A Mini-Project Report

on

USING INFORMATION RETRIEVAL FOR SENTIMENT POLARITY PREDICTION

carried out as part of the course Information Retrieval (IT362)

Submitted by

Samyak R Jain (14IT240)
Bhat Aditya Sampath (14IT209)
Rakshitha K N (14IT136)
VI Sem B.Tech (IT)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Department of Information Technology
National Institute of Technology Karnataka, Surathkal.

Jan - May 2017

CERTIFICATE

This is to certify that the project entitled "Using Information Retrieval for Sentiment Polarity Prediction" is a bonafide work carried out as part of the course **Information Retrieval (IT362)**, under my guidance by

1. Samyak R Jain (14IT240)
2. Bhat Aditya Sampath (14IT209)
3. Rakshitha K N (14IT136)

students of VI Sem. B.Tech (IT) at the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, during the academic semester of Jan-May 2017, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, at NITK Surathkal.

Place: NITK, Surathkal

(Signature of the Guide)

Date: 16-04-2017

DECLARATION

We hereby declare that the project entitled "Using Information Retrieval for Sentiment Polarity Prediction"submitted as part of the partial course requirements for the course Information Retrieval (IT362) for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal during the Jan-May 2017 semester has been carried out by us. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Further, we declare that we will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Instructor.

Name and Signature of Students:

1. Samyak R Jain (14IT240) _____
2. Bhat Aditya Sampath (14IT209) _____
3. Rakshitha K N (14IT136) _____

Place: NITK, Surathkal

Date: 16-04-2017

Abstract

Twitter is one of the social networks used by millions of people to express their opinions on a diverse range of topics. Thus, social media like Twitter are examined periodically by sentiment analysis algorithms which classify the tweets (posts) as containing positive or negative sentiment. Since the variety of posts is vast and length of the posts is short, the machine learning approach by using the words as features results in poor sentiment polarity prediction. In this project, we use features derived from the ranking generated by an Information Retrieval System in response to a query consisting of the post that needs to be classified as one containing positive or negative sentiment.

This system has very few features as compared to the other approach which uses the words as features resulting sparse vectors, thus, not requiring expensive resources. The idea is to use the ranking generated by the information retrieval system to extract information about the class of the similar posts.

Contents

1 Introduction **1**

2 Literature Survey **2**

2.1 Background 2

2.2 Outcome of Literature Survey 2

2.3 Problem Statement 3

2.4 Objectives 3

3 Methodology **4**

3.1 Information Retrieval System 4

3.2 Feature Generation 5

3.3 Classification 6

4 Implementation **7**

4.1 Work Done 7

4.2 Results and Analysis 8

4.3 Innovative Work 8

4.4 Individual Contribution 8

5 Conclusion and Future Work **9**

References **10**

List of Figures

1 An overview of SABIR 4

1 Introduction

With millions of posts everyday, Twitter has evolved to become a major forum for expressing personal opinions on a variety of topics. Because of its popularity, this social media service has been the target of a number of research studies from a broad range of areas including Psychology, Sociology, Marketing, and Computer Science.

Sentiment analysis, also called Opinion Mining, is dedicated to the computational study of opinions and sentiments expressed in text. This area of research has been gaining increasing attention from the research community. Out of the different aspects of opinions that can be studied, the polarity of sentiments is the most well investigated. It consists in predicting whether the opinion expressed in the text is positive or negative.

While most of the research concentrates on product reviews, of late, a number of studies on Twitter posts (tweets) have emerged. Sentiment Analysis on Twitter can be done at three different levels: (i) entity, (ii) tweet, or (iii) expression. Entity-level analysis deals with finding the overall opinion about an entity or topic, tweet-level analysis detects the polarity of individual tweets, and expression level analysis deals with specific phrases within a tweet. Our focus is on the second – tweet-level analysis. The added challenge of analysing tweets (compared to product reviews) is their shorter length – at most 140 characters – which results in very sparse vector representations. In addition, the variety of topics, and the informal vocabulary, characterised by slangs, abbreviations, and misspellings, pose added difficulties to its computational treatment.

