Faculty of Electrical Engineering, Mathematics and Computer Science
Network Architectures and Services

Estimating Popularity by Sentiment and Polarization Classification on Social Media

K. Charalampidou
(4040422)

Committee members:
- Supervisor: Dr. C. Doerr
- Mentor: N. Blenn
- Member: Dr. Ir. F.A. Kuipers
- Member: Dr. A.C.C. Lo

July 1, 2012
M.Sc. Thesis No: PVM 2012-75

Estimating Popularity by Sentiment and Polarization Classification on Social Media

Master's Thesis in Computer Science

Network Architectures and Services Group
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Kassandra Charalampidou

July 1, 2012

Author
Kassandra Charalampidou

Title
Estimating popularity by sentiment and polarization classification on social media

MSc presentation
June 28, 2012

Graduation Committee

Dr. C. Doerr	Delft University of Technology
N. Blenn	Delft University of Technology
Dr. Ir. F.A. Kuipers	Delft University of Technology
Dr. A.C.C. Lo	Delft University of Technology

Abstract

Mass processing of social media posts has been brought to scientists' attention during the last decade. The massive growth of online social networks, like Twitter and Facebook, have created a need for determining peoples' opinions and moods through these means. This thesis constitutes a research on measuring users' sentiment upon a particular subject by analyzing their posts. Establishing an efficient sentiment measurement technique, can be used into estimating popularity of products or persons. For separating subjective from objective posts, a hybrid classifier based on the syntax analysis of texts, is proposed, performing clearly better than existing classifying tools. Moreover, a new sentiment evaluation technique for measuring the polarity and magnitude of posts' sentiment is described and tested over different social media. Results are compared to various real ratings and show that this approach can have a promising accuracy on sentiment establishment of online posts.

Contents

Preface

This thesis is the final result of a graduation project and completes the master's degree programme Computer Science of the Faculty of Electrical Engineering, Mathematics and Computer Science at Delft University of Technology.

Even though my specialization is on Parallel and Distributed systems, I was always interested in the social network analysis area and thus I immediately seeked my involvement in this field. In the Network Architectures and Services (NAS) Group, I found plenty of challenging projects, and a group of excited and helpful people to work with.

I would like to especially thank Christian Doerr and Norbert Blenn for their valuable guidance and advice, and for their positivity which kept my motivation high throughout this thesis.

Delft, The Netherlands
July 1, 2012

Chapter 1

Introduction

Online social networks have enjoyed significant growth over the past years. Online communities such as Twitter and Facebook have been swamped with active users during the last years and much attention has been given in analyzing the social behavior and opinions of users. The wide-spread popularity of online social networks and the resulting availability of data has enabled the investigation of new research questions, such as the analysis and estimation of public opinion on various subjects.

People's sentiment towards a particular matter as expressed online, can be very useful in many cases, and its classification and estimation arises to a crucial subject. Cha, Haddadi et al. [14], have shown that the volume of discussion about products in weblogs (like Twitter) can be correlated with the product's financial performance. It is also known that social network users represent the aggregate voice of millions of potential consumers, especially for products designed for the target-group of young-aged technology users. This reveals a brand new aspect that companies should take close consideration of, while this free and high-scale feedback can give them the opportunity to understand consumers' needs and take proper action.

Additionally, a lot of effort has been given to social media analysis, regarding its power at predicting real-world outcomes, during the latest years. Some of the research that has been already made has shown the information gained from social networks can be indeed used to make quantitative predictions on some specific domains. Apart from that, aggregation of opinion of a collective population may be useful to gain insights about their behavior as well as predicting future trends on technology, arts and other domains. This can be also helpful for markets when they design marketing and advertising campaigns.

A lot of studies have been done for implementing automatic techniques

that classify social media posts according to their sentiment. However, most of these approaches are based on machine learning algorithms (Naive Bayes, Maximum Entropy, SVM) using lists of common positive and negative words. For implementing such an algorithm, it is required that a human will provide a list of training words or sentences with their sentiment, in order to train the selected model according to them. The algorithm is then able to predict the sentiment of any given text, by assign it into the category of either positive or negative meaning.

This means that a whole sentence can be evaluated as positive or negative, only because of the appearance of a specific word with strong sentiment. Without examining the relation of this word to others and determining the actual meaning of the phrase, this technique often leads to wrong classification. Additionally, word's sentiment can change from text to text, depending on its subject, its writing style and it purpose, and therefore a word with positive sentiment meaning in one kind of texts can have negative sentiment in another, something which is not taken into account by such approaches. Apart from that, additional effort and time is needed for constructing lists of positive and negative words, as well as training machine learning classifiers.

Even though there are already plenty of implementations for polarity determination of texts, very few of them introduce the concept of syntax analysis in their techniques. It is, therefore, very challenging not only to use natural language processing tools in a brand new approach of this matter but also establish particular syntax patterns that would apply to texts of any topic. This way, it is possible to build a polarity classifier that can be applicable to texts of any style, length and language complexity as long as it maintains a correct syntax structure.

This thesis describes a hybrid classifier implementing the mentioned approach, which was presented at the Networking 2012 conference [12] and compares it to existing methods. It is based on the syntactic analysis of each one of the given posts, and is seeking for specific grammatical patterns that denote the existence of sentiment over a subject. Therefore, this approach is not counting on the appearance of specific words separately but rather on particular syntax structures that reveal the actual relation between these words. This way we managed to have a clearer perception of the actual meaning of the given text, while also this classifier can be used on any content with no need to be trained on any human-made lists, as other common approaches. This algorithm contributes a 40% gain on accuracy over the existing classifiers, while it reaches about 85% of correct classified posts.

Apart from determining whether a text is objective/subjective or posi-

tive/negative, it is also extremely useful to know how positive or how negative it is - in other words what is its sentiment magnitude. Applying the hybrid classifier can ensure that subjective posts are separated from objective ones and therefore applying a sentiment estimation technique on them, instead of the initial text collection, leads to more accurate results. Our approach of that aspect is based on adjectives derived by the hybrid classifier applied to all posts before. These adjectives are initially attached with a sentiment magnitude metric according to the number of times that they occur in positive and negative labeled training texts. Then we use those results to apply a sentiment magnitude value to whole texts on a specific matter and accordingly we finally extract people's sentiment on this particular subject.

Chapter 2

Related work

2.1 Sentiment polarization estimation

Sentiment analysis, i.e. the extraction of an opinion's overall polarization and strength towards a particular subject matter, is a recent research direction, and typically approached from a statistical, or machine-learning angle. Plenty techniques have been proposed during the latest years regarding this subject and almost all of them were tested on social media posts.

A number of published works either perform machine learning on a provided corpus of human-categorized positive and negative texts such as product reviews [39, 11, 36], or use a set of keywords with positive or negative connotations to classify input [16, 26] according to its sentiment. Bo Pang and al [26] and Cilibrasi and Vitanyi [16], present a variety of machine learning techniques (Naive Bayes classification, maximum entropy classification and support vector machines) in order to classify input texts to positive and negative. For implementing these approaches, a human expert has to provide a list of training words both positive and negative, according to which the training classifier will be able to predict the sentiment of any input text. However, all these classifiers work in a target-independent way, which means that sentiment is decided for each input text regardless its subject. That may lead to incorrect sentiment predictions in many cases, while no attention is given to the specificity of the target discussed. This denotes the need of a topic-based classification, which is approached in this thesis by the hybrid classifier presented later.

Apart from that, recent publications [21, 13] introduce the use of Natural Language Processing (NLP) modules, through which the syntactical analysis of any text can be performed. Having the syntax structure of a post, makes it possible to extract the meaning of the processed text and eventually derive

sentiment out of it. However, most of the existing NLP related approaches focus on classifying texts to positive, negative and neutral, according to their syntax analysis. On the other hand, the proposed approach works in two stages: First it makes a distinction between objective and subjective posts regardless of containing positive/negative words or syntax patterns, and then performs a classification based on sentiment polarity of texts. This way, the case of attaching any kind of sentiment to objective factual texts is avoided.

Wiebe and Rillof [31, 37] focus on subjective and objective sentence classifiers, with attention to extraction of subjective patterns. A large collection of subjective sentences established as a training set to an extraction pattern learning algorithm which is then able to provide patterns of opinionated expressions. These patterns denote the grammatical structures most often used for the formulation of opinions, and the relations between these structures. Such expressions are used to identify more subjective sentences, while the absence of such a pattern in a text denotes its objective content. This concept is similar to the one proposed in this paper for subjective/objective text classification, it however, does not require the use of a pattern learning algorithm and therefore no training needs to be done.

Barbarosa and Feng [11] propose a 2-step sentiment analysis classification method, which first classifies messages as subjective and objective, and further distinguishes the subjective ones as positive and negative. For the first classification, it uses a machine learning approach based on the appearance of certain Part-of-speech tags (adjectives, verbs etc), punctuation (exclamation marks) and emoticons which denote the existence of subjectivity in texts. For the positive/negative categorization, a number of polarity labels are being used, derived by a number of training texts, similar to other existing machine-learning approaches. This study introduces a distinction between subjective/objective and sentiment classification of texts, which is also followed in this thesis. It however lacks high accuracy due to the disadvantages of machine learning approach in general, which are pointed out earlier.

Yu and Hatzivassiloglou [39] among common machine-learning approaches (SVN, Naive Bayes, maximum entropy), also propose a technique of measuring sentence similarity between given input data and texts of specific polarity. This approach explores the hypothesis that opinion sentences will be more similar to other opinion sentences than to factual ones. Using a sentence similarity tool, such as SimFinder [19], it is possible to measure the overall similarity of a sentence to both opinion or factual documents. Based on theses results, a distinction of subjective posts from the overall testing test, can be achieved. Again this method, is target-independent and it is

efficient mostly when training and testing sets have the same writing style, length, context, etc.

2.2 Sentiment magnitude establishment

Even though there is no existing research work for establishing a metric over the sentiment magnitude of posts derived from social media, similar to the work presented in this thesis, there are several techniques that focus on the semantic similarity between words. DISCO [23], a tool that retrieves the similarity between two words, is based on the detection of words co-occurrences in various corpora such as Wikipedia, British national corpus etc. This tool traces pairs of co-occurred words within an interval of three words, which gradually result in an establishment of a distributional similarity among words. DISCO offers the possibility of retrieving the semantically most similar words for an input word as well as the value of the semantic similarity between two input words.

Another common approach of sentiment similarity establishment ([27, 29, 20]), is based on the path measurement between two words included in a semantic taxonomy as derived from various lexical networks of English words (such as Wordnet [9]). The main idea of this technique is that the length of the shortest path in WordNet Similarity tool [28] between two input words denotes their semantic similarity. This way all words can be related to each other by a measurement of how similar meaning they carry.

Similar sentiment similarity estimation methods, as mentioned above, were used for the purposes of this work, in order to achieve the establishment of a sentiment magnitude measure. However, the results were not always successful, mainly because of the high difficulty of determining a word (or words) with optimal positive or negative meaning in order to be used as a basis for comparisons with all other words. For this reason the proposed algorithm is based on the word occurrences in positive or negative texts rather than on co-occurrences or semantic distances with other words.

This thesis presents an unsupervised, target-dependent approach for estimating movies' popularity, following a 2-step technique of evaluating social media posts. First a classification of subjective and objective posts is performed with the use of NLP modules that extract the syntactical structure of texts and determine their actual meaning. The classified subjective posts are then used for the estimation of the overall sentiment of people against particular subjects, without the need of any training or human check like common machine-learning techniques. This approach can be used for any

text type, and for any discussed subject while it also provides a popularity estimation of high accuracy.

The overall work flow of the technique presented in this thesis can be briefly described by figure 2.1. After a collection of social media posts over a subject is completed, the hybrid classifier based on posts' syntactical analysis, is used to distinguish subjective from objective posts. The subjective posts, that actually contain peoples' opinion, are evaluated by the proposed sentiment estimation technique, and this way a measurement on the population of entities related to this specific subject can be established.

