Perspective Identification in Informal Text

Hebatallah Elfardy

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

ABSTRACT

Perspective Identification in Informal Text

Hebatallah Elfardy

This dissertation studies the problem of identifying the ideological perspective of people as expressed in their written text. One's perspective is often expressed in his/her stance towards polarizing topics. We are interested in studying how nuanced linguistic cues can be used to identify the perspective of a person in informal genres. Moreover, we are interested in exploring the problem from a multilingual perspective comparing and contrasting linguistics devices used in both English informal genres datasets discussing American ideological issues and Arabic discussion fora posts related to Egyptian politics. Our first and utmost goal is building computational systems that can successfully identify the perspective from which a given informal text is written while studying what linguistic cues work best for each language and drawing insights into the similarities and differences between the notion of perspective in both studied languages. We build computational systems that can successfully identify the stance of a person in English informal text that deal with different topics that are determined by one's perspective, such as legalization of abortion, feminist movement, gay and gun rights; additionally, we are able to identify a more general notion of perspective–namely the 2012 choice of presidential candidate–as well as build systems for automatically identifying different elements of a person's perspective given an Egyptian discussion forum comment. The systems utilize several lexical and semantic features for both languages. Specifically, for English we explore the use of word sense disambiguation, opinion features, latent and frame semantics as well; as Linguistic Inquiry and Word Count features; in Arabic, however, in addition to using sentiment and latent semantics, we study whether linguistic code-switching (LCS)

between the standard and dialectal forms for the language can help as a cue for uncovering the perspective from which a comment was written. This leads us to the challenge of devising computational systems that can handle LCS in Arabic. The Arabic language has a diglossic nature where the standard form of the language (MSA) coexists with the regional dialects (DA) corresponding to the native mother tongue of Arabic speakers in different parts of the Arab world. DA is ubiquitously prevalent in written informal genres and in most cases it is code-switched with MSA. The presence of code-switching degrades the performance of almost any MSA-only trained Natural Language Processing tool when applied to DA or to code-switched MSA-DA content. In order to solve this challenge, we build a state-of-the-art system–AIDA–to computationally handle token and sentence-level code-switching.

On a conceptual level, for handling and processing Egyptian ideological perspectives, we note the lack of a taxonomy for the most common perspectives among Egyptians and the lack of corresponding annotated corpora. In solving this challenge, we develop a taxonomy for the most common community perspectives among Egyptians and use an iterative feedback-loop process to devise guidelines on how to successfully annotate a given online discussion forum post with different elements of a person's perspective. Using the proposed taxonomy and annotation guidelines, we annotate a large set of Egyptian discussion fora posts to identify a comment's perspective as conveyed in the priority expressed by the comment, as well as the stance on major political entities.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my gratitude to all the people who supported me during this journey.

First and foremost, I would like to express my deep gratitude to my advisor Mona Diab for her guidance and support. I consider myself lucky to have worked with an advisor who genuinely cares about her students and who always encourages them to achieve their utmost potential. As I end this journey and start another, Mona will always be my role model in my career.

I would also like to thank the rest of my thesis committee Kathleen McKeown, Smaranda Muresan, Owen Rambow and Philip Resnik for their insightful comments and feedback on this research.

My academic family made the PhD journey a more joyful one. I am grateful to my mentors: Yassine Benajiba, Rebecca Collins, Mohamed R. Fouad and Nizar Habash for their guidance and invaluable advice. Special thanks to:
Apoorv Agarwal, Daniel Alicea, Fahad Alghamdi, Sarah Alkuhlani, Nada Almarwani Sawsan Alqahtani, Mohamed Altantawy, Daniel Bauer, Erica Cooper, Hatim Diab, Ahmed El Kholy, Ali Elkahki, Ramy Eskandar, Noura Farra, Mahmoud Ghonim, Debanjan Ghosh, Weiwei Guo, Sardar Hamidian, Abdelati Hawwari, Kathy Hickey, Idrija Ibrahimagic, Faiza Khattak, Rivka Levitan, Derrick Lim, Manoj Pooleery, Vinodkumar Prabhakaran, Axinia Radeva, Mohammed Rasooli, Jessica Rosa, Sara Rosenthal, Wael Salloum, Ali

Seyfi, Shabnam Tafreshi, Boyi Xie and Aya Zerikly. I can never forget the amazing memories I shared with all of you.

I am especially grateful for my family. I am blessed with the most supportive mother and siblings one can dream of. Thanks Mom, Hadir and Wael for your love and encouragement. This journey would not have been possible without you.

To my beloved mother and in loving memory of my father.

Chapter 1

Introduction

In this thesis, we investigate the notion of "Ideological Perspective" in as much as it is reflected in the authors' texts. We build computational systems that identify such "Ideological Perspective" by examining the authors' stance on political events, leaders and ideological issues.

1.1 Overview

Nowadays, informal genres have become rich sources of information providing a plethora of documented conversations discussing a variety of topics. For example, social media and online discussion fora have become major platforms for ideological and political discussions. Through these discussions people often express their stances and concerns on different topics and entities governed by their belief systems, such as partisanship and positions on the legalization of abortion, creationism, climate change, gay and gun rights among other polarizing issues. Automatically identifying people's stances on these issues as well as the underlying perspective governing these stances is a challenging research problem that has recently started to garner interest in the Natural Language Processing (NLP) community. The successful identification of the ideological background of people has a lot of applications, such as recommendation systems and targeted advertising and even the prediction of possible future events. In general, due to the many linguistic registers used in informal genres, the problem is a significant challenge for NLP. In particular though, especially in discussion fora, the problem becomes even more complex given dynamic political

settings in which no clear taxonomy of the different perspectives presents itself, and where the perspectives are emergent and shifting.

One example of such a case is in Egypt where, after the Arab Spring, citizens often changed their positions on different political entities as they broke up and reformed their alliances. In this thesis, we investigate the specific linguistic devices that users exploite to express their ideological perspectives. Specifically, we are interested in two aspects: (1) identifying individual users' perspectives on topics governed by their belief system and (2) studying the problem from a multilingual standpoint by comparing English and Arabic. We utilize lexical and semantic linguistic cues to build supervised systems that can identify a person's perspective in English informal genres data discussing American politics and Arabic discussion fora data discussing Egyptian politics. We devise a principled taxonomy of major Egyptian community perspectives, develop annotation guidelines and gather linguistic annotations based on the developed taxonomy and guidelines. Additionally, since Arabic is a diglossic language, we address the problem of code-switch detection between Standard Arabic and Dialectal Arabic as a preprocessing step prior to building systems for identifying perspective in Arabic, while exploring its impact on identifying a person's perspective.

1.2 Contributions

In this section, we summarize our research contributions. (Table 1.1 lists the associated publications.) As mentioned earlier, we are interested in automatically identifying different dimensions of people's perspectives as reflected by their stances towards different ideological topics and political entities in English informal genres data discussing American ideological issues as well as Arabic discussion fora posts on the topic of Egyptian politics. In doing so, we solve several challenges.

For English and Arabic Perspective Identification: We build computational systems that can successfully identify the perspective from which a given informal text is written. We do the following:

- Use lexical and semantic features to build supervised computational systems that can successfully identify the stance of a person in English informal text on different topics that are determined by one's perspective such as legalization of abortion, feminist movement, gay and gun rights in addition to being able to identify a more general notion of perspective–namely, the 2012 choice of presidential candidate. We evaluate the systems intrinsically. We also participate in SemEval 2016 task 6–Detecting Stance in Tweets. Our system ranks in 6^{th} place (out of 19 participating systems) and we later improve its performance (Elfardy and Diab, 2016b).

- Build supervised systems for automatically identifying different elements of a person's perspective given an Egyptian discussion fora comment. The systems utilize several lexical and semantic features and look at whether or not identifying code-switching helps in determining a person's perspective.

- Draw insights from our work on both English and Arabic. We discuss the similarities and differences between both languages by analyzing the difference in the notion of perspective elements in both studied languages as well as analyzing which linguistic devices help in identifying the perspective in each language.

Devising a taxonomy of Egyptian political perspectives: The lack of a commonly accepted taxonomy for the most prevalent perspectives among Egyptians creates a significant challenge. In order to solve this impediment, we do the following:

- **Develop a taxonomy for Egyptian ideological perspectives:** Based on the works of "The Hariri Center at the Atlantic Council" and "Carnegie Endowment for In-

ternational Peace" (Brown, 2013; *Carnegie Endowment for International Peace*; Carothers and Brown, 2012; *The Hariri Center at the Atlantic Council*), we develop a taxonomy to quantifiably characterize the most common community perspectives among Egyptians.