Other successful approaches for the problem of polarity classification on tweets use one or more of the following: resources such as lexicons (which are sometimes manually created), expensive pre-processing such as part-of-speech (POS) tagging, numerous features, large amounts of training data, and complicated machine learning methods such as classifier ensembles. In this work, we propose a method called Sentiment Analysis Based on Information Retrieval (SABIR) which uses none of the above. We show that classification accuracy comparable to the state-of-the-art can be achieved with a single classification algorithm using only 24 features. Contrary to the prevalent approaches, we do not use the words of the posts as features. Our features are derived from the ranking generated by an Information Retrieval System in response to a query q which consists of the tweet that we wish to classify as positive or negative sentiment. The ranking has the n most

similar tweets for which we already know the class in decreasing order of similarity to the unlabelled tweet q. The idea is to extract information of the class of the similar posts to classify q.

2 Literature Survey

2.1 Background

Sentiment analysis refers to the use of natural language processing, text analysis, and other methods to identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

2.2 Outcome of Literature Survey

The survey in [5] covers techniques and approaches that promise to directly enable opinion-oriented information gathering systems. Their focus was on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis.

[4] emphasizes on the fact that it is necessary to see the amount of research being done in the field of Sentiment Analysis after the Twitter boom. It aims to be the starting point for those investigations concerned with the latest references to Twitter.

In [6], the authors emphasize on the importance of contextul information in the sentiment analysis process. What tweets lack in structure, they make up in sheer volume and metadata. This metadata includes geolocation, temporal and author information. They used this data to analyse the variation of tweet sentiments across different authors, times and locations.

[2] says that Twitter messages are being used to determine consumer sentiment towards a brand. In their research, they introduce an approach to supervised feature reduction using n-grams and statistical analysis to develop a Twitter-specific lexicon for sentiment analysis. They show that the reduced lexicon set, while significantly smaller, reduces

modeling complexity, maintains a high degree of coverage over the Twitter corpus, and yields improved sentiment classification accuracy.

Carvalho, Prado, and Plastino (2014) [1] applied a genetic algorithm to select the paradigm words that will help determine the polar- ity of other words. They found an improvement compared to ap- proaches in which the paradigm words are fixed.

2.3 Problem Statement

One of the main driving forces for research in Sentiment Analysis on Twitter data is the endless applications where Sentiment Analysis is useful. Some of these applications include tracking customer reviews, survey responses, competitors, etc. It is also practical for use in business analytics and situations in which text needs to be analyzed. More research in this domain would mean improvement in the accuracy of sentiment analysis systems. Organizations will readily and confidently incorporate sentiment analysis systems in their products if there are state of the art systems which give good practical accuracy.

The paper we are implementing [3] incorporates a completely different and novel approach to the problem of sentiment analysis. Broadly, this paper discusses the working and performance of a system called Sentiment Analysis based on Information Retrieval (SABIR). SABIR is primarily composed of two steps:

1. Obtain the n most similar posts (tweets) in relation to the tweet we wish to classify.
2. Use the information about these n posts as features to train a supervised classifier.

2.4 Objectives

The main goal (objective) of this project is to perform sentiment analysis on twitter data using an Information Retrieval system. We try our hand at multiple approaches to creating an information retrieval system. Also, we test our system using a number of classifiers and various datasets to retrieve a comparison of the various approaches. Finally, we compare the results obtained with existing state-of-the-art approaches to Sentiment Analysis to see where our approach stands. Hence, our main objective is to develop a sentiment analysis system based on information retrieval and evaluate its performance by comparing its accuracy with existing state of the art systems.

3 Methodology

As mentioned in the Problem Statement, SABIR is composed of two main steps:

1. Obtain the n most similar posts (tweets) in relation to the tweet we wish to classify.

2. Use the information about these n posts as features to train a supervised classifier.

More formally, our goal is: given a set of tweets $T = \{t_1, t_2, ..., t_m\}$ for which the class $c_i \epsilon \{positive(+), negative()\}$ is known and a set of unlabelled tweets $Q = \{q_1, q_2, ..., q_p\}$ (i.e., for which the class is unknown), we use information about the similarity of each element $q_i \epsilon Q$ in relation to the elements $t_j \epsilon T$ to predict the class of q_i. Information on the similarity of the tweets is taken from an Information Retrieval System as shown in Figure 1.