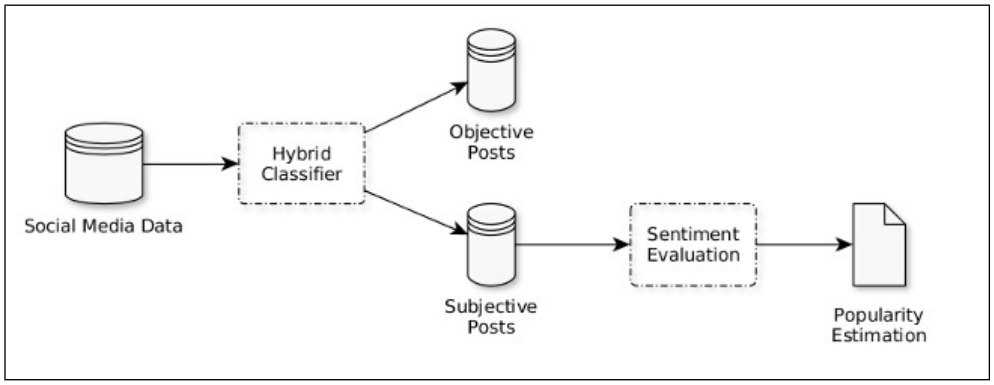


Figure 2.1 – Work scheme of popularity estimation via social media data analysis

2.3 Uses of social media sentiment

Research work [26, 10, 25], focused on the film industry domain, aims to the evaluation of data collected from different social media, like twitter, blogs etc using one of the techniques described above. By observing and plotting how buzz and attention is created for different movies and how it changes over time, one can witness whether this somehow reflects the real success and popularity of each movie. Developing techniques in sentiment analysis, can make it possible to quantify the aggregative level of positive and negative mention, while also a close look of how different opinion are propagated and influence peoples' opinion, can be also attempted. It is found that prices of the movie industry have a strong correlation with observed outcome frequencies, and therefore they are considered as good indicator of future outcomes.

Wong, Sen and Chiang [38] work specifically on the evaluation of critical

posts about movies, collected from Twitter, IMDb and Rotten Tomatoes. Using SVN machine learning they identify positive, negative and neutral posts, and based on the number of these, an evaluation of their popularity is established. This proportion of positive and negative posts is then compared to the IMDb rating and to the Box Office gross of each movie. However, in the current thesis, the tweets evaluation is based not only on the number of positive/negative posts, but also on their sentiment magnitude, which as is shown later results in a more accurate evaluation of the overall movies' popularity.

Another rapidly-rising research area is predicting outcomes on political matters, especially before important events, such as elections. Conover, Ratkiewicz, Francisco et al. make an extended research [17] on how social media shape the networked public sphere and facilitate communication between communities with different political orientations. The observation of how information is propagated and discussed between network clusters that correspond to different political parties, can be a good indicator of how people judge and correspond to different political matters according to their beliefs, which is invaluable information for politicians and analysts.

Social media data analysis was also performed on [15] for tracing the cholera outbreak of Haiti after the earthquake in January of 2010. The Haitian Ministry of Public Health (Ministére de la Santé Publique et de la Population, MSPP) has published data facilitating studies examining the evolution of the epidemic. Data collected through such health institutions and official reporting structures may not be available for weeks, hindering early epidemiological assessment, whereas data from informal media (such as Twitter) are typically available in near real-time and could provide earlier estimates of epidemic dynamics. During this study, it was found that trends in volume of informal sources were significantly correlated in time to the official case data and was available up to two weeks earlier.

2.4 Outline

The rest of this thesis is structured as follows: In chapter 3, after a brief presentation of Twitter's functionality, a hybrid classifier of objective/subjective tweets is presented, with emphasis on the NLP tools used for its implementation. This classifier was tested on a long list of tweets and its estimations were compared against those of other relevant classifiers. The construction of a sentiment magnitude estimation tool can be divided into two distinct procedures: First, the sentiment establishment of adjectives, presented in

chapter 4, and second the sentiment estimation of the whole texts and finally of certain films, which is discussed in chapter 5.

Chapter 3

Tracing subjectivity in social media posts

Measuring the popularity of subjects discussed online and estimating public opinion on various matters can be useful in many ways. Peoples' preferences are a major concern for companies' marketing departments, while users' feedback over products can essentially an important factor to make their products better. Measuring public opinion can be also very crucial into predicting the outcome of elections and other events important to society.

However, only counting the amount of discussions on a particular subject is not a safe approach for reaching to conclusions on what people approve or are opposed to. The actual peoples' sentiment on a specific matter is expressed only through subjective posts, while objective ones do not add anything into estimating their preferences. Besides, a lot of advertising posts can lead to incorrect results and should be excluded from this procedure. Consequently, there is a need to analyze the social media posts at a deeper level and generate safe estimations of peoples' opinions on a particular subject.

Since the last decade more and more people use the social networks (for example Facebook [2], Google+ [3] and Twitter [8]) to express their opinions, preferences, oppositions and thoughts on a variety of matters. Social networks can be therefore seen as a perfect source of data to be analyzed for estimating popularity of products or persons. Twitter is one of the most widely used tools, while it focuses on simple text posting over any issue.

3.1 Twitter

Launched in July of 2006, Twitter [8] rapidly gained worldwide popularity, with over 300 million users as of 2011, and became one of the highest-ranking social networking sites in January 2009, according to the Alexa rank [1]. Twitter, which enables its users to send and read text-based posts of up to 140 characters, known as "tweets", rises its usage especially during prominent events.

As a social network, Twitter revolves around the principle of followers. When you choose to follow another Twitter user that user's tweets appear in reverse chronological order on your main Twitter page. Personal updates, interesting links, music recommendations and various reviews are usually shared among followers. Users can group posts together by topic or type by use of hashtags - words or phrases prefixed with a "#" sign. Similarly, the "@" sign followed by a username is used for mentioning or replying to other users. To repost a message from another Twitter user, and share it with one's own followers, the retweet function is symbolized by "RT" in the message.

One can say that if a person follows the discussions made on Twitter he can be fully informed of what are the hot subjects at any specific time. Twitter is now a place of flowing news and users can be updated real-time for events in progress, such as sport events, elections, competitions etc. In the past, exceptionally high traffic of tweets has been noted during important sport games: For example, 2,940 tweets per second were posted during the match between Japan and Cameron on 2010 FIFA World Cup, while 3,085 tweets per second were also posted after the Los Angeles Lakers' win in the 2010 NBA Finals. The extraordinary number of 7,196 tweets per second were published during the FIFA Women's World Cup Final between Japan and the United States, two of the countries were Twitter is mostly used [24]. Apart from, sport games, Twitter seems to have unusually high traffic during celebrity events too: A major example is Michael Jackson's death, which caused a rate of 100,000 tweets per hour.

Twitter has been also used for helping people to cooperate and organize protests, riots and reactions of this kind, the so-called "Twitter Revolutions". In Moldova a big demonstration was organized through Twitter [33], on 7th April of 2009, claiming that the results of the the 2009 Moldovan parliamentary election were fraudulent. The demonstration turned to a violent riot of 10,000 people in the town of Chisinau which caused numerous reactions from the European Union and all adjacent countries.

In Tunisia, an intensive campaign was supported through Twitter (and

Facebook) [30] against the current government which was responsible for high unemployment, inflation and general corruption. This campaign led to an extended series of street demonstrations, while in January of 2011 the Prime Minister designated and free, democratic elections were held after 60 days. Protests were also held in almost all around the world by Iranian protesters against the disputed victory of the Iranian President elected on June of 2009. These protests were organized and spread through Twitter [34], which helped the development of the "Iranian Green Movement". The above cases show the importance of online social networking in modern societies and give one more reason for the need of tweets' analysis and automatic estimation of their sentiment.

3.2 Natural Language Processing

Twitter data is frequently used for measuring movies and products popularity by most of the existing approaches. However, all of these approaches are mostly based on machine learning or statistical techniques using words usually used when expressing opinions. This means that the evaluation of a post is decided only by the appearance of a specific word with strong sentiment. Without examining the relation of this word to others and determining the actual meaning of the phrase, this technique ofter leads to wrong classification.

For this reason, the need of analyzing the syntactic structure of texts appears as a way of further exploiting its essence and therefore can lead to more accurate classifications. The main idea of such an approach lies on the identification of structures that denote the existence of sentiment towards the discussing matter.

If the author expresses some form of judgment, the input can be considered a subjective statement, otherwise the data is classified as an objective claim. Example cases distinguished by such test are for example "*The King's Speech* was a really nice movie." versus, "I watched *The King's Speech*", respectively. Once the existence of a sentiment has been established, typically a classification step is performed to determine whether the speaker is expressing a positive or negative opinion over a particular subject matter.

The analysis of the input posts requires the use of a natural language processing tool, which is able to determine the structure of the text and display the relations between words. Using the grammatical patterns found in a text, it is possible to trace any subjective writing included and therefore to denote the existence of opinion within this text.

3.2.1 Stanford NLP

The Natural Language Processing (NLP) Group at Stanford University having people from both linguistics department and computer science field, has worked intensively on algorithms that process and analyse human languages, during the last years. Among other research areas that have been covered by NLP group, a complete software package on probabilistic sentence parsing and tagging has been implemented, and is now used for the purposes of this classifier.

3.2.2 Stanford NLP parser

Stanford NLP parser [22] is a natural language parser that performs grammatical analysis to documents, providing as output not only grammatical structure trees, but also dependencies between words. Using this kind of information we can get deeper into the meaning of sentences and determine the essence of what each text is presenting.

Stanford NLP parser uses a probabilistic algorithm that uses the language knowledge gained from hand-parsed sentences (tarining set) and produces the most likely analysis of given sentences. This package is written in Java and implements a factored product model, with separate PCFG phrase structure (simple syntax trees) and lexical dependencies, whose preferences are combined, using an A* algorithm. This efficient algorithm calculates all potential combinations of syntax models for a given text and returns the one with the highest likeness probability according to the training set. It has been calculated that this approach has precision 75,3% - 83,7%, and recall 70,2% - 82,1%.

The parser outputs various representations of the grammatical analysis performed on a given piece of text. One of them is the grammatical tree that represents the syntax being used, as well as the relation that each word ha to another. An example for the tweet “I liked the movie" is given in the following:

```
[I, liked, the, movie]
(ROOT
  (S
    (NP (PRP I))
    (VP (VBD liked)
      (NP (DT the) (NN movie)))))
```

The above result text shows that the sentence (S) can be divided to a noun phrase (NP) and a verb phrase (VP). The verbal phrase can be further divided into to a noun phrase which includes a determiner (DT) and a noun (NN). The above information provides a complete syntactic decomposition of the sentence, and makes it possible to determine its structure.

Another very important output, which is mostly used for the implementation of the described hybrid classifier, are the word dependencies that consist of a tuple of words (expressing a governor and dependent relation) along with an identifier of the type of their relationship. An example for the tweet "I liked the movie" is given in the following:

```
nsubj(liked-2, I-1)
det(movie-4, the-3)
dobj(liked-2, movie-4)
```

Here, the parser is reasoning that "I" is the nominal subject (nsubj) of "liked", and "movie" is the direct object (dobj) of the verb "liked". Additionally, "the" is a determiner (det) to the word "movie".

Using the above tuples we can usually determine which is the discussed object, what kind of verb is used for it and how is being characterized. This can lead us to a safe conclusion about the meaning and the sentiment of the given sentence. Additionally, NLP parser returns an extensive variety of modifiers (such as adjectival modifier, adverbial modifier, prepositional modifier) which help us extract the essence of any descriptive grammatical components found.

3.2.3 Stanford NLP POS Tagger

Stanford Part-Of-Speech Tagger (POS Tagger) [35], is another tool that can be used in order to determine the role of each word in a text and trace the appearance of sentiment in it. It reads text in and determines parts of speech to each word, such as verb, noun, adjective, etc. It deploys maximum entropy methods and returns the most likely POS tag for each word, according the the probabilities calculated based on the given training set. It presents a maximum tagger accuracy of 97.24%, while it can also trace whether a noun is in singular or plural form. As an example, the output for the tweet 'I liked the movie' is the following:

```
I/PRP liked/VBD the/DT movie/NN
```

The part of speech tagger tells us that "I" is a personal pronoun, "liked" a verb in past tense, "the" a determiner and "movie" a noun in singular. The

above approach can be used for tracking down adjectives or verbs in a given text, while this kind of words might be very useful for the determination of the polarity and sentiment of the text.

Natural Language processing in subjectivity sentiment

Using the output of the tools discussed previously, it is possible to analyze the way that a text is being structured and try to derive some conclusions regarding its meaning. Having as an aim to distinguish subjective from objective texts, it is necessary to examine the common ways that people usually express their opinions. Subjectivity signifies the existence of some kind of judge on a particular subject, which reveals either the approval or the opposition of the author against this subject. This is usually expressed by the use of words of strong sentiment, which are mainly adjectives and verbs.

For example, giving a specific characterization to a subject, like in the phrase "This *album* is amazing", is a very common way to declare an opinion regarding this specific matter. Besides, verbs can also play this role, for example "I love this *film*", is a definite statement of approval, and consist a case of subjective writing.