- **Develop guidelines for identifying Egyptian ideological perspectives:** Based on the taxonomy we created, we use an iterative feedback-loop process to devise guidelines on how to successfully annotate a given online discussion forum post with different elements of a person's perspective (Elfardy and Diab, 2016a). To the best of our knowledge, this is the first attempt at creating guidelines for collecting fine-grained multidimensional annotations of Egyptian Ideological Perspectives that tries to uncover the different underlying elements of a person's belief system.

- **Collect large-scale annotations:** Using the proposed taxonomy and annotation guidelines, we annotate a large set of Egyptian discussion fora posts to identify a comment's perspective as expressed in the salience it reflects in the comment as a proxy for the person's stance on political entities and issues.

Arabic Preprocessing: The diglossic nature of the Arabic language where the standard form of the language (MSA) coexists with the regional dialects (DA) corresponding to the mother tongue of Arabic speakers in different parts of the Arab world poses a significant challenge to processing Arabic social media text in particular. Code-switching degrades the performance of almost any MSA-only trained NLP tool when applied to DA or to code-switched MSA-DA content. In order to solve this challenge, we do the following:

- **Create Annotation Guidelines:** We build a simplified set of guidelines aimed at identifying token-level code-switching in Arabic. For a given sentence, the

guidelines address the problem of how to identify the class of each word in that sentence;

- **Collect Token-Level Annotations:** Using the proposed guidelines to annotate a corpus that is rich in DA with frequent code-switching to MSA we annotate a large set of data on the token-level;

- **Automatically Handle Token-Level Dialectal Arabic Identification:** We build the first system–AIDA–for performing automatic token-level identification of Dialectal Arabic in a given Arabic text. AIDA was evaluated intrinsically on the dataset released for the shared task at EMNLP 2014 code-switching workshop (Solorio et al., 2014) and outperformed all participating systems on the token-level dialect classification task (Elfardy, Al-Badrashiny, and Diab, 2014b);

- **Automatically Handle Sentence-Level Dialectal Arabic Identification:** We extend our token-level Dialectal Arabic identification system (AIDA) to identify whether a given sentence is predominately MSA or DA. The sentence level component was evaluated intrinsically on a standard dataset (Al-Badrashiny, Elfardy, and Diab, 2015; Elfardy and Diab, 2013) and it outperformed all baselines. Both token and sentence level components were evaluated extrinsically in the context of machine translation and resulted in BLEU score improvements (Aminian, Ghoneim, and Diab, 2014; Salloum et al., 2014). Additionally, AIDA's token and sentence level tags were used as features for identifying uncertainty cues in Arabic tweets (Al-Sabbagh, Diesner, and Girju, 2013).

Year	Ref.	Venue	Mode	Citation
2016	LAW-16	LAW (Workshop)	Long	Elfardy and Diab, 2016a
	SemEval-16	SemEval (Workshop)	Long	Elfardy and Diab, 2016b
2015	*SEM-15	*SEM (Main Conf.)	Long	Elfardy, Diab, and Callison-Burch, 2015
	CONLL-15	CONLL (Main Conf.)	Long	Al-Badrashiny, Elfardy, and Diab, 2015
2014	EMNLP-14	EMNLP (Workshop)	Long	Elfardy, Al-Badrashiny, and Diab, 2014b
	XRDS-14	XRDS (ACM Student Magazine)	-	Elfardy, Al-Badrashiny, and Diab, 2014a
	ACL-14	ACL (Main Conf.)	Short	Salloum et al., 2014
	LREC-14	LREC (Main Conf.)	Long	Diab et al., 2014
2013	NLDB-13	NLDB (Main Conf.)	Demo	Elfardy, Al-Badrashiny, and Diab, 2013
	ACL-13	ACL (Main Conf.)	Short	Elfardy and Diab, 2013
2012	COLING-12	COLING-Main Conf.	Short	Elfardy and Diab, 2012c
	EAMT-12	EAMT (Main Conf.)	Demo	Elfardy and Diab, 2012a
	LREC-12	LREC (Main Conf.)	Long	Elfardy and Diab, 2012b

Table 1.1: Contributions of the thesis

1.3 Thesis Outline

This thesis is organized as follows:

- In Chapter 2, we explain the motivation behind the work conducted in this thesis, discuss the relevant social-science work and relate the task to media framing.
- Chapter 3 focuses on Dialectal Arabic code-switch detection. We describe our system–AIDA–for token and sentence level Dialectal Arabic Identification and compare it to related work.
- In Chapter 4, we describe our work on perspective identification in Arabic discussion fora. We present our proposed taxonomy and guidelines for collecting linguistic

annotations as well as our approach and experimental results for building computational systems that can uncover the different elements of perspective.

- In Chapter 5, we present our work on automatic perspective identification in English. We describe the datasets used; as well as the followed approach; along with the experimental results.
- In Chapter 6, we discuss computational related work and compare it to the work presented in this thesis.
- Finally in Chapter 7, we summarize the main findings of this thesis, list our research contributions as well as limitations and outline future directions for automatic perspective identification.

Chapter 2

Background

In this chapter, we discuss the motivation as well as the related social science literature.

2.1 Motivation

With the pervasion of social media and online discussion fora, political and ideological discussions have increased significantly. These discussions typically reflect polarizing topics and, in doing so, convey the participants' belief systems by expressing their stance on contentious issues; namely, their "Ideological Perspective". Van Dijk, 1998 defines "Ideology" as *"the set of factual and evaluative beliefs–or the knowledge and opinions–of a group.* These beliefs influence an individual's goals, expectations, and views of the world (ibid.). Identifying the perspective of online discussion participants poses a research problem with a wide variety of applications, including recommendation systems, planning political campaigns, targeted advertising, political polling, product reviews, and the potential for predicting future events. As a matter of fact, social media played a major role in the Arab Spring. In Egypt, for example, activists and political leaders resorted to social media as an alternative to the censored and mostly biased state- and private-owned media. Social media formed the primary means of communication between the public and the activists, most of whom used social media to make announcements, campaign for elections, spread awareness of important causes and conduct polls in order to predict election outcomes (Borge-Holthoefer et al., 2015; Mansour, 2012; Siegel, 2014). While social media polls failed in many cases to provide ideal predictions of voting patterns–such as predict-

ing the winning candidate in the 2012 presidential elections, or predicting that the majority of people will vote against some suggested amendments to the constitution when actually the majority (>70%) later voted in favor of these amendments–it did provide significant clues for the common perspectives and perspective shifts in the public sphere. After the Revolution in January, 2011, alliances kept forming (and later breaking) between Islamist movements, revolutionaries, and public figures from Mubarak's regime and those from the army. The formation and break-up of such alliances often triggered a perspective shift among various segments of Egyptian society.

2.1.1 Centrality

These perspective shifts can be best explained by Converse's (Converse, 1962) concept of centrality in belief systems. Converse defines a belief system as: *"a configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence. The constraint may be taken to mean the success we would have in predicting, given initial knowledge that an individual holds a specific attitude, that he holds certain further ideas and attitudes."* For example, if we know that an American citizen supports ObamaCare, can we predict that he/she supports immigration reform? While there are Americans who support ObamaCare and oppose immigration reform, the vast majority of people either support or oppose both issues because the stance towards these two issues is always backed up by one's ideology or belief system.[1] In American politics, one's political party affiliation–whether Democrat, Republican or Neither (i.e. Independent)–is often a good indicator of one's belief system. The "Conservative Liberal" scale is another yardstick with which we, according to Converse, can position events, legislation, and political leaders. Ideal Point Models are often used to place legislators on such scale by estimating the position of politicians on different issues–their ideal points–from their voting behavior (Poole and Rosenthal, 1985) or more recently by jointly modeling legislators' votes on

[1] In this work, we use *Ideology*, *perspective* and *Belief System* interchangeably

proposed bills, the text content of bills, and the language used by legislators (Nguyen et al., 2015). Locating the mass public on the same scale often fails, in part because the belief systems of most people lack the solidity of their leaders' beliefs. Converse states that within a belief system idea elements vary in "centrality". This variation always governs what happens when the status of one of the idea elements in a belief system changes. For example, what will a self-proclaimed Republican do if the Republican Party decided to change its stance on ObamaCare and started to support it? The reaction of the person will depend on which is more central to his/her belief system: the political party affiliation or the stance on healthcare? This concept of centrality is related to how Chong and Druckman, 2007 define a person's attitude towards an entity. The authors suggest that a person's attitude towards an object or a topic is a function of his/her evaluations of the different attributes of that topic, and the weight assigned by the person to each one of these attributes.

$$Attitude = \sum v_i * w_i \tag{2.1}$$

and

$$\sum w_i = 1 \tag{2.2}$$

where v_i is the evaluation/opinion of attribute i and w_i is the salience weight.