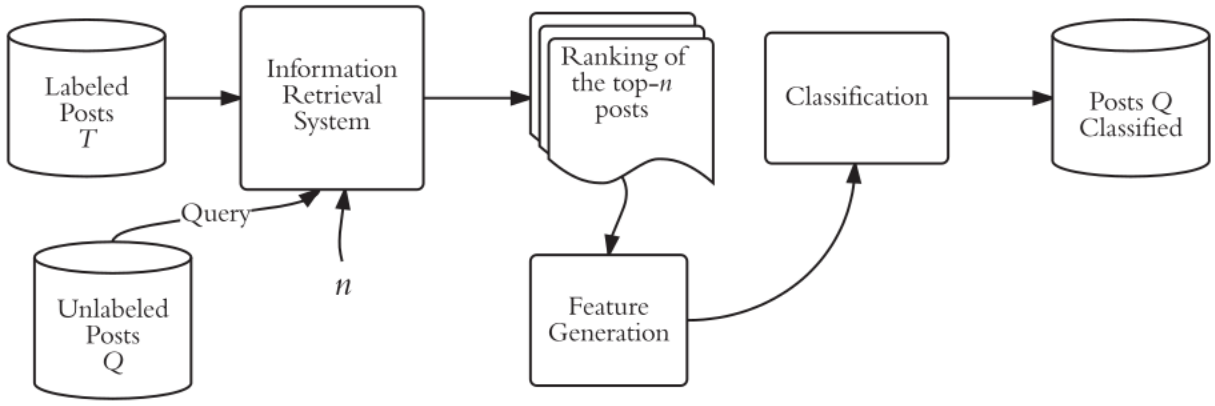

Figure 1: An overview of SABIR

3.1 Information Retrieval System

Information Retrieval Systems (IRS) apply similarity functions to rank a set of items (usually textual documents) in response to a user query. In this work, an IRS is used to index the labelled tweets in T, and rank them in relation to each unlabelled tweet in Q. The ranked list with the n highest scoring items for each q_i serves as a source of features which will later be used by a classifier. If a post q_i is closest to positive posts, then it is

likely to be positive too. The idea is analogous to the principle behind a k-NN classifier in which a test instance is classified as belonging to the majority class of its k-closest neighbours. An extremely simple but naïve approach would be taking the prevalent class in the ranking as the class of the post. However, this evidence alone is not enough. Thus, our method extracts different sources of evidence from the ranking of the n most similar posts in relation to q. SABIR considers a variety of attributes ranging from simple counts to more elaborate metrics that combine the rank and the score of the retrieved results.

The first step is to index the labelled posts (T) using an IRS. Then, each unlabelled post $q_i \epsilon Q$ is used as query and the n most similar posts are retrieved and ranked in decreasing order of similarity. In this project, we employ the widely used Okapi BM25 algorithm to generate the ranking. Although other metrics could have been used, we opted for BM25 due to its good results in IR experiments.

3.2 Feature Generation

A total of 24 features are extracted from the ranking (12 for each class). Notice that only one ranking is necessary to produce the 24 features. Aggregation functions such as average, max, min, sum, and count are used for each class (positive and negative).

In addition to these aggregation functions, we derived another feature, called ϕ , that takes into consideration the absolute and the relative (i.e., in relation to the class) ranks of each item. It is given as:

$$\phi_c = \sum_{r=1}^{n_c} \frac{rank_{rel}}{rank_{abs}}$$

where c is the class; n_c is the number of retrieved posts for that class, $rank_{rel}$ is the relative position of the post among the posts of that class, and $rank_{abs}$ is the absolute position of the post in the overall ranking.

The remaining attributes are a combination of the metric ϕ and the aggregation functions. They are: $\phi_{avg_c} = \frac{\phi_c}{avg_c}$, $\phi_{max_c} = \frac{\phi_c}{max_c}$, $\phi_{min_c} = \frac{\phi_c}{min_c}$, $\phi_{sum_c} = \frac{\phi_c}{sum_c}$, $\phi_{count_c} = \frac{\phi_c}{count_c}$. The aggregation functions avg_c, min_c, max_c and sum_c compute the average, minimum, maximum and sum of scores for class c, respectively; and, $count_c$ is the number of posts from class c in the ranking. The final feature, $\phi_{c_{positional}}$, combines the scores of the re-

trieved posts and their positions in the ranking. It is given as:

$$\phi_{c_{positional}} = \sum_{r=1}^{n} \left(\frac{rank_{rel}}{rank_{abs}} \right) \times S(q_i)$$

where $S(q_i)$ is the score for the post q_i.