Therefore, we can generate two rough rules based on the most common patterns traced in subjective texts:

1. if an adjective is referring to the subject → subjective
2. if a verb expressing emotion is referring to the subject → subjective

This means that a text can be classified as subjective if it contains one of the above patterns. Otherwise it is classified as objective, while there is no visible indicator that it contains any kind of opinion. Tracing such patterns can be easy with the use of a Natural Language Processing tool, and is the main idea of the hybrid classifier presented in this thesis.

In this study, we extracted data from the movie industry domain, firstly because a wide range of movies are being discussed every day in all social media and secondly, because a lot of viewers after watching a film, are willing to give his opinion of it, through the Internet. That gave us an excellent opportunity to collect a sufficient number of texts that not only express viewers' opinion, but also give some kind of 'goodness' levels regarding the discussed films. Twitter is the social media mostly used in this work, while it provides an enormous number of short-length texts that are meant to present opinions in a clear and straight-forward way. However, apart from

Twitter posts, data from IMDb and Rotten Tomatoes are also used in the establishment and estimation of sentiment magnitude, as described in the next pages.

3.3 Classification Methodology

The main idea behind this hybrid classifier (yielding to a categorization of objective and subjective text) is that the actual meaning of a statement can be found if we derive its grammatical structure and examine its components one by one. That would lead us to safer results while this method does not depend on any training corpus, does not need any human provision and can be applied to any type of text (tweets, e-mails, wiki documents).

As noted before, people usually express their feelings using statements of specific patterns: For example verbs like 'love/like/hate' are directly expressing an emotion and therefore are used in subjective sentences which include a person's opinion on a matter rather that some general information. Additionally polarity is also very often expressed with a use of an adjective that is related to the topic of the sentence (or any other subject that we examine).

For this approach we use the Stanford NLP parser that was discussed previously, and especially the generated word dependencies that denote the relations between words and the type of them. For example having the sentence: "Black swan is a very beautiful film", NLP parser generates the following dependencies:

```
nn(swan-2, Black-1)
nsubj(film-7, swan-2)
cop(film-7, is-3)
det(film-7, a-4)
advmod(beautiful-6, very-5)
amod(film-7, beautiful-6)
root(ROOT-0, film-7)
```

The semifinal dependency "amod(film-7, beautiful-6)" reveals that there is a connection between the words "film" and "beautiful" and that the latter word ("beautiful") plays the role of an adjectival modifier (amod) to the word "film". This fact is sufficient to conclude that the subject "film" is characterized by an adjective (or adjectival term) and therefore the phrase includes some kind of judgment. Consequently, having analyzed the syntactical structure of the text we can safely claim that it can be classified as a subjective one, towards the subject "film".

Another common example has the following form: "I really loved Inception", which generates the following dependencies:

```
nsubj(loved-3, I-1)
```

```
advmod(loved-3, really-2)
root(ROOT-0, loved-3)
dobj(loved-3, Inception-4)
```

The last dependency "dobj(loved-3, Inception-4)" denotes that there is a direct object (dobj)- "Inception", for the verb: "loved". Being able to identify the verb "love" as one of strong sentiment, we can safely conclude that the subject "Inception" is directly linked to a verb of judgment. That can lead us to the claim that the above text expresses an opinion on the subject "Inception" and therefore can be safely classified as subjective.

However, the subject of the text can be expressed with multiple words. For example the above sentence can also be written in a pattern similar to: "Inception: I really loved this movie", which will generate the relation:

```
dobj(loved-3, movie-4)
```

In order to be able to catch all possible expressions of this statement, a list of nouns that could refer to the focusing subject, is given as an input. In the case of films, such nouns can be: movie, film, picture, the movie title etc. That way, not only every form of such statement can be traced but this list also enables the expansion of the tool to different domains and makes it a useful method for a variety of subjects.

A similar list of verbs that can be used to show some sentiment towards the discussed subject, is also given as an input. Some examples of such verbs are: like, love, hate, prefer etc. Adding both present and past forms of such verbs to the list, would be sufficient to trace the majority of subjective writings of this pattern.

In order to be able to trace the adjectives being used to characterize the subject, we use the Stanford NLP Tagger, in patterns where an adjectival modified is linked to the discussed subject. That way, it is possible to avoid incorrect classifications when having misleading syntactic structures but also collecting all adjectives is helpful for the sentiment estimation of Tweets, described in the next chapters.

However, there are many cases of slang terms instead of normal adjectival clauses that the tagger fails to recognize. For this reason a list of addition adjective that might be used in also given as input. As an example, the phrase "Inception, what a top film!" generates the relation:

```
amod(film-6, top-5)
```

The Stanford NLP Tagger, classifies the word 'top' as noun , even though it is very often used as an adjective, mostly in informal language. It has been

observed that after including such potential terms, a much higher precision on subjective tweets classification can be reached, something totally expected if we consider the writing style of most of the tweets posted.

Taking in consideration the above cases, one can easily end up with a list of syntax patterns that reflect all (or at least most of) the forms of subjective writing. Having a close up look to such syntax patterns, we built a list of simple rules to detect whether a message is a subjective or objective statement, that can be described as following:

1. adjectival modifier either traced by the Stanford NLP Tagger or included in the 'adjectives list' is related to a word included in the 'subjects list' → subjective

2. a verb included in the 'sentiment verbs list' is related to a word included in the 'subjects list' → subjective

The whole procedure of the hybrid classifier is simple: For each one of the input tweets grammatical analysis is performed using the tools described in a previous chapter (build by Stanford NLP Group). As a result of it, a list of the word dependencies found is maintained. After, having these syntax relations between words we check whether any of the above rules apply in the specific tweet. In such case, we categorize the tweet as subjective (with respect to a particular subject), or objective in any other case.

The overall work flow of this technique, can be briefly described by the figure 3.1.

3.3.1 Limitations

Even though Twitter is flooded by simple and short messages, that usually do not contain sophisticated syntax structures or complicated meanings, they often contain slang terms, internet writing style, acronyms or even internet jokes and commonly-used web phrases. That of course may lead us to wrong syntax analysis of texts (while NLP parser is being trained by normal English writings), which may become a reason for the hybrid classifier to miss a subjective pattern or even proceed to a wrong classification of a tweet. On the other hand, this factor can be significantly reduced if we use the adjectives list that was described before.

Another problem that can be often faced is wrong spelling of words or wrong syntax of sentences. Even though, this factor cannot be completely corrected by any algorithm (or even by any human), the hybrid classifier does not need entirely correct sentences, while it only examines a few key-words

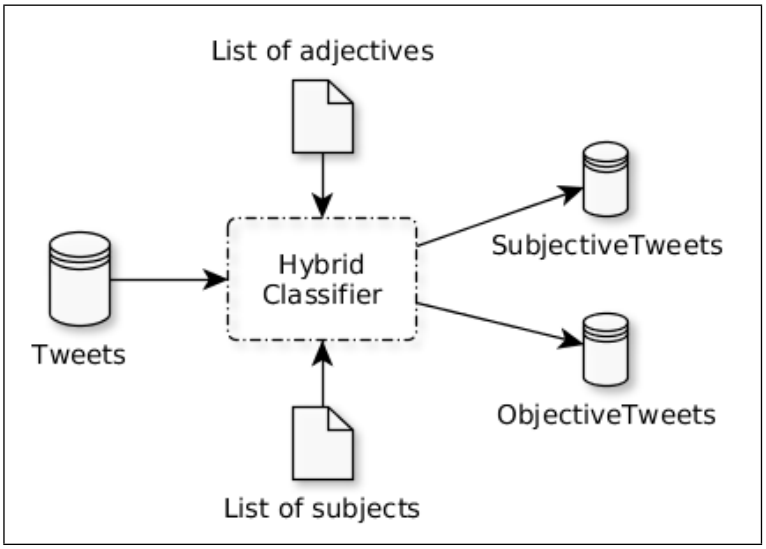


Figure 3.1 – Work scheme of adjectives sentiment establishment

from the entire tweet. For example if a (correctly-spelled) adjective can be associated with the particular subject that we examine, the tweet will be categorized as subjective even if it contains a small typo (which pretty often the case in internet texts).

As one can easily imagine, this kind of classifying algorithms cannot predict the cases of irony (ie. statements that mean the exact opposite of what they are saying). Such statements require a prior knowledge on the subject in order to trace some kind of exaggeration that would imply the fact that some text has an ironic meaning. This kind of research is out of the scope of this thesis, and is left as future work for the additional improvement of natural language classifiers.

3.4 Sentiment classification performance

To evaluate the performance of established sentiment classifiers and create a benchmark for our developed solution, we randomly sampled a set of 1,000 messages from the microblogging platform Twitter. This corpus was tested upon existing classifiers and the prososed hybrid classifier, and their results were then compared.

For our evaluation, we collected a data-set of more than 1,000 randomly chosen tweets related to the five most popular films of the 83rd Academy Awards. We used the language detection library of Cybozu Labs [32], in order to eliminate the tweets written in any language other than English, while we also tried to remove advertising tweets out of the set. Multiple retweets of the same text were also removed to prevent performance over- or underestimation, as well as unnecessary tokens like link urls, "@" tags for mentioning a user, 'RT' tags etc. Each tweet of this test-set was classified by hand before the begin of the evaluation into an objective or subjective statement.

3.4.1 State-of-art-classifiers

For our analysis, we focus on those approaches for which the original authors made a reference implementation available to us, specifically we compare the classification accuracy with the following classifiers:

Twitter Sentiment We used the bulk classification service available on Twitter Sentiment website [18] in order to classify our test-set. This tool attaches to each tweet a polarity value: 0 for negative, 4 for positive and 2 for neutral- therefore we consider the first two describe subjective tweets, while neutral is for objective tweets. The main idea behind Twitter Sentiment approach is the use of emoticons as noisy labels for the training data which is shown that it increases the accuracy of different machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM). It is noted that the web service of Twitter Sentiment uses a Maximum Entropy classifier.

Tweet Sentiments Our test-set was also tested through the API of Tweet-Sentiments [7], a well known tool for analysing Twitter data and provide sentiment analysis on tweets. TweetSentiments is based on Support Vector Machines (SVM) and is using the LIBSVM library developed at Taiwan National University. It classifies tweets as positive, negative or neutral and these values are treated as stated previously.

Lingpipe We also used the Sentiment Analysis tool of the LingPipe [5] package which focuses on the subjective/objective (as well as positive/negative) sentence categorisation especially on the movie-review domain. This approach uses the usual machine learning algorithms (Naive Bayes, Maximum Entropy, SVM) and a Java API of the classifier is available online. Even though it comes with its own training set, we used the half of our hand-classified set to train the classifier, in order to have better results. The other half was used as test-set and results were compared to the corresponding hand-classified tweets.

3.4.2 Results and Evaluation

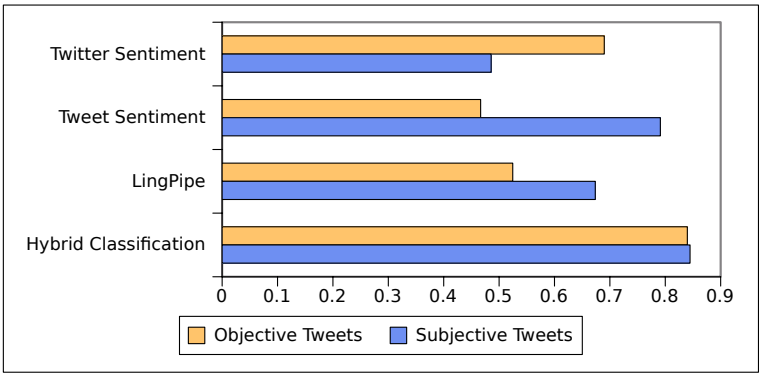
Comparing the output against the previous human classification, the overall accuracy of the automatic classifiers in distinguishing subjective from objective statements was measured, as shown in figure 3.2(a). Figure 3.2(b) shows the overall performance in correctly and incorrectly classified statements.

As can be seen in the figure, the classification accuracy of all statistical sentiment analyzers is between 55 and 60%, whereas the proposed statistical-grammatical hybrid approach yields a correct classification accuracy of about 85%, a 40% gain of previous work. Note also that the accuracy of existing system also varies significantly between the type of input data: Twitter Sentiment [18] for example is much stronger in identifying objective statements compared to subjective ones, while Tweet Sentiment [7] shows exactly the opposite behaviour. The proposed hybrid solution on the other hand does not any significant difference.