Accordingly, we expect that the more central elements of a person's belief system will have higher salience weights.

After the 2011 Revolution, as the stance of political leaders towards the Army, the Police, and the Islamists–among other political entities–kept changing, many Egyptians faced choices that conveyed the most central elements to their belief systems. This stance-change among the leaders often triggered shifts in perspective among the mass public towards those entities less central to their belief systems.

2.2 Framing

From a social science viewpoint, the notion of "Perspective" is related to the concept of "Framing" (Entman, 1993). Entman gives the following definition for framing: *"Selecting some aspects of the perceived reality and making them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described."* According to Entman, frames can perform the following functions: (1) define problems; (2) diagnose causes; (3) make moral judgments; and (4), suggest remedies. By increasing the salience of certain ideas and by suppressing/concealing other ideas, framing can alter people's perceptions, which is how the media often directs people to think about certain topics. Framing is also related to bias, which in itself can have several different meanings. Entman, 2010 distinguishes between three facets of bias in the context of news' coverage:
(1) *distortion bias* focuses on news that falsifies the truth; (2) *content bias* applies to cases where the news' sources do not provide a balanced treatment to both sides in a political conflict; (3) *decision-making bias* refers to the reasons why journalists produce biased content. Elmasry, 2009 and Elmasry et al., 2013 provide examples of the second type of bias–content bias–by studying the Israeli-Palestinian conflict. In Elmasry, 2009, the author studies how two American newspapers, The New York Times and Chicago Tribune, frame the events in favor of the Israeli side; however, in Elmasry et al., 2013, the authors study how two Arabic TV networks, Al-Jazeera and Al-Arabiya, use opposite framing mechanisms of salience and victim personalization to highlight the Palestinian perspective when covering the same conflict. Our hypothesis is that, similar to media framing, people use selection and salience techniques when discussing ideological topics in order to convince the reader with their perspective. In many cases, people will neglect the other side's stance when discussing the topic, or will discuss both sides, but will highlight and promote their own views. These selection and salience mechanisms, often reflected on different levels of

linguistic representation, can be strong indicators of the writer's stance–whether supporting or opposing or indifferent–on different topics.

On the most basic level, these decisions are expressed in the lexical choice (Gentzkow and Shapiro, 2006; Monroe, Colaresi, and Quinn, 2008). For example, a person who opposes gun rights is more likely to use words that emphasize "death", while a supporter of those rights is more likely to use ones that promote "self-defense". As the saying goes, "One man's terrorist is another man's freedom fighter". Perspective is also expressed on the syntactic and semantic levels. Greene and Resnik, 2009 show that the syntactic structure can be a strong indicator of bias. For example, using the passive voice puts less emphasis on the doer than the use of a transitive verb. This is particularly important when the verb is sentiment bearing. In such a case, the passive voice is less likely to associate the sentiment with the doer. For example, saying *"Egyptian Revolutionists were killed during the protests"* does not identify the causal agent as opposed to saying *"The police killed Egyptial Revolutionists during the protests."*

Sentiment also serves as another important clue for identifying a person's perspective since it shows this person's opinion on different topics. In fact, from a computational point of view, the work on perspective identification closely relates to subjectivity and sentiment analysis. One's perspective normally influences the sentiment towards different topics or targets. Conversely, identifying the sentiment of a person towards multiple targets can serve as a clue for identifying his or her perspective. For example, we expect a prototypical Democrat to express positive sentiment towards Obama, universal healthcare, immigration reform, gun control, Planned Parenthood and gay rights; on the other hand, a prototypical Republican is expected to express negative sentiment towards all of these topics while expressing positive ones towards gun rights and fiscal conservatism. Similarly, while we expect a typical Egyptian Revolutionist to express positive sentiment towards social justice, freedom of speech, and the Revolution's public figures, that same Revolutionist is expected to show negative sentiment towards Mubarak and authoritarian regimes. As mentioned

earlier, these different issues vary in how important–or central–they are to the belief system of each person and the stance on the less central elements will vary across time.

The notion of stance is akin to the notion of "*tone of text*" in political science literature (Barberá et al., 2016; Boydstun et al., 2013; Boydstun and Gross, 2014). A given text can have a (1) Positive/Pro (whether implicit of explicit), (2) Negative/Anti (whether implicit of explicit) or (3) Neutral tone. For example, when discussing immigration, a positive tone will portray immigrants' right in a more sympathetic manner, a negative tone will portray them in a non-sympathetic way, while in a neutral tone both positive and negative tones will balance each other out (Boydstun et al., 2013).

2.3 Social media challenges

We hypothesize that we can automatically identify people's stances on different issues, as well as people's underlying perspectives, by building computational systems that integrate the linguistic devices that those people use as a way of highlighting and promoting their views.

While automatically identifying people's perspective and stance on contentious issues is a challenging task, in part, due to those people's use of nuanced sentiment when expressing the stance on these topics, it is more challenging in such a dynamic setting. Moreover, in general, when the studied genre is social media or online discussion fora, the challenges are exacerbated. In such media sources, there are no constraints or clear sets of rules on users' language. Furthermore, in threaded conversations, discussions typically deviate from the main topic and participants tend to form side-conversations with multiple dynamic shifts within the same thread among participants. Arabic poses more challenges than other languages due to its diglossic nature. Arabic exists in two forms: Modern Standard Arabic (MSA) and Dialectal Arabic (DA). MSA is the language used in education, scripted speech, and official settings, while DA corresponds to the

native tongue of the speakers of Arabic. Even though these DA variants have no standard orthography, they are used in unofficial written communication and are increasingly seen in social media text. Code-switching between MSA and DA happens both intrasententially and intersententially. Most of the available Arabic NLP tools are not targeted towards MSA-DA code-switched text and when applied to such input, their performance drops dramatically. This creates a need for an extra preprocessing step for the detection and handling of MSA-DA code-switching when dealing with Arabic input. Additionally, models that look into how and when this code-switching occurs in Arabic might provide insights that can help identify a person's perspective. For example, Arab Nationalists naturally view DA as a threat to the unity of Arabs; hence, they refrain from using it as much as possible. In this thesis, we are interested in studying whether code-switching in Arabic does indeed help in identifying the underlying perspective from which a given text was written.

2.4 Summary

In this chapter, we have provided formal definitions for the notions of ideology and belief system, discussed the motivation behind our work on ideological perspective identification, and related the task to media framing.

Chapter 3

Enabling Technology for Arabic: AIDA

The main focus of this thesis is to build computational systems that can automatically identify ideological perspective in Arabic and English informal text. Arabic informal text presents a significant challenge that goes beyond the problems posed by English informal text. Arabic is a diglossic language (Ferguson, 1959) where the standard form of the language (MSA) and the regional dialects (DA) live side by side and are closely related. We look into the phenomenon of linguistic code-switching between these two forms to study whether it can be used as an indicator in identifying the perspective of a given comment.

In this chapter we describe our approach for building state-of-the-art systems for detecting code-switching between MSA and DA in the course of an utterance and across different utterances. We begin by giving a brief overview of the problem in Section 3.1. We then describe the datasets used to build and evaluate our systems in Section 3.2. The token-level and sentence-level components and the latest version of the system (AIDA-2) are described in Sections 3.3, 3.4 and 3.5 respectively. Finally we discuss related work in Section 3.6 and summarize the chapter in Section 3.7.

3.1 Background

With the rise of social media, the use of informal language mixed with formal language became increasingly common. The degree of mixing formal and informal language-registers varies across languages. The problem is quite pronounced in Arabic where the formal modern standard Arabic (MSA) and the informal dialects of Arabic (DA)

differ on all levels of linguistic representation–morphologically, lexically, syntactically, semantically and pragmatically. MSA is used in formal settings, edited media, and education, while the informal dialects correspond to the native tongue of Arabic speakers and are also being used in written communications in social media and other informal genres. There are multiple dialects corresponding to different parts of the Arab world. Habash, 2010 classifies Arabic dialects into the following broad categories: (1) Egyptian Arabic, (2) Levantine Arabic, (3) Gulf Arabic, (4) North African Arabic, (5) Iraqi Arabic, (6) Yemenite Arabic and (7) Maltese (which in some cases is not considered Arabic). For each one of these dialects, sub-dialectal variants exist. Arabic speakers/writers normally "code-switch" between the two forms of the language both inter- and intra-sententially. *Linguistic Code-Switching* refers to switching from one language to the other in bilingual communities in the course of an utterance (Joshi, 1982). Joshi suggests that in such setting there is a "matrix language"–where the mixed sentence is coming from–and an "embedded language". When the matrix and embedded languages are variants of the same language, and both languages share the same character-set and are closely related, the problem becomes "Dialect Identification". Automatically identifying code-switching between variants of the same language (Dialect Identification) is quite challenging due to the lexical overlap and significant semantic and pragmatic variation, yet it is crucial as a preprocessing step before building any Arabic NLP tool.