An important aspect of creating a classification system using IR based features is that the dimensionality is greatly reduced and, at the same time, classification quality is improved. Rather than using the words themselves as features, which would lead to several thousand features even in small datasets, we are able to classify with high accuracy using only 24 attributes. Although our features have no linguistic meaning, the intention is that they should capture latent discourse and semantic properties of the text while Okapi BM25 captures the lexical properties.

3.3 Classification

The second stage in SABIR involves using a supervised machine learning algorithm. Based on labelled training instances, a learning model is generated. The goal is to obtain a model that is able to generalise and classify unseen data. Since it is well-known that there is no single machine that is better than all the others on all tasks, a number of classification algorithms were tested to identify the ones that are best suited to our attributes.

4 Implementation

4.1 Work Done

Datasets. In order to test our approach, we ran experiments using two datasets composed of real tweets, which have also been applied to related works -

- **Stanford-Twitter Sentiment corpus (STD).** This dataset is divided into two subsets - STD-train and STD-test. STD-train is composed of 1.6 million automatically labelled tweets. The automatic labelling method is known as noisy-labelling, and it basically consists in assigning a sentiment to the tweet based on the sentiments associated to the emoticon present in the tweet. This method eliminates the manual effort in labelling, but the resulting dataset contains noise. Tweets with irony and sarcasm, for example, could be misclassified. STD-test, on the other hand, has 359 tweets which were manually labelled.

- **Stanford Sentiment Gold Standard (STS-Gold).** This dataset contains tweets for which three annotators have agreed whether the sentiment label was positive or negative.

Preprocessing. The raw text in the tweets was processed as follows: the tweets are broken down into tokens, URLs from tokens are removed, punctuation and numbers are removed, stemming is done and stopwords are removed. All the tweets present in the testing and training dataset are preprocessed in the same way.

Information retrieval system. We created our own Information Retrieval system. We used Okapi BM25 to generate the rankings for the queries. The similarity score between tweets t_j and q_i is calculated using the following formula-

$$BM25(q_i, t_j) = \sum_{w \epsilon q_i} log\left(\frac{m - f_w + 0.5}{f_w + 0.5}\right) \times \frac{\big(k_1 + 1\big) f_{w,t}}{K + f_{w,t}}$$

where w is a word in the tweet, m is the number on indexed tweets, f_w is the number of occurrences of word w in the dataset, and $f_{w,t}$ is tthe number of occurrences of w in t_j. K is $k_1\big(\big(1-b\big) + b \times \frac{L_t}{avgL}\big)$ where L_t is the length of t_j and $avgL$ is the average length of

the indexed tweets. The default values for the constants k_1 and b were used (1.2 and 0.75 respectively).

Classification Algorithms. We used the classifiers implemented in Scikit-learn according to their default parameters. Our results have shown that the best algorithms differed across datasets.

Performance Evaluation. The standard evaluation metrics like accuracy, precision, recall and F-1 Score are calculated for all datasets and all classification algorithms. The results are discussed in the next section.

4.2 Results and Analysis

(Algorithm still running, results not yet computed.)

4.3 Innovative Work

4.4 Individual Contribution

5 Conclusion and Future Work

SABIR performs well to classify tweets as containing positive or negative sentiment without the need for expensive resources. It uses merely 24 features to classify the tweets, which are solely based on the ranking generated by an IR system. A number of alternatives are open for exploration:

- The number of results to retrieve (n) is a parameter of SABIR, and can be varied to obtain different observations.
- Ranking functions (other than BM25) can be explored and/or combined.
- SABIR can also be explored and used for other sentiment classification datasets, apart from Tweets.

References

[1] J. Carvalho, A. Prado, and A. Plastino. A statistical and evolutionary approach to sentiment analysis. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 110–117, Aug 2014.

[2] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266 – 6282, 2013.

[3] Anderson Uilian Kauer and Viviane P. Moreira. Using information retrieval for sentiment polarity prediction. *Expert Systems with Applications*, 61:282–289, 2016.

[4] Eugenio Matinez-Camara, M. Teresa Martin-Valdivia, L. Alfonso Urena-Lopez, and A Rturo Montejo-Raez. Sentiment analysis in twitter. *Natural Language Engineering*, 20(1):1–28, 2014.

[5] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[6] Soroush Vosoughi, Helen Zhou, and Deb Roy. Enhanced twitter sentiment classification using contextual information. *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2015.