph on the right is a representation of YouGov data over time. Computed using Equation 2 described in section 5.1.4

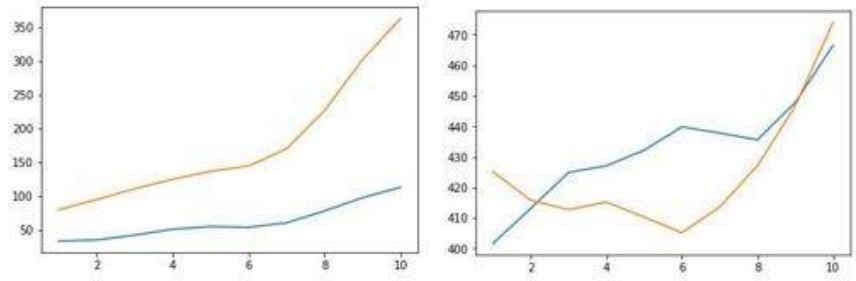


Figure 5.3: graph on left is a representation of the system analysis over time, graph on the right is a representation of YouGov data over time. Computed using Equation 3 described in section 5.1.4

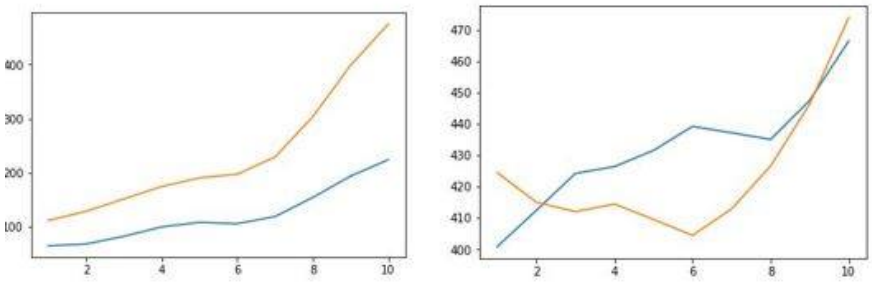


Figure 5.4: graph on left is a representation of the system analysis over time, graph on the right is a representation of YouGov data over time. Computed using Equation 4 described in section 5.1.4

5.2 ANALYSIS

5.2.1 CLASSIFIER ACCURACY

In the first experiment we found that the accuracy, 71%, was lower than that obtained by Hutto and Gilbert [2014]; 98%. The statements which were classified incorrectly contained semantics that would require a more intelligent system to determine the semantics. Statements like, lebron IS THE BOSS or, Lebron is a beast nobody in the NBA comes even close were given a score of 0.0 using the VADER lexicon. These statements do not seem to contain an opinion however a human reader can easily recognize them as positive statements. The positive labels on the tweets therefore did not match the neutral label given by the classifier. Some other statements which were misclassified were statements that expressed both a positive and negative opinions. For example, good news, just had a call from the Visa office saying everything is fine.what a relief! I am sick if scams out there! Stealing! This statement was given a score of -0.7701 by the classifier however the overall sentiment expressed is positive. The results showed that analysing a statement is harder when the

language used is informal and contains made up words and expressions.

5.2.2 FEATURE EXTRACTION

Feature extraction produced the most effective results for the classifier. Hashtags on tweets are a very accurate indicator of the sentiment being expressed by the user. The results showed a similarity between the tweets that were classified, and the online polls carried out by YouGov. It should be noted that the result obtained were not similar to other polls; telephone and in-person, and that the overall results of the Referendum were not reproduced by this classifier. This could be an indicator that online users display similar characteristics in terms of how they express their opinions

5.2.3 CLASSIFICATION USING THE VADER LEXICON

In general using the polarity of a tweet to determine sentiment in general can be effective for example when dealing with review or single arguments. For this investigation, the nature of the topic meant that negative and positive opinions cold be expressed from either side if the debate. For example, tweets like I support #VoteLeave because I want to live in a democracy again and restore freedom of speech.. is a positive statement in support of Leave. If you are under or have kids, you should #VoteLeave #EU is destroying youth unemployment is a negative statement in support of Leave. The same is true for tweets supporting Remain; #Remain The entire nation are vested benefactors of the EU is a positive statement while, Huge cohesive case been written for #Remain but blind as a bo selecting bat. No wonder #StrongerIn is a negative statement. This made classification based on polarity ineffective and inaccurate.