Our system for "Automatic Identification of Dialectal Arabic"–AIDA–addresses the problem of token- and sentence-level dialect identification in Arabic by detecting code-switching between Egyptian DA (EDA) and MSA. The token-level component of AIDA can be thought of as a word sense disambiguation tool that can identify for each word in a given sentence whether the most plausible sense is the MSA or the EDA one. This is particularly helpful because MSA and EDA have a lot of "faux amis" which are words that look exactly the same in both languages but have different meanings. Moreover, MSA

and EDA have different tools for tokenization, part of speech tagging, lemmatization, etc. The sentence-level component, which decides whether a given sentence is mostly MSA or EDA, helps us in deciding which set of tools to use. Moreover, the level of dialectalness–or lack thereof–of a given text can give us better insights about the authors, which can prove beneficial in identifying their backgrounds and perspectives.

3.2 Datasets

In this section we describe the datasets that we use for training and evaluating both token and sentence-level components of our system.

3.2.1 Token-Level Datasets

Since AIDA was the first system to automatically handle token-level dialect identification in Arabic, we established a set of guidelines and collected token-level annotations for EDA-MSA data curated from Egyptian discussion fora, commentaries and Wikipedia articles (Elfardy and Diab, 2012b). Developing the guidelines was an iterative process that aimed at reaching a higher inter-annotator agreement. In the final version of the guidelines, we asked the annotators to perform a totally contextual annotation for the words in each given post. If a word can be used in both MSA and EDA, the class is chosen based on the class of the context it appears in. In rare cases where the context itself is ambiguous, the annotators were instructed to use a "Both" class indicating that the word can belong to either class. During the annotation, the annotators assigned each word to one of the following classes:

- **MSA**: If the token is contextually MSA or if it is only used in MSA (ex. الواقع "*AlwAqE*" [1] meaning "The reality")
- **EDA**: If the token is contextually EDA or if it is only used in EDA (ex. مش "*m$*" meaning "Not", معلش "*mEl$*" meaning "Never Mind")
- **Both**: If the given context is not sufficient to identify the token as MSA or EDA (ex. السلام عليكم "*AlslAm Elykm*" meaning "Hello/Peace be upon you")
- **NE**: If the token is a named-entity such as:
  - Terms of Address (ex. أستاذ " *>stA∗*" meaning "Sir", عم "*Em*" meaning "Uncle");
  - People's Names (ex. Arabic/Foreign person's name and foreign entity names);
  - Organizational Names (ex. الأمم المتحدة "*Al>mm AlmtHdp*" meaning "The United Nations" البيت الأبيض "*Albyt Al>byD*" meaning "The White House");
  - Company names;
  - Country names.

- **Foreign**: If the token is not (originally) part of the Arabic language whether it is spelled in Arabic or in Latin script. (ex. جيلاتو "*jylAtw*" meaning "Gelato", كانلوني "*kAnlwny*" meaning "Cannelloni")
- **Typo**: If the word is misspelled such as:

  - Splits: If a word is split into several consecutive words (i.e., the word has extra spaces);
  - Merges: If a word seems to be multiple words stuck together.

As mentioned earlier, DA has no standard orthography, hence we do not consider

[1] We use Buckwalter transliteration scheme: `http://www.qamus.org/transliteration.htm`

	MSA	EDA	Both	NE	Foreign	Typo	Unknown	**Total**
Dev.	19, 955	9, 769	9	2, 581	257	330	149	33, 050
Test.	15, 462	16, 242	5	2, 434	408	312	44	34, 907

Table 3.1: AIDA Token-Level dataset: Tag Distribution and total number of tokens in the development and test sets of the first evaluation dataset

inconsistent spelling (with respect to the MSA homograph cognate) or the use of speech effects (consecutive repeated characters) as typos.

- **Unknown**: If the annotator does not know the word at all.

Punctuations, URLs and numbers were not assigned a class. Table 3.1 shows the statistics of this dataset.

Later, another dataset was released for the *First Workshop on Computational Approaches to Code Switching* (Solorio et al., 2014). This newer dataset was annotated with a refined set of our guidelines in order to simplify the task. These guidelines instruct annotators to assign each word to one of the following six tags:

- **lang1/MSA**: If the token is MSA *(Same as the first dataset)*;
- **lang2/EDA**: If the token is EDA *(Same as the first dataset)*;
- **NE**: If the token is a named entity *(Same as the first dataset)*;
- **Ambig**: If the given context is not sufficient to identify the token as MSA or EDA *(Similar to "Both" class in the first dataset)*;
- **Mixed**: If the token is of mixed morphology (ex. المألوشون "*Alm>lw$wn*" meaning "the ones that were excluded or rejected" where the word is EDA but the suffix is MSA.);
- **Other**: If the token is or (is attached to) numbers, punctuation, Latin characters, emoticons, etc.

	ambig	MSA	EDA	Mixed	NE	Other	**Total**
Training	1, 066	79, 134	16, 291	15	14, 112	8, 699	119, 317
Test 1	11	44, 594	141	1	5, 994	3, 991	54, 732
Test 2	119	10, 459	14, 800	2	4, 321	2, 940	32, 641
Surprise Genre Test	110	2, 687	6, 930	3	1, 097	1, 190	12, 017

Table 3.2: AIDA Token-Level Dataset: Tag Distribution and total number of tokens in the training and test sets released for the shared task at EMNLP's 2014 Code Switching workshop

	MSA Sentences	EDA Sentences	MSA Tokens	EDA Tokens
Train	12, 160	11, 274	300, 181	292, 109
Test	1, 352	1, 253	32, 048	32, 648

Table 3.3: AIDA Sentence-Level Annotation: Statistics of the training and test sets.

We only present the results on the second dataset in order to compare the performance of AIDA to other systems that participated in the shared task. Table 3.2 shows the statistics of this dataset.

3.2.2 Sentence-Level Dataset

For the sentence-level component, we use the code-switched EDA-MSA portion of the crowd source annotated dataset by Zaidan and Callison-Burch, 2011. The dataset consists of user commentaries on Egyptian news articles. We split the dataset into 90% training set and 10% test set. The data is almost balanced; 51.9% of the sentences are MSA and the rest are EDA. Table 3.3 shows the statistics of this dataset.

3.3 Token-Level Dialect Identification

We use a hybrid approach that relies on Language Models (LM), MADAMIRA–a tool for morphological analysis and disambiguation for Arabic (Pasha et al., 2014)–and gazetteers to tag each word in a given Arabic sentence. The decisions from these three components

are then used by a combiner module that makes the final decision with respect to the class of each given word (Elfardy, Al-Badrashiny, and Diab, 2013, 2014b; Elfardy and Diab, 2012c).

3.3.1 Preprocessing

We experiment with two preprocessing techniques:

1. **Basic/Surface**: In this scheme, no significant preprocessing is applied to the text apart from the regular initial clean-up, which includes normalizing all punctuation; URLs; numbers and non-Arabic words to *PUNC*, *URL*, *NUM*, and *LAT* keywords respectively. Additionally, all word-lengthening effects are normalized by reducing all redundant letters in a given word to a standardized form in order to reduce the sparsity of the data; for example, the elongated form of the word كتير "*ktyr*" meaning "a lot", which could be rendered in the text as كتتتتييير "*kttttyyyyr*", is reduced to كتتييير "*kttyyyr*". We choose to restrict word-lengthening effects to three letters–as opposed to two–in order to maintain a signal that there is a speech effect which could be a DA indicator.
2. **Tokenized**: In this scheme, in addition to basic preprocessing, we use MADAMIRA toolkit to tokenize words. Arabic has a set of clitics that get attached to words when written and that can be segmented in different levels of detail. In order to reduce the sparseness of the data, we use the most detailed level of tokenization provided by the tool (D3) (Habash and Sadat, 2006) which splits off:

    - conjunction clitics و "*w*" and ف "*f*";
    - particles ك "*k*", ب "*b*" and س "*s*";

- pronominal clitics;
- the definite article ال "*Al*".

Using this scheme, وبالفريق "*wbAlfryq*" meaning "and by the team" becomes و ب ال فريق "*w b Al fryq*", and وبفريقهم "*wbfryqhm*" meaning "and by their team" becomes و ب فريق هم "*w b fryq hm*" after tokenization.

