Ain Shams University
Faculty of Computer and Information Sciences
Information systems Department

Time Tagging for Enhancing Opinion Mining Prediction

Submitted in partial fulfillment of the Requirements for the Degree of master

By

Ghada Hafez Hassan Diab

B.Sc in Information Systems,
Faculty of Computer and Information Sciences,
Helwan University

Supervised By

Professor. Omar.H.Karam

Professor , Information System Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Rasha Isamil

Associate Professor, Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Cairo 2019

Abstract

Nowadays, opinion mining becomes one of the most important fields and it attracts the interest of many researchers. The 'electronic Word of Mouth' (eWOM) statements that are expressed on the web, are important for business and service industry to enable customers share their point of view. In the last one and half decades, research communities, academia, public and service industries are working rigorously on opinion mining -which is also called, sentiment analysis to identify and categorize opinions from a piece of text. One key use of sentiment analysis is to extract and analyze public moods and views. Researchers used sentiment analysis in different ways. For example, to determine the market strategy that improve customer service.

One of the key challenges of sentiment analysis is how to extract temporal synsets from text. Temporal synsets may be events, dates, times, or even Explicit lyrics. Tempowordnet is one of the attempts to building a lexicon that may help in finding temporal synsets.

It is noticed that temporal word net contains 117598 word, 100512of them are classified atemporal, which means it doesn’t have any temporality in it. 12053of those atemporal words were found to be verbs, and this shows that they were wrongly classified to be atemporal. From this point we developed our system to work on those atemporal verbs and classify them.

This thesis presents a framework for enhancing opinion mining process. The framework presents two parts: The first part is for discovering temporal verb references (future, past and present) from opinions and using them to build accurate prediction models. The proposed framework improved the percentage of discovering (past, present and future verbs)over the tempowordnet.To enhance existing methods and make them more efficient, an algorithm that targets verbs is proposed. It extracts verbs based on tempowornet and classified them as past, present, and future. Experimental results showed that The accuracy for the proposed framework was 0.7%, 1.1%, 0.2%for the past, present and future respectively over that of the tempowornet.

The second part introduced in the proposed framework is enhancing the extraction of aspects and sentiment words by detecting shortcuts, emojis and sarcasm based on time tagging. The proposed framework are applied to a mobile reviews data set and compared to different systems based on theaccuracy, recall, precision and F1 measure. The results showed that our proposed work improved The accuracy of the proposed framework is 0.23%, 0.4% , 0.14 %and 0.24 %for total past, present , future and atemporal respectively over that of the compared systems, also improved the recall, and precision and f1 measure.

Acknowledgment

What a journey!

First and foremost, all praise and thanks be to almighty Allah, the Most Gracious, the Most Merciful, for giving me the strength and determination to finish this work.

I first thank and express my sincere gratitude to my supervisor, Dr. Rasha Ismail, whose ultimate support, encouragement, and guidance helped me to develop my thinking as a scientist. I thank her for believing in me and encouraging me to publish in good places.

I would like to express my gratitude too to my supervisor, Prof. Omar Karam, for his support and gentle guidance during my research.

Words cannot express my gratitude to my family. I thank my mother for her prays, support, and bearing me during this time ,my father for his support and my husband for his support especially financially and time

I would like to thank my friends and sisters Soha Ahmed Ehssan and Naja Ahmed for their support and encouraging.

I thank also my colleagues and professors who helped me through my journey Thanks to them all I am really virtue to have them in my life.

Table of Contents

List of Tables

List of figures

Chapter 1

Introduction

The amount of available information nowadays makes actual systems more concerned with how to handle information overload and ensure that the user will have access to the best sources with the least effort. In recent years, special attention was given also to the amount of produced user-generated content. The e-commerce sector is one of the most affected by the amount of data produced by customers, which increased dramatically during the phase known as Web 2.0. Customers’ opinions represent a valuable unique type of information which was given much attention by the research community. Thus, this work emphasizes the need of special mechanisms that aim to provide the community with better methods to take full advantage of this data.[9]

From the customer's perspective, considering others' opinions before purchasing a product is a common behavior long before the existence of the Internet. In the era of the digital world, the difference is that a customer has access to thousands of opinions, which greatly improves decision making. Basically, customers want to find the best product for the lowest price , they search for products that best fulfill their needs inside a price range that they are willing to pay.[5]

It is important to emphasize that the benefit of analyzing other opinions, comes from their neutral nature since they usually not linked to any organization or company.They represent the voice of ordinary consumers, and that differs greatly from ads ("advertisements are biased and tend to favor the product, emphasizing the positives aspects and concealing the negatives ones") [8].

From the e-commerce perspective, receiving consumers' feedback can greatly improve its strategies in order to increase profits of the sector. For example, an online shop can place smart ads by measuring the level of satisfaction of consumers for a given product. Also , if a product has a low level of satisfaction, a smart strategy would be placing a competitor ad inside this page. It is common to find products with thousands of opinions, thus it could be a hard task for a customer to analyze all of them. It could be very tiresome work to find opinions about just some features (aspects) of a product, and usually requires an experienced customer [4].

Temporality and opinion mining are two fields that are given great attention in the fields of Natural Language Processing (NLP) and Social Networks. Tempowordnet is one of the attempts to building a lexicon that may help in finding temporal synsets. It contains words extracted from wordnet, and probabilities of each word which shows if it is a "future","present", "past",or "atemporal. Temporal classifiers are learned from a set of time sensitive synsets and then applied to the whole WordNet to give rise to TempoWordNet. So, each synset is augmented with its intrinsic temporal value.

1.1 Problem Definition

In order to mine opinions according to time tagging (a tag is a keyword or term assigned to a piece of information) we based our work on Tempowordnet (time lexicon based on wordnet) . Some problems are encountered when applying opinion mining this set:

First: Most of the words are atemporal , Table 1.1 shows statistics of the dataset and about 79 % of the words are detected as atemporal, and this is a high percentage.