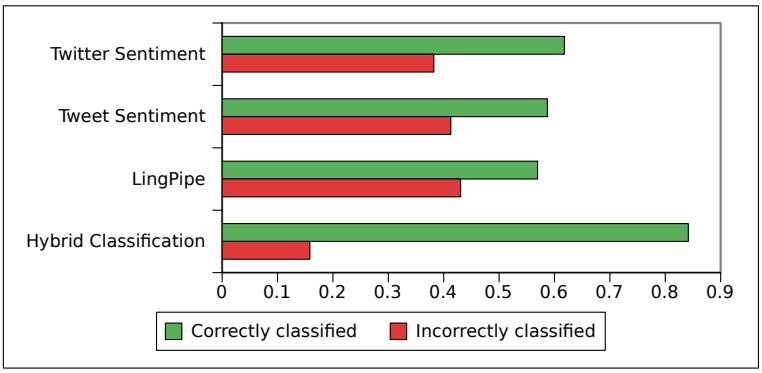
As noted before the reason for these classification failures of the existing systems is mainly the fact that they evaluate words independently and not as parts of a phrase. That leads to incorrect estimation of the tweet meaning and finally an incorrect classification of its sentiment. Below there are some examples of the above claim:

'Now who will I see Black Swan with? I was sick when all my friends went.'

The above tweet was classified incorrectly as subjective by lingpipe and tweet sentiment, while the word 'sick' has a negative meaning and therefore there is the assumption that the whole tweet contains a negative sentiment. As a result, the above tweet was wrongly considered as subjective, while on the other hand, the hybrid classifier found no relation between the adjective 'sick' and the object 'Black Swan' and therefore it classified the tweet



(a)



(b)

Figure 3.2 – Classification accuracy of different sentiment analyzation methods.

correctly as objective. There is a large number of such cases that mislead the existing classifying tools into incorrect estimations, while contrary the syntax analysis that is performed by the discussed hybrid classifier, limits or even extinguishes these cases.

'What an awesome day with my lovely wife! We had lunch at Olive Garden & saw the matinee of king's speech'

The above tweet was also classified incorrectly as subjective by most of the used classifiers, while 'awesome' and 'lovely' are some clearly positive words that give to the whole text a positive sentiment. The tweet is therefore classified as subjective, even though, in fact, there is no connection between the above words and the film that is included in the text. This

clearly shows, that the syntax analysis which is performed by the proposed approach, contributes a lot to a correct classification of tweets to objectives and subjectives and gives a major reasoning of the accuracy difference that was observed in the above measurements.

Chapter 4

Adjectives sentiment establishment

After having developed an efficient way to separate subjective from objective texts, it is now necessary to extract the sentiment polarity and their magnitude. In other words, after collecting the opinion of each user over a matter, expressed by a short post, there should be a way to determine whether this post is positive or negative, and additionally how positive or how negative it is.

The sentiment polarity and magnitude over a particular subject is usually expressed by the use of adjectives. This holds also for the film critics field, to which this analysis is focused. For example, people usually express their opinion concerning a film they watched, mostly by characterizing it as amazing, terrible or just good. If we ask a group of people to give an adjective for a particular movie, we can then have an idea of how positive or how negative public opinion is on this film.

Therefore, one essential step to be taken is to establish a metric of how positive or how negative an adjective is, in other words the magnitude of the sentiment. For instance, a movie characterized as 'brilliant' is much more appreciated than a movie characterized as just 'good'. The purpose of this chapter is to numerate the 'goodness' of adjectives so that any comparison and sorting among them can be possible.

4.1 Methodology

While it is necessary to establish an automatic approach for assigning sentiment to adjectives (and words in general), we used an unsupervised approach based on word correlations. This approach is inspired by the way

a person is learning to judge which words have a positive or negative meaning, which is essentially a result of a lot of exposure to speech and written text, from which the learner infers which words appear in a positive or negative context.

Specifically, positive words (adjectives in our case) usually occur in sentences of positive sentiment, while contrary negative words are commonly found in texts of negative sentiment. That way words (and adjectives in particular) are obtaining a sentiment notion which would be very useful to extract.

The same basic principle, inferring which words appear together in a positive or negative context, can be easily implemented in a simple algorithm. Counting how often a particular adjective has been encountered with a positive meaning compared to the frequency it has been observed with a negative connotation is all that is needed. More precisely, the desired value here is calculated by subtracting the total number of shows of a particular adjective in a positive context divided to the number of positive texts processed, by the corresponding frequency for negative texts.

$$sent_{adj} = \frac{\sum adj_{positive}}{|positive|} - \frac{\sum adj_{negative}}{|negative|}$$

Afterwards, this value is been normalized by the total number of words in the corpus: It is multiplied by the total times of occurrences of all words divided to the number of occurrences of the particular subject. That way, uncommon adjectives like "brilliant" get a higher ratio that the common ones like "good":

$$sent_{adj} = sent_{adj} \times occur_{allwords}/occur_{adj}$$

To begin such an automatic classification, some notion of what is deemed positive or negative will be necessary. In order to proceed to a sentiment estimation of adjectives, a list of texts with an indication of their sentiment should be provided. Having a collection of texts with positive and negative context makes it possible to measure word occurrences in contexts of different polarity, an therefore to establish a sentiment metric for these words (fig. 4.1).

Additionally, a list of adjectives to be evaluated is also provided. The hybrid classifier presented in the previous section not only can be used to distinguish subjective texts from objective ones, but also collects the adjectives used in the opinion posts, referring to the examined subject. Therefore, a list of adjectives used to describe this particular matter is already available.

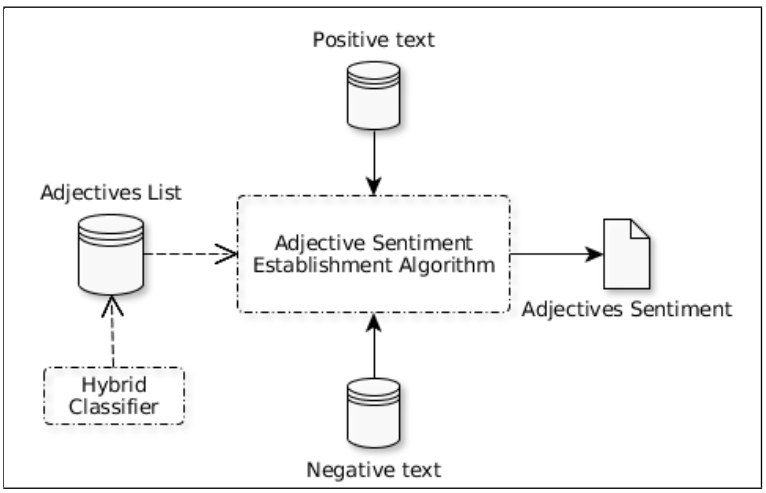


Figure 4.1 – Work scheme of adjectives sentiment establishment algorithm

4.2 Corpus

In our approach we chose movie-related posts that are already characterized as positive or negative as for their content. Internet Movie Database (IMDb) [4] and Rotten Tomatoes [6], two Internet websites focused on film reviews, provide a large collection of movie critics marked according to their polarity. In IMDb each one of the critics is accompanied by a number of stars denoting the writer's opinion against this movie, while Rotten Tomatoes follows an even clearer policy: Every critic is either characterized as 'fresh' if it is positive, or 'rotten' if it is negative, while additionally they divide their reviews to those written by professionals and to those written by regular users. These websites also include the following ratings for each movie:

- IMDb rating [4]: This is the average score of all ratings that users submit for this movie. All visitors are able to rate films on a scale of one to ten, even if they are not IMDb registered members. IMDb is one of the most popular online entertainment sites, with over 100 million unique users each month launched in 1990.

- Rotten Tomatoes rating [6]: Rotten Tomatoes distinguishes its critics to those written by professionals (journalists, website editor etc), and to those written by regular users. Accordingly, the website publishes two different ratings for each movie: the percentage of approved professional critics that gave a positive review which will be called professional rating in our analysis, and the percentage of users that

rated the particular movie with more than 3,5 stars, which we call audience rating.

We collected 122,656 audience and 22,920 professional critics from Rotten Tomatoes and 27,030 reviews from IMDb, for almost 140 movies released in the years 2011 and 2012 from those two websites, and measured the frequency that adjectives were encountered in both a positive and a negative context.

A number of Twitter posts was also used for adjectives' sentiment establishment. However, in Twitter there is no clear indicator of which tweet has a positive content and which has a negative one. For this reason, the polarity indicator in this case is the use of smilies. We collected a number of 1,014,414 tweets containing smilies, and we consider positive those having a positive smilie such as :-), :), =) and accordingly we did the same for negative smilies as :-(, :(, =(too. A representation of positive/negative posts distribution of this corpus can be found in figure (4.2).

Using the techniques discussed above, we dissect all individual statements and count the number of occurrences in posts with positive or negative content. This results in a relative assessment of a particular word to appear in a positive or a negative context, where context is defined either by indicators of each website or by the appearance of smilies.

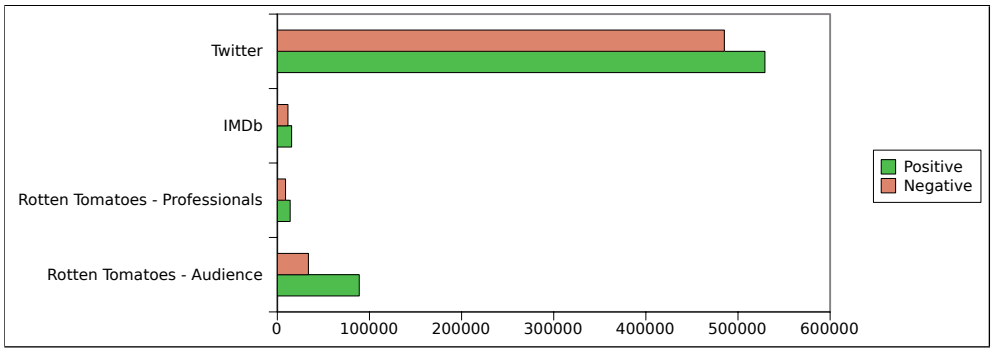


Figure 4.2 – Number of posts from each website used as training copus

4.3 Adjectives selection

In this approach, we use the list of adjectives found in input tweets related to movie reviews, by the hybrid classifier. As was described in the previous section, these adjectives are found to have a direct relation to the subject of the posts, and therefore are used by users in order to attach a sentiment to the particular targets.

However, having a small training corpus, might need a more careful selection of adjectives. For example, if the adjectives 'western' or 'dark' were included in the final list, the results might be misleading, if the training set happened to contain really bad or really good critics for a classic or a dark film. However, these cases are eliminated, when having a long list of positive and negative texts.

Apart from that, different contexts require different adjectives for their evaluation. Therefore, the adjectives list used for sentiment evaluation here, should be derived from the subjective classification procedure of texts related to the same subject. If, for example, this study was focused on peoples' opinion about American presidential candidates, the adjective 'entertaining' which is chosen for the film case, would not be the most appropriate. Instead, we would definitely seek for the words 'capable' or 'trustful', which apparently make no sense for movies.

Again, an automatic way to establish a list of adjectives appropriate for any matter is by the use of the proposed hybrid classifier. That would only require a number of texts of any length or writing style, related to this particular subject.

4.4 Ground truth

In order to trace peoples' sentiment towards particular words, we asked 50 students to assign to some of the most commonly used adjectives for describing movies, a grade from 1 to 10 according to how positive its meaning is to them. This way, we can make comparisons to the adjectives sentiment estimations derived by the use of texts from different websites, as therefore appraise the results of the proposed estimation algorithm. Figure 4.3 shows the average of students' answers for each given adjective, shifted five points so that base is zero, and normalized to sum to one. They are sorted according to their calculated sentiment magnitude.

The collected answers were very similar to each other, and all students agreed on the polarity of them. For instance, all answers gave to the word 'awful' as much more negative notion than the word 'good'. Another interesting finding is that students agreed also on the sentiment magnitude of adjectives. The word 'good', for example received much smaller grades than the word 'amazing' by the majority of the asked students. However, the five higher ranked adjectives ('amazing', 'incredible', 'brilliant' etc), received slightly different grades between students' answers and some minor differences in the sorting of these adjectives were observed.