3.3.2 Language Model

The '*Language Model*' (LM) module uses the preprocessed training data to build a 5-grams LM. All tokens in a given sentence in the training data are tagged with either "MSA" or "EDA" depending on whether the sentence was collected from MSA or EDA source. Using SRILM (Stolcke, 2002) and the tagged datasets, a 5-gram LM is built with a modified Kneser-Ney discounting. We build two variants of the language model corresponding to the two preprocessing schemes mentioned earlier–Surface and Tokenized. For the tokenized LM, we use D3 tokenization. All D3 tokens of a word are assigned the same tag of their corresponding word (ex. if the word بالفريق *bAlfryq* meaning "by the team" is tagged as MSA, then each of ب "*b*", ال "*Al*" and فريق "*fryq*" gets tagged as MSA.

The prior probabilities for all MSA and EDA words are calculated based on their frequency in the MSA and EDA corpora, respectively. For example, the EDA word كتير "*ktyr*", meaning *"a lot"*, will have a probability of 0 for being tagged as MSA since it would not occur in the MSA corpora, and a probability of 1 for being tagged as EDA. Other words can have different probabilities depending on their unigram frequencies in

both corpora.

The LM and the prior probabilities are then used as inputs to SRILM's *disambig* utility which uses them on a given untagged sentence to perform a lattice search in order to return the best sequence of tags for the given sentence. Thus for any new untagged sentence, the '*Language Model*' module uses the already built LM and the prior probabilities via Viterbi search (Forney, 1973) to find the best sequence of tags for the given sentence. If there is an out-of-vocabulary word in the input sentence, the '*Language Model*' leaves it untagged. In the tokenized case, if the tags assigned by the LM to the prefixes, suffixes and stem of a word are not the same, we set "*isMixed*" flag to *true*.

We build the Language Model using the following data:

1. **Shared-task's training data (*STT*)**: The training dataset described in Table 3.2 (119, 317 words).
2. **LDC Web-log training data (*WLT*)**: Eight million words, half of which comes from *lang1/MSA* corpora while the other half is from *lang2/EDA* corpora. We define a corpus as being MSA if it was collected from an MSA source (such as an MSA forum) and as being EDA if it was collected from an EDA source. While this method yields noisy and not true labels, we do expect a higher percentage of MSA content in these MSA sources and a higher percentage of EDA content on the EDA sources. We did a manual inspection of the MSA data and found that while most of it is indeed MSA, it contains some highly dialectal sentences. In order to filter the dialectal content, we built a lexicon of highly dialectal Egyptian words and excluded all the sentences where any of these words occur. We then labeled all tokens in the sentence/comment according to the dialect of the source (MSA or EDA) it was collected from.[2]

[2]The LDC numbers of these corpora are 2006{E39, E44, E94, G05, G09, G10}, 2008{E42, E61, E62, G05}, 2009{E08, E108, E114, E72, G01}, 2010{T17, T21, T23}, 2011{T03}, 2012{E107, E19,

Since the size of *STT* is very small compared to *WLT* (1.5% of *WLT* size), the existence of six different tags in this corpus is expected to add noise to the already weakly labeled *WLT* data. To make *STT* consistent with *WLT*, we change the labels of *STT* as follows:

- If the number of MSA tokens in the tweet exceeds the number of EDA tokens, we assign all tokens in the tweet "MSA" tag;
- Otherwise, all tokens in the tweet are assigned "EDA " tag.

3.3.3 MADAMIRA

MADAMIRA is a publicly available tool for morphological analysis and disambiguation of MSA and EDA text (Pasha et al., 2014). Using MADAMIRA, each word in a given untagged sentence is tokenized, lemmatized and POS-tagged. Moreover, the MSA and English glosses for each morpheme of the given word are provided. MADAMIRA uses two underlying morphological analyzers. The first morphological analyzer SAMA (Maamouri et al., 2010) analyzes the words that MADAMIRA deems MSA, while the second one CALIMA (Habash, Eskander, and Hawwari, 2012) analyzes those that are deemed EDA.
We use MADAMIRA to tag each word in the input sentence as being MSA or EDA according to the morphological analyzer upon which it is disambiguated. Out-of-vocabulary words are tagged as "Unknown".
Additionally, for the tokenized preprocessing variant of the system, we use MADAMIRA to tokenize both of the language models and the input sentences using D3 tokenization-scheme in order to maximize the coverage of the Language Models (LM).

3.3.4 Gazetteers

We use the ANERGazet (Benajiba, Rosso, and Benedruiz, 2007) to identify named-entities. ANERGazet consists of the following gazetteers:

E30, E51, E54, E75, E89, E94, E98, E99}.

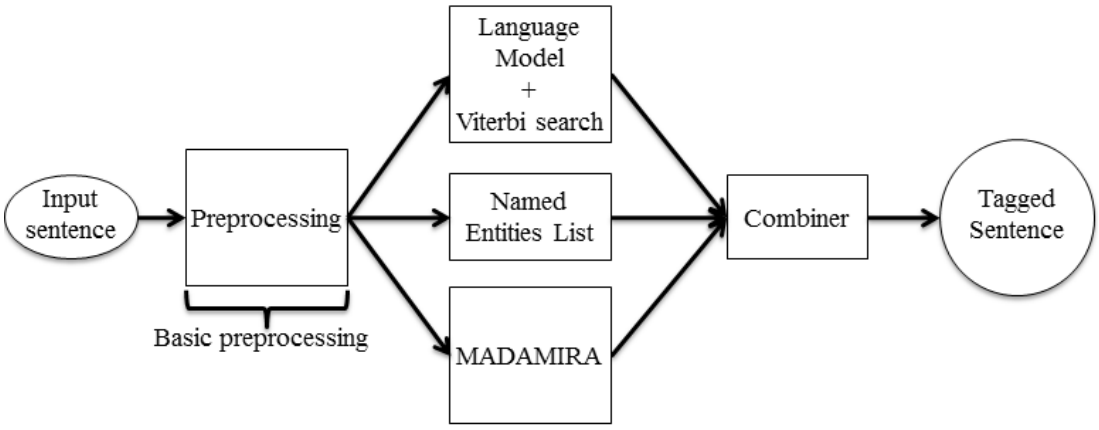

Figure 3.1: AIDA: Pipeline using the basic preprocessing scheme

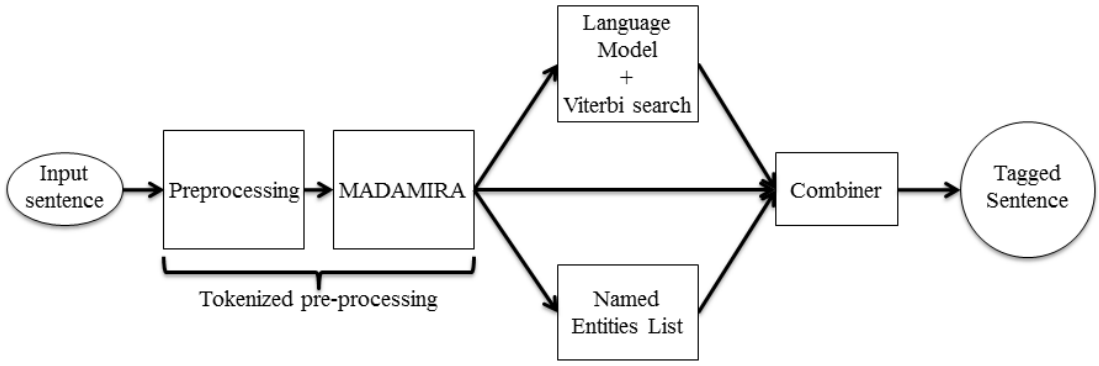

Figure 3.2: AIDA: Pipeline using the tokenized preprocessing scheme

- **People:** 2, 100 entries corresponding to names of people;
- **Organizations:** 318 entries corresponding to names of organizations such as companies and football teams;
- **Locations:** 1,545 entries corresponding to names of continents, countries, cities, etc.

Moreover, all NE tokens in *STT* are used to enrich our named-entity list.

3.3.5 Combiner

Each word in the input sentence can get different tags from each module. The *"Combiner"* module uses all of these decisions, and the following set of rules, to assign the final tag to each word in the input sentence. The final set of rules are:

1. If the word contains any numbers or punctuation, it is assigned *"Other"* tag;
2. Else if the word is present in any of the gazetteers, or if MADAMIRA assigns it *noun_prop* POS tag, the word is tagged as *"NE"*;
3. Else if the word is (or all of its morphemes in the tokenized scheme are) identified by the LM as either *"MSA"* or *"EDA"*, the word is assigned the corresponding tag;
4. Else if the word's morphemes are assigned different tags, the word is assigned the *"Mixed"* tag;
5. Else if the LM does not tag the word (i.e. the word is considered an out-of-vocabulary word by the LM) and:
    - If MADAMIRA retrieves the analysis from SAMA, the word is assigned *"MSA"* tag;
    - Else if MADAMIRA retrieves the analysis from CALIMA, then the word is assigned *"EDA"* tag;
    - Else if the word is still untagged (i.e. non-analyzable), the word is assigned *"EDA"* tag.