Table 1.1: Statistics of the Tempowordnet Dataset

	> 0.5
Past	**2508**
Present	**820**
Future	**13758**
Atemporal	**100512**
Total	117598

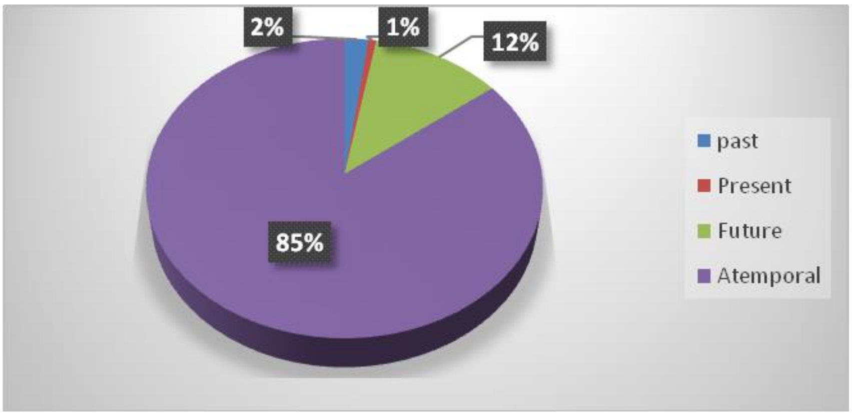

Figure 1.1 Percentage of atemporal words in the Tempowordnet Dataset

Second: Verbs classified as atemporal are 12053 "as shown in Table 1.2 and Figure 1.2

Table 1.2: classification of (pos)words as Noun, Verbs „etc

Pos	*CountOfAtemporal1*
A	6768
N	68859
R	3201
s	9626
v	12058
	100512

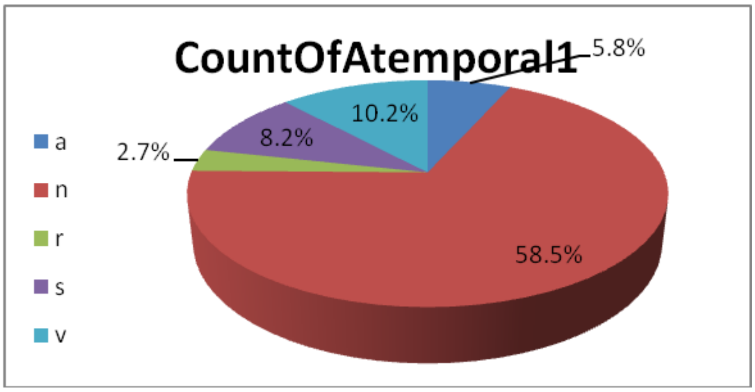

Figure 1.2 : Percentage of Classification for each POS

Third: Abbreviations , slang words and shortcuts are neglected most of the time and are not considered in the mining process, although they may give weight to the opinion orientati

Fourth :Emojis are commonly used these days by social media users. They are used to express users' feelings about a product, event, movies, etc. Emojis are used in clarifying sarcastic versus literal intent[49] . Sarcasm (and irony in general) is especially likely to be misunderstood in written communication , as it involves deciphering a meaning that is often the opposite of what is said [61].

1.2 Objectives

The aim of this work is to propose a framework for enhancing opinion mining prediction and discovering temporal verb references (future, past and present) from opinions by enhancing the temporality of tempowordnet and extracting emojis and shortcuts and also taking into consideration the case of sarcasm . This was done by:

i. A framework for enhancing the opinion mining prediction. The framework relies on employing several modules . To enhance opinion mining the first module is for enhancing the temporality of tempowordnet by increasing the number of temporal words it contains by Classifying verbs as (past , present and future) with the proposed Extract Verbs algorithm (EV).

ii. Creation of a sentiment analysis framework that is capable of analyzing opinions about mobile phones on a feature level using suggested words, abbreviation sentiment identification, emoji detection and sarcasm.

iii. Usage of lexicon based approaches to label training data in order to eliminate manual classification.

iv. Exhibiting the importance of techniques such as suggested words and abbreviation sentiment identification when analyzing opinions for sentiments.

v. Demonstrating the value of emoji detection which shows the need to improve opinion mining based sentiment analysis tools by adding the feature of emoji detection.

vi. Merging the results for modules to make a prediction model contain the aspects for a certain product, sentiment , shortcuts, emojis , sarcasm and the time orientation of the opinion (past, present or future).

Research Plan

Including this chapter, this work is divided into six chapters . The remaining five chapters are as follows:

Chapter 2: Illustrates Opinion Mining as a widespread and important field of Information Retrieval, Information Extraction and Web Mining. This chapter provides a basic background about Tempowordnet which is a temporal ontology which contribute to the success of time-related applications.

Chapter 3: Reviews research related to Opinion Mining and temporal mining fields. The research mainly demonstrates sentiment analysis, feature identification and time trends. This chapter explains the different strategies used for different approaches to sentiment analysis, feature identification and time tags extracting from opinions.

Chapter 4 : Illustrates the first part of the Framework which is responsible for extracting verbs and classifying them as (past, present and future). In this chapter, an algorithm is proposed for classifying verbs as "past, present and future " in order to decrease the number of atemporal words in tempowordnet and then enhancing it , A comparison is made between the results after and before extracting verbs from opinions.

Chapter 5: Illustrates the second part of the framework which is responsible for extracting aspects of a certain product , sentiment words, shortcuts and emojis, and sarcasm identification .

Chapter 6: Summarizes the work presented in this thesis and concludes the results. It also proposes some ideas for future work