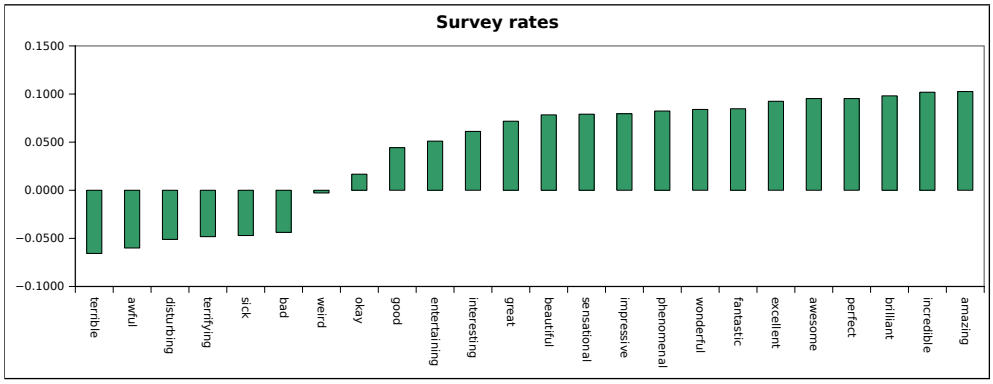


Figure 4.3 – Survey results on adjectives sentiment

4.5 Adjectives sentiment estimation from social media posts

Figure 4.4, shows the adjectives sentiment estimation derived after evaluating IMDb critics, figure 4.5 and figure 4.6 for Rotten Tomatoes critics and figure 4.7 for Twitter posts. The adjectives are sorted according to the resulted sorting of the ground truth, in order to make comparisons easier.

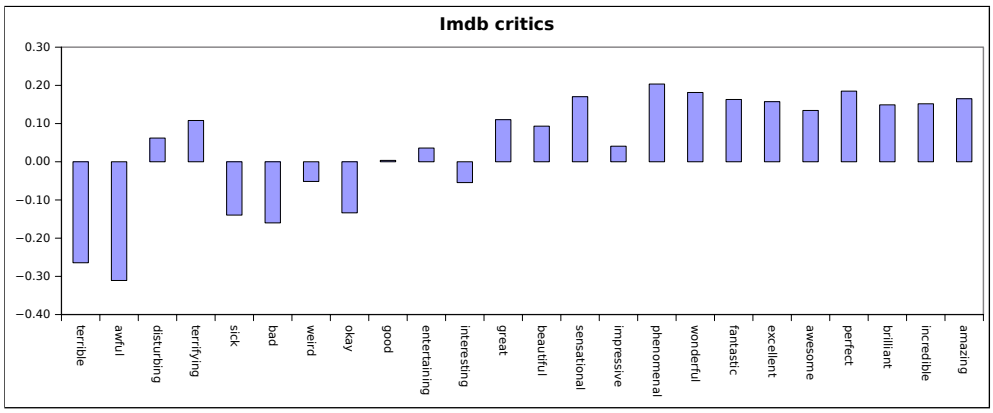


Figure 4.4 – Adjectives sentiment derived from IMDb critics

Even though there are differences among the values of particular adjectives between different social media and the ground truth, this type of adjective estimation seem to give reasonable results in positive/negative meaning of words, as well as their sorting in most cases. According to these results, the estimation algorithm is able to trace the differences in sentiment polar-

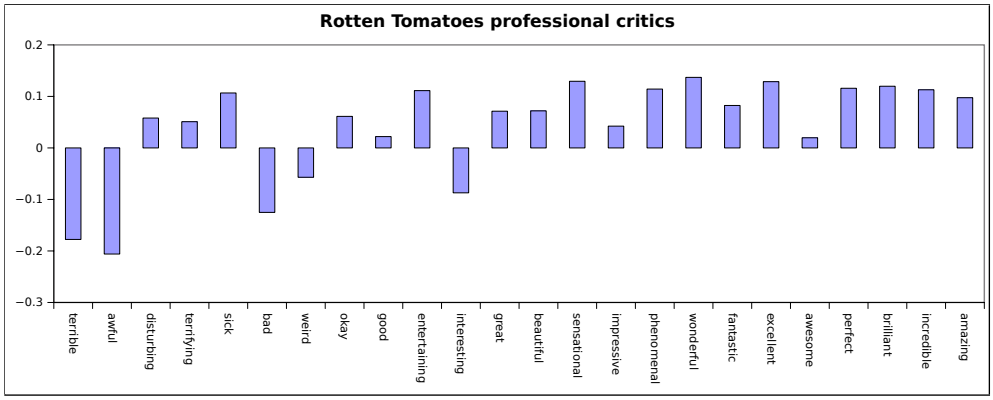


Figure 4.5 – Adjectives sentiment derived from Rotten Tomatoes professional critics

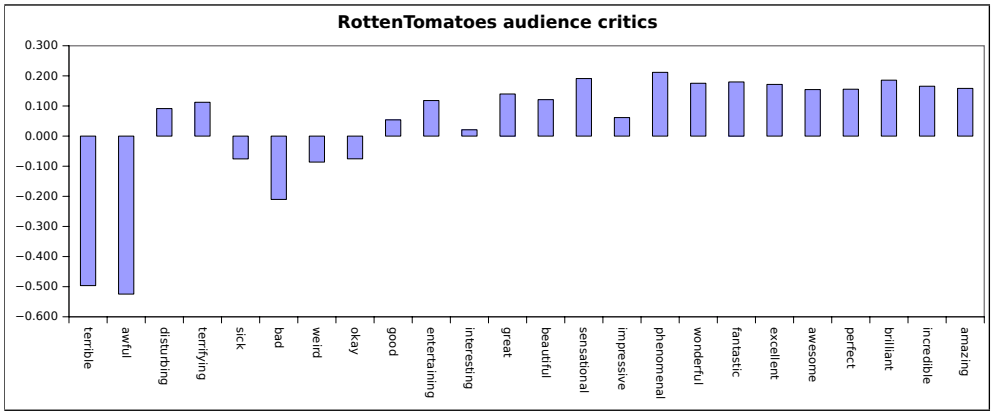


Figure 4.6 – Adjectives sentiment derived from Rotten Tomatoes audience critics

ity. For example 'terrible' and 'awful' appear to be words of definite negative meaning in all cases, while adjectives with positive sentiment, such as 'brilliant', 'worderful' and 'amazing' are given a positive score. The algorithm also succeeded in tracing differences in sentiment magnitude. For instance, 'excellent' is found to have a much more positive meaning than 'good', which ensures that the proposed algorithm is able to attach accurate sentiment estimations to given words.

On the other hand, one can notice that in IMDb and Rotten Tomatoes evaluation, the words 'disturbing' and 'terrifying', have surprisingly a positive sentiment, even though most people that answered the survey gave a definite bad meaning in both words. This is mainly because a sufficient large

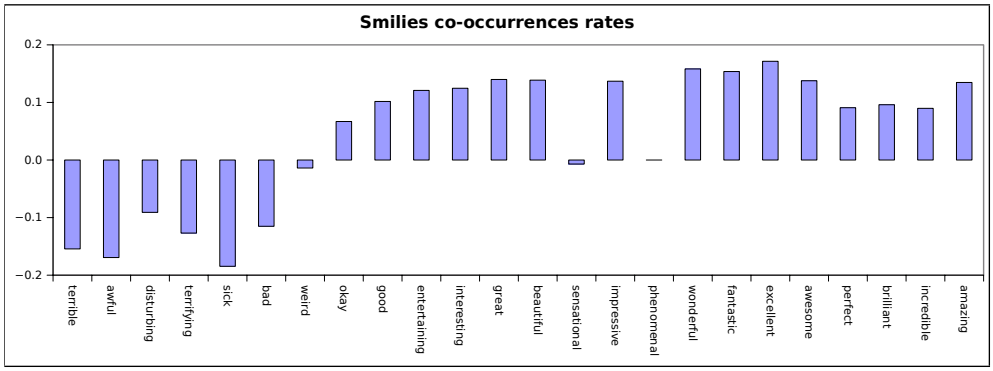


Figure 4.7 – Adjectives sentiment derived from twitter posts

number of positive IMDb and Rotten-Tomatoes critics, use those words to describe movies that receive positive rating scores. This is not unexpected, as for some kind of movies (eg. Thriller or horror films), 'terrifying' is in fact a quite positive characterization to make. For example, the following posts from Rotten Tomatoes are marked an positive (fresh), yet they contain the above seemingly negative adjectives:

- "While the all-star cast and the story, about a deadly virus, sounds similar to 1995's Outbreak, 'Contagion' is more insightful, moving, and *disturbing* with first rate entertainment"
- "A non-stop, over-the-top, intentionally ridiculous smorgasbord of violence and bloodshed that ceases being *disturbing* and just becomes pure, 100% fun"
- "Perverse, *terrifying*, hilarious in exactly the right way; smart enough, emotional enough, and at the end uniquely satisfying in any number of hard-to-quantify ways"
- "If you take Gibson's performance on its own merits, it's one of the finest of his career; touching, *terrifying* and admirably understated throughout"

This demonstrates the semantic differences one word can have in different subjects and contexts, which can be in some cases so intense that may turn a negative meaning to positive, as we have in this case. This denotes the need of separate words sentiment evaluations in different contexts, or even in different writing styles, while it is clear that the semantic of a word is not universal but can be adapted in different domains and subjects.

Using the Earth Mover's Distance (EMD) we measured the distance of each one of the above distributions to the ground truth. The EMD is a measure of similarity between two probability distributions. The two distributions can be presented by signatures and the EMD is defined as the minimum amount of work needed to change one signature into the other. Therefore the smaller the value, the more similar the distributions. Results were similar for all of the estimations compared to the ground truth and are displayed in the below table:

Source of training sets	EMD to ground truth
IMDb	2,668
Rotten Tomatoes (professional critics)	2,546
Rotten Tomatoes (audience critics)	2,379
Twitter	2,445

The distribution derived by the audience critics of Rotten Tomatoes has a slightly smaller distance to the ground truth than the others, however the differences appear to be insignificant.

Since, we already have some trustful estimations of the adjectives' sentiment, it is reasonable to explore the possibility of using them for an overall sentiment estimation of texts. The question that appears here is which one of the above estimations for adjectives is the most suitable to be used in order to calculate the sentiment of specific targets. This answer to this question is analysed in the following chapter, while all the above adjective sentiments are applied and evaluated for test sets of different sources.

Chapter 5

Measuring movies popularity

An efficient way to measure words sentiment magnitude was presented in the previous chapter. Being able to determine how good or how bad each word is, we can evaluate whole sentences according to their sentiment by simply looking at the words used. As already discussed, peoples' opinion in subjective texts is expressed mostly with the use of adjectives, and therefore using an estimation of their sentiment would lead us to an assessment of peoples' overall feelings to a particular subject. This way, the sentiment values of adjectives calculated in the previous chapter, makes it possible to establish a measurement of the public opinion regarding specific subjects, and consequently gives an indicator of their popularity.

For calculating the popularity of each movie, we used critics from both IMDb and Rotten Tomatoes (on the same movies), as well as Twitter posts having a reference to the particular movies' titles. Critics are usually texts written from users who watched the movie (either professional critics or regular users), and express their opinion by adjectives and other descriptive phrases. Having a number of more than 100 critics per movie we can have an indicator of the popularity of each movie, exactly like we would have done in our every day life.

5.1 Methodology

It is clear that just counting the number of positive and negative critics could give as a rough idea of which movie is in general really bad or acceptably good. In figures 5.1 and 5.2 one can see the relation of number of IMDb critics for each movie towards their rating given by IMDb website users and the Box Office gross respectively. Similar results hold for Rotten Tomatoes and Twitter as well. Clearly, there is no strong correlation (the correlation

coefficient is calculated round 40% to 42%) between the number of critics posted on IMDb and any of these ratings, which shows that posts number is not a reliable indication of the exact sentiment related to each one of the movies. Besides, it is reasonable that calling a movie 'exceptional' has a different importance than just call it 'good', even though both adjectives have a definite positive sentiment. The establishment of a more accurate way for evaluating critics and social media posts in general, is therefore needed.