The rules were designed in a way that (a) follows the annotation guidelines of the evaluation dataset (ex. tagging any word that is attached to a number or punctuation as "Other"), (b) increases the coverage of the named entities and (c) gives higher precedence to the (contextual) language models over the (non contextual) morphological analyzers for all classes other than "NE". Figures 3.1 and 3.2 show the pipeline of the token-level component using (a) Surface and (b) Tokenized preprocessing schemes respectively.

3.3.6 Token-Level Experiments

Tuning We split *STT* training data into 90% training set and 10% development set in order to find the best configuration for the system. We experiment with both Surface and Tokenized preprocessing schemes.

%	Ambig.	MSA	EDA	Mixed	NE	Other	Weighted Avg $F_{\beta=1}$
Tokenized-1	0.0	79.5	71.5	0.0	**83.6**	**98.9**	77.5
Tokenized-2	0.0	**79.6**	**71.6**	0.0	**83.6**	**98.9**	**77.6**
Tokenized-8	0.0	79.5	71.4	0.0	**83.6**	**98.9**	77.5
Surface-1	0.0	76.0	65.4	0.0	**83.6**	**98.9**	73.5
Surface-2	0.0	76.1	65.6	0.0	**83.6**	**98.9**	73.7
Surface-8	0.0	76.2	65.5	0.0	**83.6**	**98.9**	73.7

Table 3.4: AIDA: Token-Level tuning results on *STT-Dev*. (-1, -2, and -8) means that *STT-Tr* is replicated 1, 2, or 8 times respectively before adding it to *WLT*

As mentioned earlier, since the size of *STT-Tr* is much smaller than that of *WLT*, this causes both datasets to be statistically incomparable. We try increasing the weights assigned by the LM to *STT-Tr* by duplicating *STT-Tr*. We experiment with one, four, and eight copies of *STT-Tr* for each of the basic and tokenized experimental setups.

We use the shared task's evaluation script to evaluate each setup. The evaluation script produces two main sets of metrics. The first one specifies the accuracy, precision, recall, and $F_{\beta=1}$ score on the tweet-level, while the second group uses the same metrics but for each tag on the token-level. We focus on the second set of metrics since it aligns with the main goal of the token-level component of AIDA. We add an extra metric corresponding to the weighted average of the $F_{\beta=1}$ scores in order to rank the results.

As shown in Table 3.4, in all experiments the tokenized setup outperforms the surface setup, which is quite expected because it increases the coverage of the LM. Duplicating the data has no significant effect on the results but having two copies of *STT-Tr* yields a slightly better performance than the other two setups. Accordingly, we use the **Tokenized-2** setup as the standard configuration for the system.

%	MSA	EDA	Ambig	Mixed	NE	Other	**Weighted Avg-F$_{\beta=1}$**
CMU	**89.9**	**81.1**	0.0	0.0	72.5	98.1	86.4
A3-107	86.2	52.9	0.0	0.0	70.1	84.2	76.6
IUCL	81.1	59.5	0.0	0.0	5.8	1.2	61.0
MSR-India	86.0	56.4	**0.7**	0.0	49.6	74.8	74.2
AIDA	89.4	76.0	0.0	0.0	**87.9**	**99.0**	**86.8**

Table 3.5: AIDA: Token-Level evaluation averaged across the three test-sets released for the shared task at the first workshop for computational approaches to code-switching

Testing We use the three test sets described in Table 3.2 to evaluate the system. To make the comparison easier, we calculate the overall weighted F$_{\beta=1}$ score for all systems that participated in the shared task using the three test sets together. Table 3.5 shows the F$_{\beta=1}$ score of each system averaged over all three test-sets. [3] AIDA outperforms all other systems in the token-level evaluation. One interesting observation is that AIDA outperforms all other systems in identifying named-entities. On the other hand, CMU performs slightly better on MSA and much better on EDA.

Error Analysis Tables 3.6, 3.7, and 3.8 show the confusion matrices of all six tags over the three test sets. The rows represent the gold labels while the columns represent the classes generated by AIDA. In all three tables, it is clear that the highest confusability is between MSA and EDA classes. In Test-Set 1, since the majority of words (81.5%) have an MSA gold label and a very tiny percentage (0.3%) has an EDA gold label, the percentage of words that have a gold label of MSA and get classified as EDA is much larger than in the other two test-sets and much larger than the opposite case where the ones having a gold label of EDA get classified as MSA. [4] Table 3.9 shows examples of the words that were misclassified by AIDA. All of the shown examples are quite challenging. In example 1, the misclassified named-entity refers to the name of a TV show, but the word also means

[3]The results of the other participating systems were obtained from the workshop's website.

[4]We choose to include the results on this test set despite its skewed distribution, in order to conduct a fair comparison with all participating systems in the task.

		AIDA (Predicted)					
	%	Ambig	MSA	EDA	Mixed	NE	Other
Gold	Ambig	0.0	0.0	0.0	0.0	0.0	0.0
	MSA	0.0	74.4	5.7	0.0	1.3	0.0
	EDA	0.0	0.1	0.2	0.0	0.0	0.0
	Mixed	0.0	0.0	0.0	0.0	0.0	0.0
	NE	0.0	1.5	0.3	0.0	9.1	0.1
	Other	0.0	0.0	0.0	0.0	0.0	7.3

Table 3.6: AIDA: Token-Level confusion matrix for the best performing setup on *Test1* set.

		AIDA (Predicted)					
	%	Ambig	MSA	EDA	Mixed	NE	Other
Gold	Ambig	0.0	0.3	0.1	0.0	0.0	0.0
	MSA	0.0	28.8	2.8	0.1	0.2	0.1
	EDA	0.0	16.4	28.3	0.5	0.2	0.1
	Mixed	0.0	0.0	0.0	0.0	0.0	0.0
	NE	0.0	1.0	0.6	0.0	11.5	0.2
	Other	0.0	0.0	0.0	0.0	0.0	8.9

Table 3.7: AIDA: Token-Level confusion matrix for the best performing setup on *Test2* set.

		AIDA (Predicted)					
	%	Ambig	MSA	EDA	Mixed	NE	Other
Gold	Ambig	0.0	0.6	0.3	0.0	0.0	0.0
	MSA	0.0	19.0	2.9	0.0	0.5	0.0
	EDA	0.0	14.5	42.7	0.0	0.5	0.0
	Mixed	0.0	0.0	0.0	0.0	0.0	0.0
	NE	0.0	0.5	0.6	0.0	8.0	0.0
	Other	0.0	0.0	0.0	0.0	0.0	9.9

Table 3.8: AIDA: Token-Level confusion matrix for the best performing setup on *Surprise* set.

	Sentence	Word	Gold	AIDA
1	Allylp AlEA$rp w AlnSf msA' s>kwn Dyf AlAstA∗ Emrw Allyvy fy brnAmjh **bwDwH** ElY qnAp AlHyAp الليلة العاشرة و النصف مساء سأكون ضيف الاستاذ عمرو الليثي في برنامجه بوضوح على قناة الحياة	bwDwH بوضوح	NE	MSA
2	wlsh mqhwr yA EynY mn **vAbt** bA$A AlbTl wSAlH bA$A slym AllY AvbtwA An nZrthm fykm SH ولسه مقهور يا عينى من ثابت باشا البطل وصالح باشا سليم اللى اثبتوا ان نظرتهم فيكم صح	vAbt ثابت	NE	MSA
3	kfAyh $bEnA mnk AgAnyky Alqdymh jmylh lkn AlAn **lAnTyq** Swtk wlA Swrtk hwynA bqh كفايه شبعنا منك اغانيكى القديمه جميله لكن الان لانطيق صوتك ولا صورتك هوينا	lAnTyq لانطيق	MSA	EDA
4	AlrAbT Ally byqwl >ny Swrt **Hlqp** mE rAmz jlAl gyr SHyH . dh fyrws ElY Alfys bwk . rjA' AlH∗r الرابط اللي بيقول شأشني صورت حلقة مع رامز جلال غير صحيح . ده فيروس على الفيس بوك . رجاء الحذف	Hlqp حلقة	EDA	MSA
5	wAnt sAyb flAn w flAn w mAsk fy **dh** وانت سايب فلان و فلان و ماسك في ده	dh ده	MSA	EDA

Table 3.9: Examples of the words that were misclassified by AIDA

"clearly" which is an MSA word. Similarly, in example 2, the named-entity can mean *"stable"* which is again an MSA word. An example of a Mixed word that was correctly classified by AIDA is حتؤدي "*Ht&dy*" meaning *"will lead to"* where the main morpheme تؤدي "*t&dy*" meaning "lead to" is MSA and the clitic ح "textitH" meaning "will" is EDA. Examples 3 and 4 show instances of the confusability between MSA and EDA classes. Both words in these two examples can belong to either class depending on the context. We also found instances where the misclassifications were due to errors in the gold labels where a word gets classified correctly by AIDA but is assigned the wrong label in the gold data. Example 6 shows an instance of such a case where the highly EDA word *dh* has an MSA class in the gold data. Wrong gold labels such as this case, unfairly penalize the system for correct classifications. Fortunately, we only found a few instances of gold errors.