5.2.4 COMPARABILITY WITH OPINION POLLS OVER TIME

In experiment 4, the results showed that the classification of tweets over time by the system was inaccurate compared to the online polls taken over time. There are several reasons why there may have been a difference. The polls collected by YouGov included different methods of carrying out the polls and some may not reflect in comparison to tweets. Another reason why the tweets may not match is user demographics. Different groups of people may be more vocal on different platforms and surveys. The users on twitter may no participate in other surveys and polls like telephone and in-person surveys, and vice versa. Thus the results may not represent the general view of the public as a whole, rather a specific sub-group.

Chapter 6

CONCLUSION

Text mining applications have become increasingly important in recent years with the rise of web enabled applications that lead to the creation of text-based data. This area has been explored by many communities such as data mining, machine learning and inofrmation retreival. (Aggarwal and Zhai [2012])

6.1 OBJECTIVES ACHIEVED

A lexicon based approach has been developed for this research project to investigate opinion mining and sentiment analysis of Twitter data concerning the UK Referendum in 2016. Different standard linguistic approaches have been used to carry out the task. The project aimed to reproduce the results obtained in the election and determine if contemporary social media platform can be used to predict the outcome of such events. The system classifier that was developed was able to reproduce some of the results obtained from opinion poll data. The results showed a correlation between the tweets that were analysed and the online opinion polls collected by YouGov.

6.1.1 OBJECTIVES NOT ACHIEVED

The system was not able to predict the overall outcome of the election.

6.2 KEY FINDINGS

Following the development and testing of the classifier, results show that using feature extraction, contextualized for the domain being investigated, improves the performance of the classifier. In addition, for sentiment analysis, a domain specific lexicon would increase the accuracy and effectiveness of a classifier.

6.3 FURTHER WORK

In the field of Natural Language Processing there are general trends and directions in the development of the area.

6.3.1 DOMAIN ADAPTATION

Text mining tasks using supervised learning require large amounts of data to work effectively. Creating, processing and storing large amounts of data is expensive and tedious. Domain adaptation and transferring learning to related domains when performing related tasks to make it easier work with different domains of data. Development of effective methods for domain adaptation and transfer learning methods is necessary for text mining. (Aggarwal and Zhai [2012])

6.3.2 EXTENDING MODELS FOR NATURAL LANGUAGE PROCESSING

Currently, textual analysis processes rely on the bag-of-words model. Information extraction methods use supervised learning methods and require training data in order to perform sufficiently. More robust techniques are necessary to scale up the task that can be carried out to perform classification tasks independent of the domain. (Aggarwal and Zhai [2012])

For this project additional development concerns development of a domain specific Lexicon Social Media Data is context specific and the text is domain dependant. Using a general lexicon decrease the effectiveness of the classifier. A domain specific lexica would allow us to extract more information and determine more informative sentiments from the data. For example, if the Lexicon was able expand the meaning of leave or remain a more effective classifier could be developed. In addition a subjectivity classification could be carried out when collecting the data. The data collected for the project was collected based on the mention of specific key words, hashtags. Any information that contained the key words was collected such that the tweets were not always subjective in nature. Some tweets were links to articles and news stories. Some tweets were objective statements such quotes from politicians about the Referendum. A subjectivity test can be carried out while collecting the data to ensure only relevant data is collected thus improving the efficiency of the classifier.

Bibliography

Farzindar, Atefeh and Inkpen, Diana. Tutorial applications of social media text analysis. `http://www.emnlp2015.org/tutorials/3/3`$_{O}$*ptionalAttachment.pdf*, 2015. *Accessed* : $2019 - 4 - 10$.

Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

Omar Al-Harbi. A comparative study of feature selection methods for dialectal arabic sentiment classification using support vector machine. *arXiv preprint arXiv:1902.06242*, 2019.

Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.

M Dorothy and S Rajini. The various approaches for sentiment analysis: A survey, 2016.

Tristan Fletcher. Support vector machines explained. *Tutorial paper*, 2009.

Gaizauskas, R. Text processing, 2018. Accessed: 2018-10-30.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12):2009, 2009.

Tanvi Hardeniya and DA Borikar. An approach to sentiment analysis using lexicons with comparative analysis of different techniques. *IOSR Journal of Computer Engineering (IOSRJCE)*, 18(3):53–57, 2016.

Matt Huenerfauth, Guido van Rossum, and Richard P Muller. Introduction to python. *Harvard University, Oct*, 2009.

Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

Kemp, S. Digital in 2018: Worlds internet users pass the 4 billion mark. `https://wearesocial.com/blog/2018/01/global-digital-report-2018`, 2018. Accessed: 2019-4-10.

David D Lewis et al. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, volume 95, pages 246–254. Citeseer, 1995.

Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

SM Vohra and JB Teraiya. A comparative study of sentiment analysis techniques. *Journal JIKRCE*, 2(2):313–317, 2013.