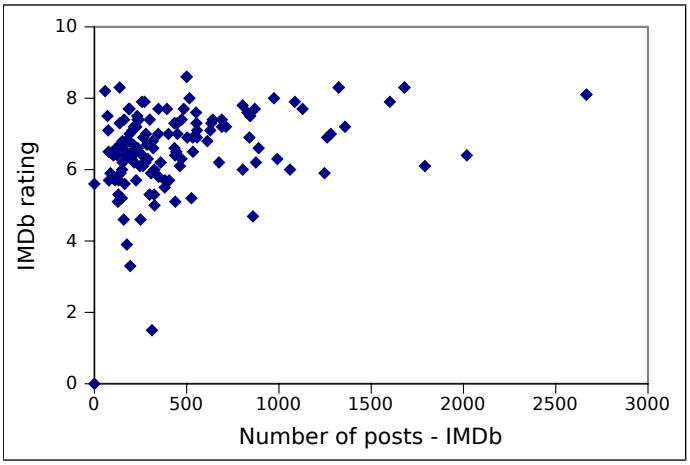


Figure 5.1 – Number of IMDb critics to IMDb ratings

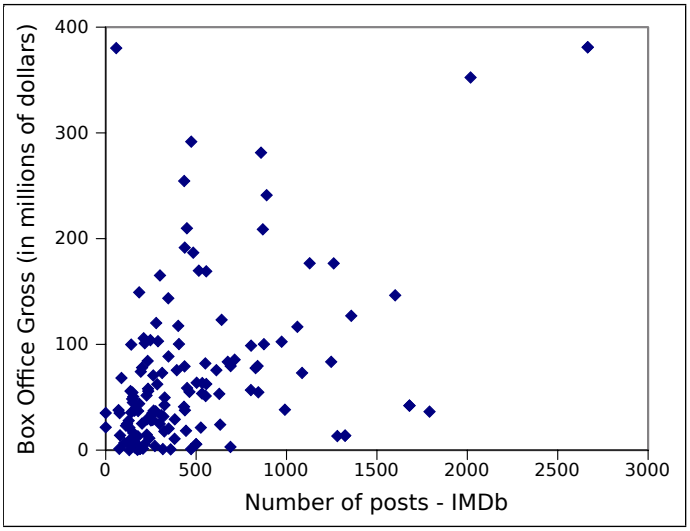


Figure 5.2 – Number of IMDb critics to IMDb Box Office gross

The technique used here is very simple: We go through every critic and

collect all adjectives contained to the texts. Afterwards, we add together the sentiment rates of all these adjectives (depending on which adjectives' sentiment estimation is being used) and then we divide this value to the total number of adjectives found:

$$sent_{film} = \frac{\sum_{adj \in film} sent_{adj}}{|adj \in film|}$$

The overall work flow of this technique, is visualized in figure 5.3. The algorithm uses an adjectives sentiment estimation generated by the technique described in chapter 4, and a list of subjective posts on specific movies. The hybrid classifier presented in chapter 3, can classify a list of given social media posts according to their subjectivity, and therefore such list of subjective posts is immediately available.

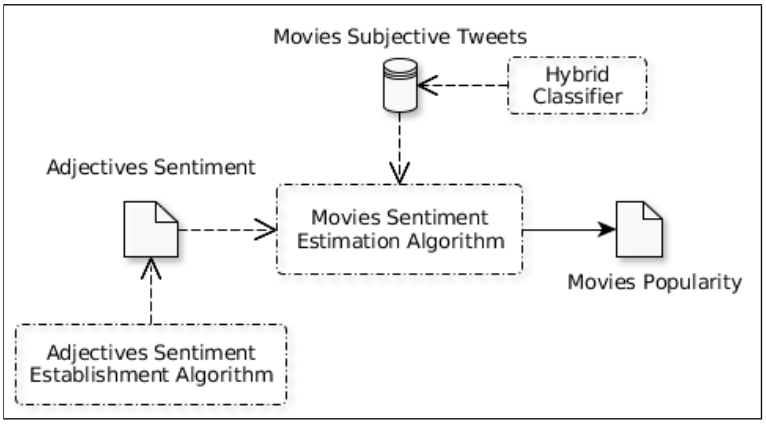


Figure 5.3 – Work scheme of movies popularity estimation algorithm

Following this procedure, we can establish a classification of the 'goodness' of any movie, only by evaluating the adjectives that are used to describe them. For instance, a popularity estimation of the 140 movies of our corpus is presented in figure 5.4. Movies are sorted from the one that has received the highest sentiment estimation score to the one with the lowest one.

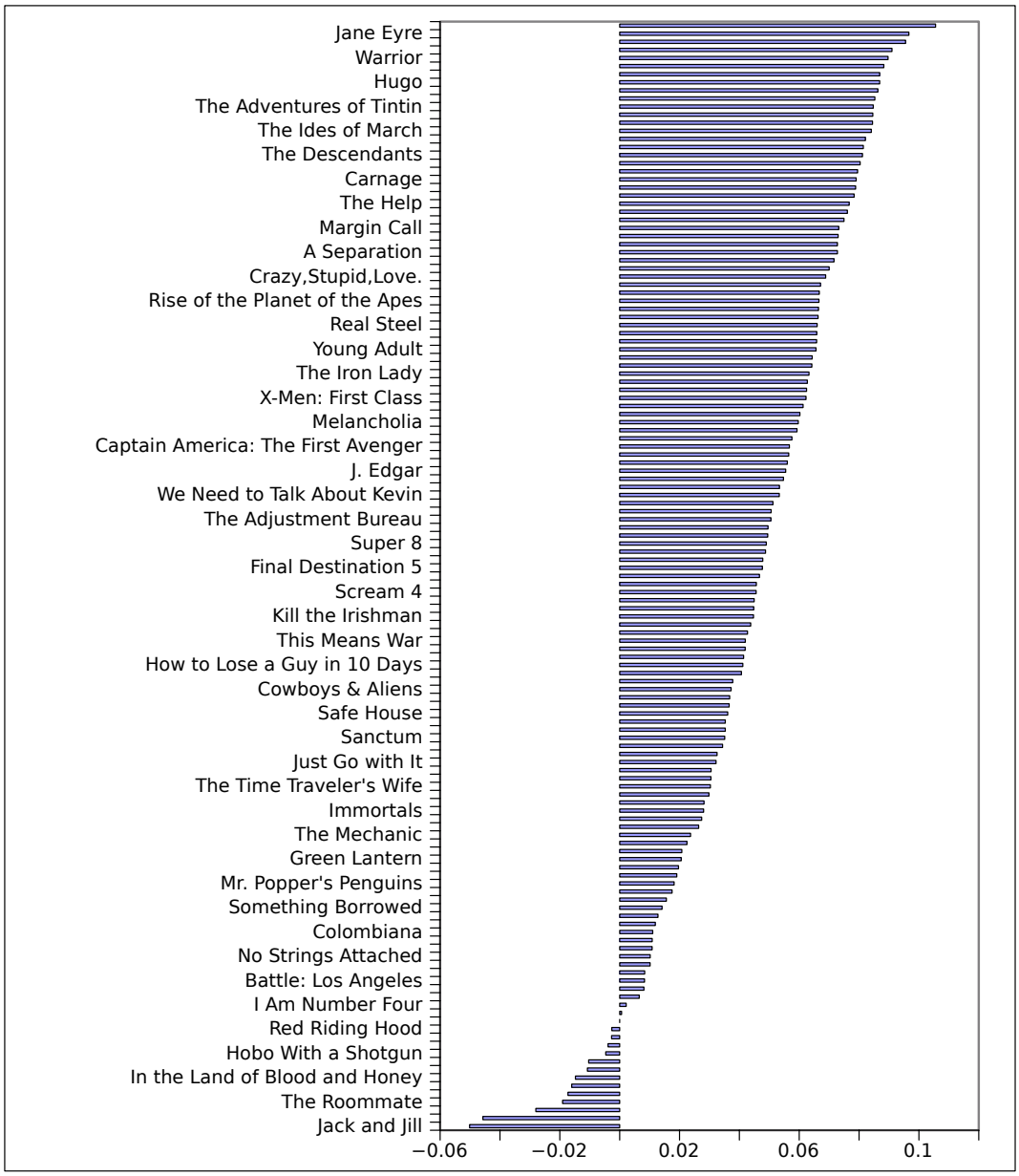


Figure 5.4 – Adjectives sentiment derived from rotten tomatoes critics

5.2 Movies sentiment estimation

The most important aspect that needs to be covered, is whether the results of the described evaluation for movies are actually corresponding to reality. In order to test the described approach, we used reviews for 140 randomly selected movies of the years 2011-2012, collected from the IMDb and Rotten Tomatoes website, as well as Twitter posts that included the titles of these movies. Using an estimation of adjectives sentiment as presented in chapter 4, we evaluated the adjectives that describe the corresponding movies. This way an estimation of their popularity was established.

Additionally, we also used different adjectives sentiment estimations, as calculated in the previous chapter. These estimations were derived by sentiment-tagged texts from IMDb, Rotten-Tomatoes, as well as tweets with smilies. This way, an evaluation of different adjectives sentiment estimations is attempted.

Below we present the relation of the estimated popularity of each movie to the ratings that these movie received in IMDb and Rotten Tomatoes websites, as well as their Box Office gross. In order to measure the correlation degree we used the Pearson correlation coefficient, which measures the linear dependence between two variables. A correlation between 30% and 50% is characterized as medium, while a correlation above 50% is considered as strong.

5.2.1 Evaluation of IMDb critics

The figure 5.5 presents the correlation of movies popularity estimated using adjectives sentiment trained on IMDb rewiews, to IMDb ratings. Accordingly, in figure 5.6 an adjectives estimation trained on the ground truth is used, while in figure 5.7 and figure 5.8 Rotten Tomatoes professional and audience critics are used as training set, respectively.

From these plots it is clear that there is a strong correlation between the sentiment estimation of these movies and the actual IMDb rating, which is considered to be one of the most accurate online film ratings. Additionally, the presented results show that the correlation of estimated movies popularity to the IMDb ratings, remain the same regardless the chosen adjectives sentiment estimation. Specifically the difference between the resulted correlation coefficients is at most 4%, which justifies that the slight differences observed in sentiment estimations of adjectives trained on different texts, are almost eliminated when used for an evaluation of a text set.

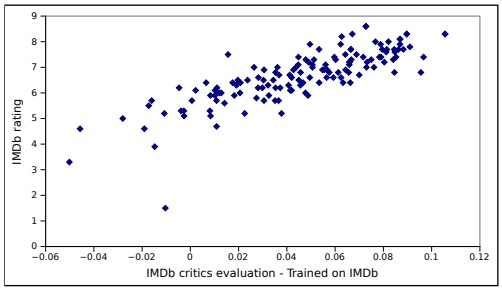


Figure 5.5 – Trained on IMDb critics compared to IMDb ratings - R^2=82%

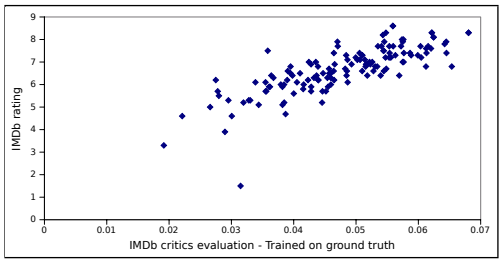


Figure 5.6 – Trained on ground truth compared to IMDb ratings - R^2=79%

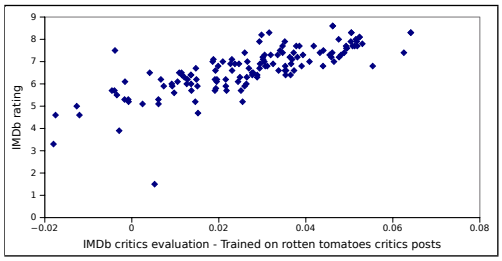


Figure 5.7 – Trained on Rotten Tomatoes professional critics compared to IMDb ratings - R^2=78%

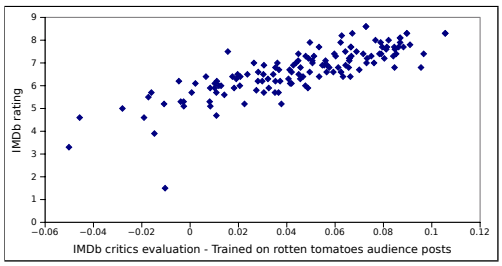


Figure 5.8 – Trained on Rotten Tomatoes audience critics compared to IMDb ratings - R^2=81%

The above estimated movies popularity was also compared to Rotten Tomatoes ratings. In the table below, the correlation coefficient derived by comparisons to both IMDb and Rotten Tomatoes audience ratings is shown:

Trained on	Compared to	Correlation Coefficient
IMDb	IMDb	82%
IMDb	Rotten Tomatoes	81%
Rotten Tomatoes	IMDb	82%
Rotten Tomatoes	Rotten Tomatoes	78%

Again, the differences in correlations are very small and the above results denote the fact that movies popularity are highly correlated to both ratings, regardless what is the training set of the adjectives sentiment.

5.2.2 Evaluation of Rotten Tomatoes critics

Rotten Tomatoes reviews (written by both professionals and audience) were also used to determine movies popularity. In this case, the adjectives' sentiment estimation trained on Rotten Tomatoes was used. In figure

5.9 the correlation between the sentiment estimation of professional critics and the professional Rotten Tomatoes ratings, along with the correlation to the IMDb ratings (figure 5.10) is shown. Additionally, figure 5.11 and 5.12 presents the sentiment estimation of audience critics compared to the audience Rotten Tomatoes ratings and to the IMDb ratings, respectively.