3.4 Sentence-Level Dialect Identification

The sentence-level component of AIDA uses the decisions from the token-level component along with other features to train a supervised system in order to decide upon the class of a given sentence (Elfardy and Diab, 2013). We divide the features into (1) Core Features and (2) Stylistic Features.

3.4.1 Core Features

These features indicate how dialectal–or non dialectal–a given sentence is. They are further divided into:

Token-based Features: We use the token-level decisions to estimate the percentage of EDA words and the percentage of OOVs for each sentence. These percentages are then used to derive the following features:

- **diaPercent1:** the percentage of words tagged as EDA or unknown in the input sentence to the number of all Arabic words.
- **diaPercent2:** the percentage of words tagged as EDA in the input sentence to the number of all Arabic words.
- **diaPercent:** the percentage of words tagged as EDA in the input sentence to the number of all words including the non-Arabic tokens. (ex: punctuation, numbers, Latin, etc.)

We also use the following features:

- The percentage of the words found in a canonical MSA morphological analyzer's dictionary "SAMA"
- the percentage of words found in a state-of-the-art EDA morphological analyzer's dictionary "CALIMA"
- the percentage of highly dialectal words. We create this set of features by building a lexicon of highly Egyptian words and calculating the percentage of words in the given post that exist in this lexicon.

Perplexity-based Features: We run each sentence through each of the MSA and EDA LMs and record the perplexity for each of them. The perplexity of a language model on a given test sentence $S(w_1, .., w_n)$ is defined as:

$$perplexity = (2)^{-(1/N)\sum_i \log_2(p(w_i|h_i))} \tag{3.1}$$

where N is the number of tokens in the sentence and h_i is the history of token w_i. The perplexity conveys how confused the LM is about the given sentence, so the higher the

perplexity value, the less probable that the given sentence matches the LM.

3.4.2 Stylistic Features

These are the features that do not directly relate to the dialectalness of words in the given sentence but rather estimate how informal the sentence is and include:

- The percentage of punctuation, numbers, special-characters and words written in Roman script.
- Number of words and average word-length.
- A set of binary features indicating whether–or not–the sentence has (1) consecutive repeated punctuation, (2) exclamation mark, (3) question mark, (4) word-lengthening effects, (5) diacritics, (6) emoticons and (7) decoration-effects (ex. writing ****).

3.4.3 Model Training

We use the WEKA toolkit (Hall et al., 2009) and the derived features to train a Baysian Network classifier. [5] The classifier is trained and cross-validated on the gold training data described in Table 3.3.

3.4.4 Sentence-Level Results

We conduct two sets of experiments. In the first one, we split the data into a training set and a held-out test set. In the second set, we use the whole dataset for training without further splitting in order to compare our results to those produced by Zaidan and Callison-Burch, 2011. Similar to the token-level component, we experiment with Surface and Tokenized preprocessing schemes and find that Tokenized scheme yields better results.

[5] We tried different classifiers and Baysian Networks yielded best results.

	Cross-Val. (90%)	Held-Out Test (10%)	Cross Val. (100%)
Maj-BL	51.9	51.9	51.9
Token-BL	79.1	77	78.5
Ppl-BL	80.4	81.1	80.4
OZ-CCB-BL	N/A	N/A	80.9
AIDA	**85.3**	**83.3**	**85.5**

Table 3.10: AIDA: Sentence-Level cross-validation and held-out test performance (measured in accuracy) compared against the baselines

We use four baselines. The first of which is a majority baseline (Maj-BL) that assigns all the sentences the label of the most frequently observed class in the training data (MSA). The second baseline (Token-BL) assumes that the sentence is EDA if more than 45% of its tokens are dialectal; otherwise, it assumes it is MSA.[6] The third baseline (Ppl-BL) runs each sentence through MSA and EDA LMs and assigns the sentence the class of the LM, yielding the lower perplexity value; while the last baseline (OZ-CCB-BL) is the result obtained by Zaidan and Callison-Burch, 2011 which uses the same approach of our third baseline, Ppl-BL. Table 3.10 shows the best setup's cross-validation and held-out test results against the baselines. The sentence-level component of AIDA outperforms all baselines, which indicates the robustness of the underlying approach.

3.5 AIDA-2

Both token- and system-level components of AIDA were recently improved with the introduction of the latest version of the system–AIDA-2–(Al-Badrashiny, Elfardy, and Diab, 2015). AIDA-2 outperforms AIDA on both token- and sentence-level tasks. We highlight the changes in the algorithm together with a comparison between the performance of AIDA and that of AIDA-2.

[6]We experimented with different thresholds (15%, 30%, 45%, 60% and 75%) on the cross validation set and the 45% threshold setting yielded the best performance

3.5.1 AIDA-2: Token-Level Component

The main difference in the token-level component of AIDA and AIDA-2 lies in the "Combiner" module. While AIDA uses a rule-based combiner, AIDA-2 uses a CRF-based combiner, hence treating the token-level dialect identification problem as a sequence labeling task. Similar to AIDA, AIDA-2 relies on the decisions from MADAMIRA, LM and gazetteers to decide upon the class of each token in the given sentence.

Additionally AIDA-2 uses a modality lexicon (ModLex) (Al-Sabbagh, Diesner, and Girju, 2013). ModLex is a manually compiled lexicon of Arabic modality triggers (i.e. words and phrases that convey modality). It provides the lemma with a context and the class of this lemma (MSA, EDA, or Both) in that context. The intuition behind this feature is that if a word is used in its modal sense then it can be a clue of whether the MSA or EDA meaning is the intended one. In our approach, we match the lemma of the input word that is provided by MADAMIRA and its surrounding context with an entry in ModLex. Then we assign this word the corresponding class from the lexicon. If we find more than one match, we use the class of the longest matched context. If there is no match, the word takes an "Unknown" tag. For example, the word صدق "*Sdq*", which means "told the truth", gets the class "Both" in this context أفلح إن صدق "*>flH An Sdq*" meaning "He will succeed if he told the truth". Using MADAMIRA, LM, gazetteers and ModLex, the following features are generated for each word:

- **MADAMIRA**:
  - the non-tokenized input word;
  - the prefixes, stem and suffixes;
  - the part of speech tag assigned by MADAMIRA to the surface-level word;
  - MADAMIRA's class (Based on whether the word's analysis is retrieved from SAMA or CALIMA).

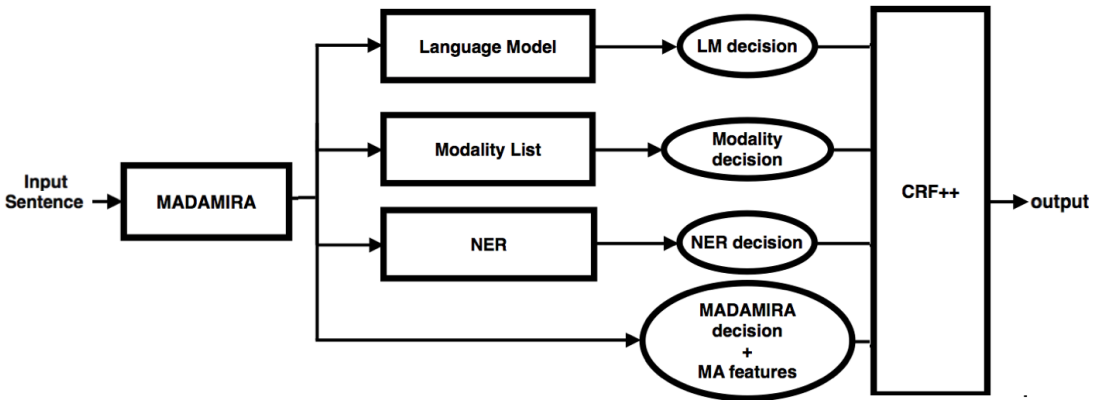

Figure 3.3: AIDA-2: Token-Level Pipeline

%	MSA	EDA	Ambig	Mixed	NE	Other	**Weighted Avg-F**$_{\beta=1}$
CMU	89.9	81.1	0.0	0.0	72.5	98.1	86.4
A3-107	86.2	52.9	0.0	0.0	70.1	84.2	76.6
IUCL	81.1	59.5	0.0	0.0	5.8	1.2	61.0
MSR-India	86.0	56.4	**0.7**	0.0	49.6	74.8	74.2
AIDA	89.4	76.0	0.0	0.0	87.9	99.0	86.8
AIDA-2	92.9	82.9	0.0	0.0	89.5	99.3	90.6
AIDA-2+	**94.6**	**88.3**	0.0	0.0	**90.2**	**99.4**	**92.9**