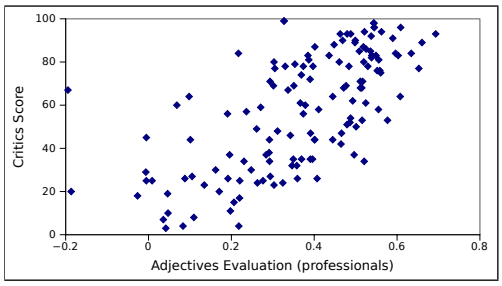


Figure 5.9 – Rotten Tomatoes professional critics compared to RT professional ratings - R^2=67%

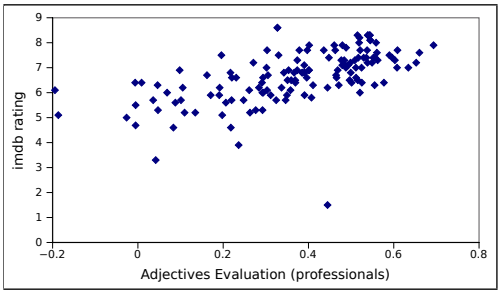


Figure 5.10 – Rotten Tomatoes professional critics compared to IMDb ratings - R^2=58%

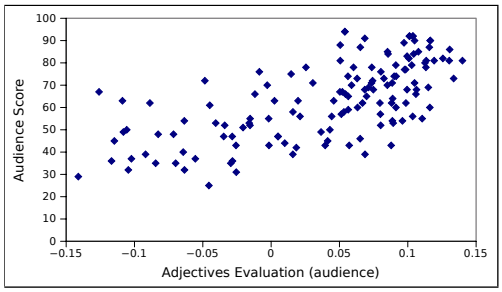


Figure 5.11 – Rotten Tomatoes audience critics compared to RT audience ratings - R^2=66%

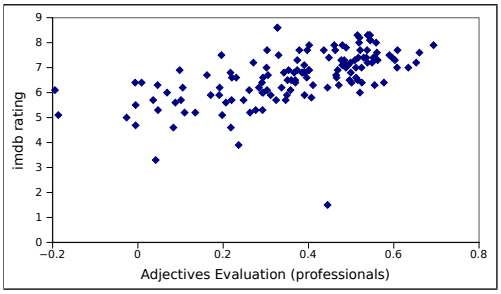


Figure 5.12 – Rotten Tomatoes professional critics compared to IMDb ratings - R^2=58%

According to the above results, there is again a high correlation between sentiment estimation of Rotten Tomatoes reviews and the above mentioned ratings. However, it is shown that the presented estimations are correlated in a higher degree to the Rotten Tomatoes rating than to the IMDb ones. While the reviews from Rotten Tomatoes are a lot shorter than the reviews of IMDb, they are expected to contains less adjectives. Therefore, the sentiment estimation towards a movie is based on less input data than with IMDb critics, and this might reduce accuracy in evaluating peoples' opinion for the particular movie. On the other hand, IMDb critics which are usually 3-4 paragraphs long, contain plenty of adjectives for describing any movie, and therefore these texts generate a more accurate estimation towards both

ratings.

5.2.3 Evaluation of Twitter posts

Below we present our results after analyzing collected tweets that contain the movies' titles we are working on. Since Twitter has no specific structure of its posts, unlike IMDb and Rotten Tomatoes, it is crucial to make sure that the posts are indeed referring to the specific movies, and moreover that the adjectives used are also related to them. For this reason we removed all duplicates created by retweets and we captured all adjectives connected to film critics with the use of the hybrid algorithm we described.

The retrieval of posts concerning particular movies in Twitter, is possible only by searching for tweets containing their titles, contrary to movie reviews websites like IMDb and Rotten Tomatoes. In some cases, however, this search returns plenty of irrelevant tweets, which even though contain the movie title, they have nothing to do with the actual movie. This case is very often with titles that contain common words, which are usually used in other context apart from the movie itself.

In order to determine such titles, we measured the number of results (in millions), that the Google search engine returned when searching for a movie title and when searching for the same movie tile plus the word 'movie'. That way, we are aiming to specify what is the percentage of the returned results from the first query, that are actually referring to the corresponding movie. The results of this measurement are shown in the below table, where, the first column contains the results number of querying the movie title and the second column represents these numbers when putting word 'movie' in the query:

Movie title	Number of results	Number of results (plus the word 'movie')	Percentage of results referring to the movie
King Arthur	140	21,3	15%
Hugo	458	83,6	18%
Anonymous	851	197	23%
Rise of the Planet of the Apes	27,2	26,5	97%
The Green Hornet	12	11,3	94%
Scream 4	32,9	27,1	84%

The first three titles are actually names that do not characterize specif-

ically the movie, but are widely used for irrelevant subjects. This fact, completely justifies the observation that the number of results returned after the first search is surprisingly higher than the second one. On the other hand, the last three titles are more unique and are usually used when referring particularly to the corresponding movies, thus the resulted percentage is much higher. These Google searches are pretty much the same procedure as Twitter searches and therefore some of the movies with ambiguous titles had to be excluded from this testing. Following the above, 72 movies with appropriate titles were carefully selected .

The tweets on these selected movies, were evaluated by using the adjectives sentiment derived by Twitter (fig. 5.13), IMDb (fig. 5.14) and the ground truth (fig. 5.15), and were compared to the IMDb rating.

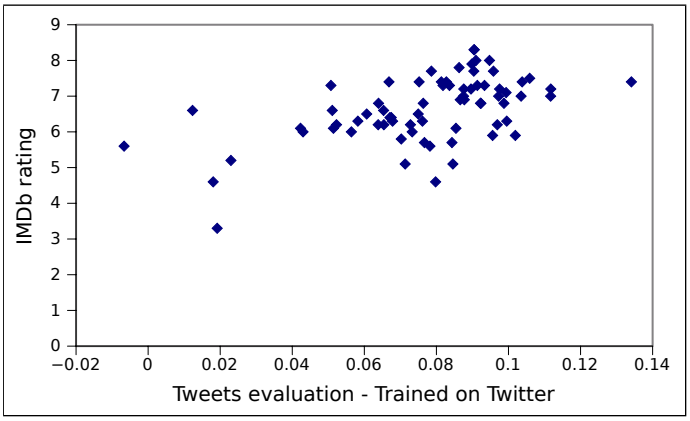


Figure 5.13 – Trained on Twitter compared to IMDb ratings - R^2=54%

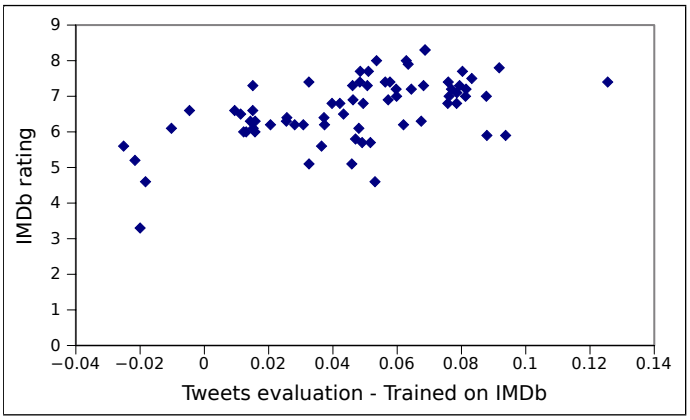


Figure 5.14 – Trained on IMDb compared to IMDb ratings - R^2=56%

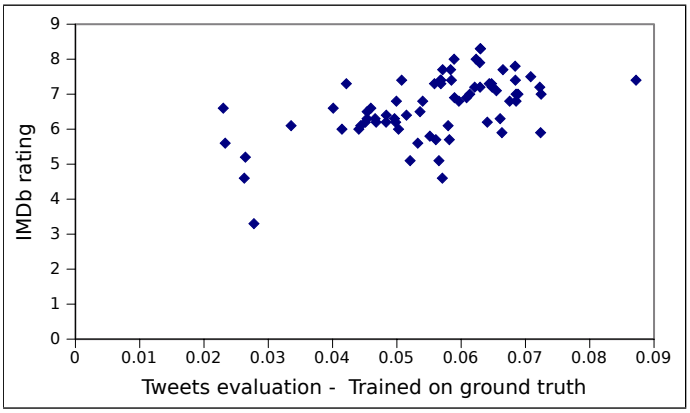


Figure 5.15 – Trained on ground truth compared to IMDb ratings - R^2=55%

It can be seen that again there is a strong correlation in all plots, even though Twitter do not aim to provide actual critics on movies, and prevents the posting of descriptive and detailed posts due to its limitation of 140 characters per tweet. This means that the Twitter short posts contain less adjectives compared to IMDb and Rotten Tomatoes reviews and provide a much shorter description towards the discussed movie. Apparently, this can reduce the accuracy of the popularity estimation performed on such kind of short text.

Additionally, the retrieval of relevant tweets can be also tricky and create problems too. As already mentioned, it is possible that the test set may contain posts irrelevant to the subject, even though they include the title of a movie. Additionally, some movies' titles can be written in various forms and abbreviations. For example the movie "Pirates of the Caribbean: The Curse of the Black Pearl" can be written as "Pirates of the Caribbean", "Pirates of Caribbean" or just "Pirates" and can be even misspelled. This make it hard to ensure that all tweets referring to a particular movie are included to the set, which can lead to less accurate estimations.

Despite all these obstacles, the correlation rates derived by movies' popularity estimation from Twitter posts to the IMDb ratings, are still sufficient and show that this evaluation offers a good approach to the actual popularity of movies. It is therefore clear that this technique is much more trustful that simply counting Twitter posts, and it would have lead to even better results if a way of efficient retrieval of Twitter posts was established.

5.2.4 Correlation with Box office

Finally, the correlation between all generated movies' sentiment estimation and box office gross of the corresponding films, was also examined. Our clear conclusion is that there is no relation between adjectives evaluation and box office revenues, which can be fully justified with a quick look of figures 5.16, 5.19 and 5.18 representing the correlation of IMDb, Rotten Tomatoes and Twitter posts' estimation towards the worldwide box office gross. Apart from that, we also compared box office success of each movie with ratings obtained from the IMDb website (figure 5.17), which is another proof of the lack of correlation between critics' ratings and commercial success. This point was also presented by a recent study by Wong, Sen and Chiang [38], related to the difficulty of predicting the Box Office revenues through Twitter.

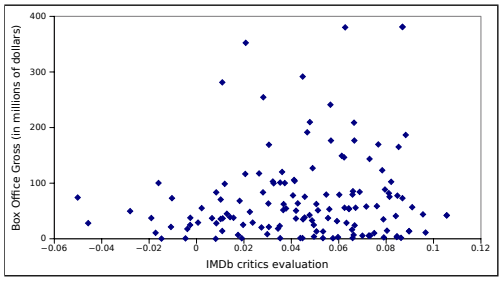


Figure 5.16 – IMDb critics compared to worldwide Box Office gross - R^2=11%

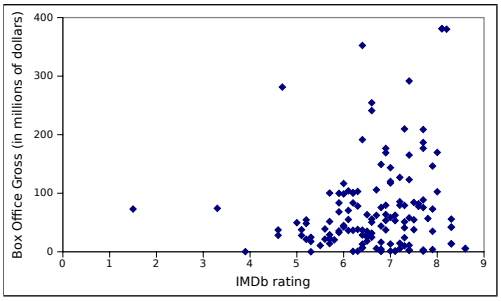


Figure 5.17 – IMDb ratings compared to Box Office gross - R^2=17%

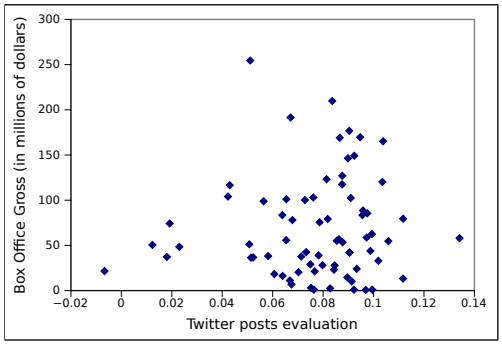


Figure 5.18 – Twitter posts compared to worldwide Box Office gross - R^2=5%

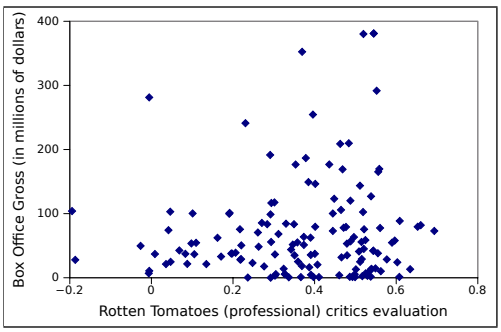


Figure 5.19 – Rotten Tomatoes critics compared to worldwide Box-Office gross - R^2=9%

5.3 Results analysis and Evaluation

From the above plots, it is shown that the choice of an adjectives' sentiment estimation training on either IMDb, Rotten Tomatoes or Twitter does not noticeably effect the correlation of the movies popularity to a rating. In the previous chapter it was found that the above mentioned sources generate some slight differences to the calculated adjectives sentiment, mostly because of the different writing styles of their posts. However, these differences are eliminated when the estimated adjectives' sentiments are used for the evaluation of a high number of texts. This fact, denotes the flexibility of the proposed algorithm while the training set can be any text related to the discussed subject, regardless its length, writing style or vocabulary.

Additionally, the above measurements show that the evaluation of IMDb posts leads to a higher correlation with both IMDb and Rotten Tomatoes ratings, contrary to the evaluation results of Rotten Tomatoes and Twitter posts. This is mainly because of the different style of reviews available on the IMDb website, which consist of much longer and more detailed texts. On the other hand, Rotten Tomatoes usually includes texts of one or two sentences, while Twitter consists of posts at most 140 characters long.

Figure 5.20, shows the number of positive and negative adjectives found in texts of IMDb, and figures 5.21 and 5.22 for Rotten Tomatoes and Twitter posts, respectively. The IMDb posts are shown to clearly contain more adjectives than Rotten Tomatoes and Twitter, mainly due to their length and writing style. Therefore, one can say that IMDb offers a more precise indication of the people's opinion on a movie, while it includes more descriptive, detailed texts. For this reason, the movie sentiment estimation derived by IMDb critics reach a higher precision level when compared with real film ratings, than the estimation derived by the two other sources.

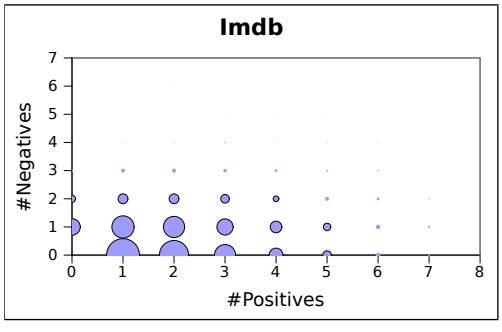


Figure 5.20 – Number of positive/negative adjectives found in IMDb critics

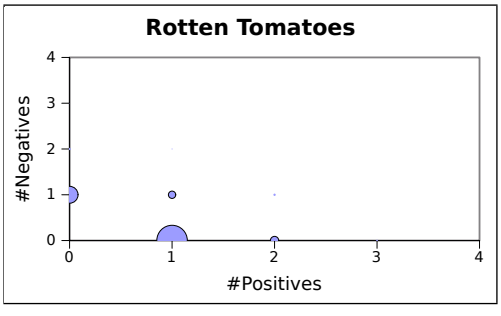


Figure 5.21 – Number of positive/negative adjectives found in Rotten Tomatoes critics

On the other hand, Rotten Tomatoes and Twitter results had also a strong correlation with official ratings. Specifically, Twitter faces an additional obstacle compared to Twitter and IMDb, due to its difficulty into retrieving all relevant posts and eliminating irrelevant ones. However, it still reaches a high correlation to all ratings, and proves that the proposed popularity estimation technique is much more efficient that naive posts counting. It is therefore, undoubted that a sentiment estimation technique, as the one that was just presented is necessary for the efficient evaluation of critics and of subjective texts in general.

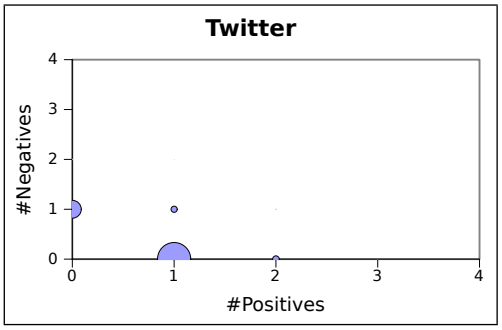


Figure 5.22 – Number of positive/negative adjectives found in Twitter posts

Chapter 6

Conclusion and Future work

6.1 Summary

In this thesis, a new classifying technique based on NLP analysis of the text, is proposed. This new hybrid classifier uses the relations between words as derived by the syntax analysis performed, and aims to trace particular subjective patterns that ensure the existence of a user's opinion upon a subject.

This proposing method has the advantage of getting deeper into the actual meaning of a text, whereas the usual classifying tools, which are mostly based on machine learning algorithms (Naive Bayes, Maximum Entropy, SVM), are seeking the existence of specific words with strong sentiment in a post. However, this approach is proven to be much less accurate than the proposed one, mostly because the use of these words are often completely unrelated to the subject of the post, or are even used in an entirely different sense. Having a careful look at the structure of the text, though, gives us the opportunity to have a safer conclusion about its subjectivity, and as shown, it contributes a 40% gain on accuracy and about 85% of correct classified posts.

Apart from detecting the presence of subjective or objective pattern in a text, the establishment of the polarity and especially of the magnitude of the sentiment found in subjective texts is also a crucial matter discussed in this work. An adjective-based approach is explained in this work, which extract people's sentiment on this particular subject by estimating the overall sentiment of their subjective posts on that. Initially a sentiment magnitude metric is assigned to a number of adjectives, according to the number of times that they occur in positive and negative labeled texts, and afterwards these values are used to evaluate entire posts and consecutively, whole subjects.

Comparing sentiment estimations to various ratings derived by real people's voting, a strong correlation was observed between them at most of the cases, which denotes that this approach can be highly successful.

6.2 Future Work

Even though, the proposed algorithms appeared to have high accuracy, there are some points of improvement which can lead to even more precise estimations. Bellow there are some point of improvement that are not sufficiently resolved in the present thesis, and would have contributed to the overall accuracy of the proposed methods.

6.2.1 Efficient tweets retrieval

A common problem of Twitter data analysis in general, is the selection and retrieval of posts that are indeed relevant to the searching terms. Since Twitter, implements a concept of totally free writing without a specific subject or rules (contrary to IMDb or Rotten Tomatoes which were also used in this study), it is definitely hard to make sure that all selected tweets over a particular subject, contain no irrelevant, needless posts. However, the use of hashtags can categorize posts to some degree, but again not everyone uses them and moreover, because of their loose nature, there is no safe way to determine all words or phrases used as hashtags on a specific matter. Improving the way of collecting Twitter, can improve the quality of information one can get out of them, and thus their overall evaluation in terms of content or sentiment would be improved. Apart from that, Twitter contains a lot of advertising and spamming posts which can distort the real results of a popularity estimation. Tracing such posts, would be another important step into improving classifying and estimating success of both proposed algorithms.

6.2.2 Tweets content classification

As for the hybrid classifier, which was presented in a previous chapter, the major reason for incorrect classifications was undoubtedly because of spelling and syntax mistakes that existed in many posts of Twitter. Establishing a way to correct, or at least point out texts containing such errors, would definitely result in more accurate classifications while the NLP analysis would be done correctly and with less ambiguities. Apart from that, tweets often contain slang terms and common phrases which usually created errors in the

generation of syntax dependences, on which the classifier relies. Recognizing such common phrases and other terms used widely in the internet (such as acronyms, smilies, slang etc), would help into having a clearer perception of what is their role and meaning in a sentence, and consequentially we would be able to extinguish such problematic cases.

6.2.3 Tweets sentiment estimation

The sentiment estimation approach has also plenty of open subjects, which can be examined for the improvement of this technique. Fist of all, the sentiment evaluation of more adjectives should definitely result in more accurate sentiment magnitude evaluation of posts and therefore more precise sentiment estimation of a particular subject. The selection of adjectives should be however, careful and related to the specific domain that is being examined. It is possible that a specific adjective can carry much heavier sentiment for a specific case than for another, and therefore such selection should necessarily contains such important adjectives. This meaning differentiation among different subjects, types of writings or even author ages, is also another interesting study field, that can give us a clearer idea of how sentiment magnitude is scaling between all these different cases and would help us produce more precise estimations.

Besides taking into account as many adjectives as possible into establishing a sentiment estimation of posts, considering other grammatical terms, such as verbs, would also contribute to more precise text analysis. Verbs like 'love', 'hate' or 'adore', undoubtedly contain some kind of emotion that gives us a clearer idea of what is the author's opinion on a particular matter, and therefore, should be included in an overall evaluation of its sentiment. However, the selection of verbs to be considered should be again careful, in order to avoid the case where irrelevant verbs (or other terms), unrelated to the subject, distort the final estimations.

Bibliography

[1] Alexa: The web information company. http://www.alexa.com/. visited on 23-06-2012.

[2] Facebook. http://www.facebook.com. visited on 23-06-2012.

[3] Google +. http://plus.google.com. visited on 23-06-2012.

[4] Imdb: Internet movie database. http://www.imdb.com. visited on 23-06-2012.

[5] Lingpipe. http://alias-i.com/lingpipe. visited on 23-06-2012.

[6] Rotten tomatoes. http://www.rottentomatoes.com. visited on 23-06-2012.

[7] Tweet sentiments. http://tweetsentiments.com. visited on 23-06-2012.

[8] Twitter. http://www.twitter.com. visited on 23-06-2012.

[9] Wordnet: A lexical database for english. http://wordnet.princeton.edu/. visited on 23-06-2012.

[10] S. Asur and B. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[11] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[12] N. Blenn, K. Charalampidou, and C. Doerr. Context-sensitive sentiment classification of short colloquial text. In R. Bestak, L. Kencl, L. E. Li, J. Widmer, and H. Yin, editors, *Networking (1)*, volume 7289 of *Lecture Notes in Computer Science*, pages 97–108. Springer, 2012.

[13] E. Cambria, A. Hussain, C. Havasi, and C. Eckl. Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems. *Development of Multimodal Interfaces: Active Listening and Synchrony*, pages 148–156, 2010.

[14] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[15] R. Chunara, J. Andrews, and J. Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.

[16] R. Cilibrasi and P. Vitanyi. Automatic meaning discovery using google. *Manuscript, CWI*, 2004.

[17] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *Proc. 5th Intl. Conference on Weblogs and Social Media*, 2011.

[18] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[19] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M. yen Kan, and K. R. McKeown. Simfinder: A flexible clustering tool for summarization. pages 41–49, 2001.

[20] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997.

[21] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. *Proc. 49th ACL: HLT*, 1:151–160, 2011.

[22] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, pages 3–10. MIT Press, 2002.

[23] P. Kolb. Disco: A multilingual database of distributionally similar words. 2008.

[24] C. Miller. Sports fans break records on twitter, 2010. http://bits.blogs.nytimes.com/2010/06/18/sports-fans-break-records-on-twitter/, visited on 16-06-2012.

[25] G. Mishne and N. Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[26] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[27] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet: : Similarity - measuring the relatedness of concepts. In D. L. McGuinness and G. Ferguson, editors, *AAAI*, pages 1024–1025. AAAI Press / The MIT Press, 2004.

[28] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet: : Similarity - measuring the relatedness of concepts. In D. L. McGuinness and G. Ferguson, editors, *AAAI*, pages 1024–1025. AAAI Press / The MIT Press, 2004.

[29] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453. Morgan Kaufmann, 1995.

[30] W. P. Review. Tunisia: Twitter revolution vs. twitter impeachment, 2011. http://www.worldpoliticsreview.com/trend-lines/7584/tunisia-twitter-revolution-vs-twitter-impeachment, visited on 16-06-2012.

[31] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[32] N. Shuyo. Language detection library. http://code.google.com/p/language-detection, 2010. visited on 16-06-2012.

[33] Spiegel. 'twitter revolution': Fearing uprising, russia backs moldova's communists, 2009. http://www.spiegel.de/international/europe/twitter-revolution-fearing-uprising-russia-backs-moldova-s-communists.html, visited on 16-06-2012.

[34] T. W. Times. Editorial: Iran's twitter revolution, 2009. http://www.washingtontimes.com/news/2009/jun/16/irans-twitter-revolution, visited on 16-06-2012.

[35] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003.

[36] P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[37] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, pages 486–497, 2005.

[38] F. M. F. Wong, S. Sen, and M. Chiang. Why watching movie tweets won't tell the whole story? *CoRR*, abs/1203.4642, 2012.

[39] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.