Table 3.11: AIDA-2: Token-Level comparison of AIDA, AIDA-2 and all other systems averaged over the three test sets released for the first Code-Switching workshop

- **LM**:
  - the classes and associated confidence scores assigned by the LM to the prefixes, lexeme and suffixes;
  - "*isMixed*" flag, which is set to "true" if the prefixes, suffixes and lexeme of a word are assigned different classes by the LM.
- **Modality**: the ModLex decision.
- **NER**: a flag indicating whether the word is assigned a *noun_prop* POS by MADAMIRA or is found in the gazetteer.
- **Stylistic:** "*isOther*" is a binary flag that is set to "true" only if the input word is a non-Arabic token. And "*hasWordLengthening*", which is another binary flag set to

"true" only if the input word has word-lengthening–speech–effects

We then use these features to train a CRF classifier using CRF++ toolkit (Sha and Pereira, 2003) and set the window size to 16. Figure 3.3 shows the pipeline for the token-level component of AIDA-2.

We compare the token-level results of AIDA and AIDA-2 using the same datasets. (Table 3.2) Additionally, we explore whether adding more training data improves the results. We add one more setup (AIDA-2+), which uses extra training data (171,419 words) that is manually annotated using the same guidelines of the shared task. We use this dataset as an extra training set in addition to the shared task's training data to study the effect of increasing the training data size on the system's performance. As the results show, using a CRF based combiner yields better results than using a rule-based system. This result is consistent with our assumption that the class of each token depends on the context from the surrounding words. Another quite expected result is that increasing the size of the training data improves the performance. Unlike the original version of AIDA, AIDA-2 outperforms CMU's system on MSA and EDA classes even without the use of extra training data. Adding more training data (AIDA-2+) results in 24.5% error reduction.

3.5.2 AIDA-2: Sentence-Level Component

Instead of relying on a single classifier, AIDA-2 relies on a classifier ensemble approach. Two sets of features are used to train two decision-tree classifiers. The generated classes and confidence scores from each of these classifiers are then used to train a third decision-tree classifier, which is responsible for making the final decision on whether a given sentence is MSA or EDA. We use with the following features for the two classifiers:

- **Comprehensive Classifier (Comp-Cl)**: The first classifier is intended to explicitly model detailed aspects of the language and includes the following features:

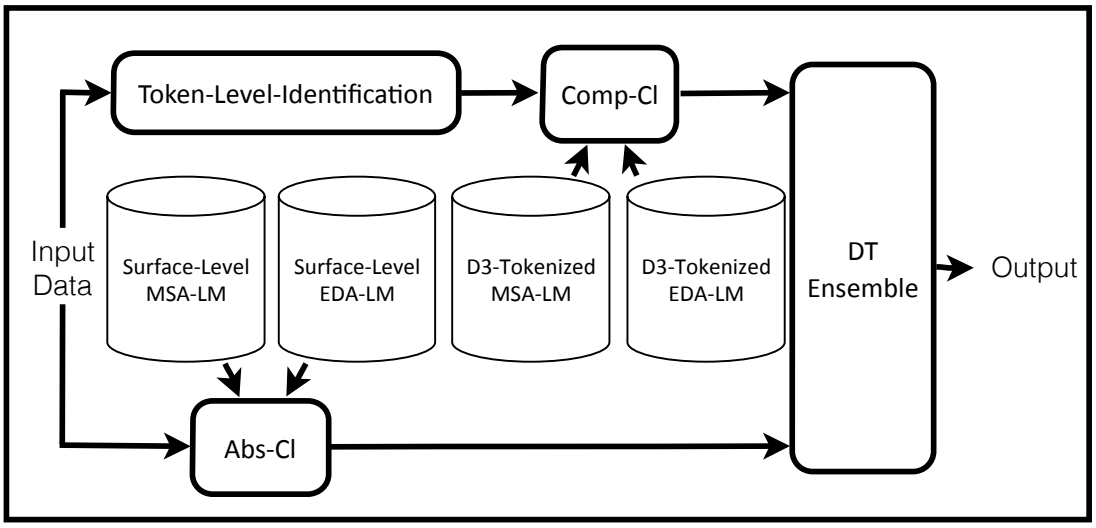

Figure 3.4: AIDA-2: Sentence-Level Pipeline

  - The same core and stylistic features described earlier in Section 3.4. (*Tokenized language models are used to derive the features for this classifier.*)
  - Modality features; the percentage of words tagged as *"EDA"*, *"MSA"*, and *"Both"* using the "ModLex" component described in Section 3.5.1

- **Abstract Classifier: (Abs-Cl)** This classifier is intended to cover the implicit semantic and syntactic relations between words. It runs the input sentence in its surface form without tokenization through a surface form MSA and a surface form EDA 5-gram LMs to get sentence probability from each of the respective LM (msaProb and edaProb). This classifier complements the information provided by Comp-Cl. While Comp-Cl yields detailed and specific information about the tokens as it uses tokenized-level LMs, Abs-Cl is able to capture better semantic and syntactic relations between words since it can see longer context in terms of the number of words compared to that seen by Comp-Cl. On average, a span of two words in the surface-level LM corresponds to almost five words in the tokenized-level LM (Rashwan et al., 2011). For each one of these two classifiers, we use a heuristic to select the most informative features.

	Cross-Val. (90%)	Held-Out Test (10%)	Cross Val. (100%)
Maj-BL	51.9	51.9	51.9
Token-BL	79.1	77	78.5
Ppl-BL	80.4	81.1	80.4
OZ-CCB-BL	N/A	N/A	80.9
AIDA	85.3	83.3	85.5
AIDA-2	89.9	87.3	**90.8**
AIDA-2+	**90.6**	**87.5**	90.1

Table 3.12: AIDA-2: Sentence-Level evaluation

- **Decision-Tree Ensemble: (DT Ensemble)** In the final step, we use the classes and confidence scores of the preceding two classifiers on the training data to train a decision tree classifier. Accordingly, an input test sentence goes through *Comp-Cl* and *Abs-Cl*, where each classifier assigns the sentence a label and a confidence score for this label. It then uses the two labels and the two confidence scores to provide its final classification for the input sentence.

Figure 3.4 shows the pipeline of the sentence-level component in AIDA-2. As Table 3.12 shows, AIDA-2 significantly outperforms AIDA and all baselines. Adding more token-level training data (AIDA-2+) yields almost the same performance as AIDA-2, which suggests that while increasing the token-level training data improves the performance of the token-level system, it is less impactful on the sentence-level component.

3.6 Related Work

Until quite recently there has been, with few exceptions (Chan et al., 2004; Diab and Kamboj, 2011; Joshi, 1982; Manandise and Gdaniec, 2011; Solorio and Liu, 2008a,b), little research in computational approaches to deal with Linguistic Code-Switching (LCS). Predictive models of how and when LCS typically occurs were not developed. A major barrier to research on LCS has been the lack of large, consistently and accurately annotated

corpora of LCS data. In fact, there has been very little discussion even of how such data should be collected and annotated to best support the interests of both the theoretical and the computational communities. However, the task gained interest with the first and second workshops for developing computation approaches to code-switching that were held recently (Molina et al., 2016; Solorio et al., 2014).

For Dialect Identification in Arabic, the task has also recently gained interest among Arabic NLP researchers. Early work on the topic focused on speech data (Biadsy, Hirschberg, and Habash, 2009) while more recent work targets textual data. The main task for textual data is to decide the class of each word in a given sentence; whether it is MSA, DA or some other class such as Named-Entity or punctuation, and whether the whole sentence is mostly MSA or DA.

The most recent work on sentence-level Dialectal Arabic identification–beside ours–includes those of Zaidan and Callison-Burch, 2011, Cotterell and Callison-Burch, 2014 and Darwish, Sajjad, and Mubarak, 2014. Zaidan and Callison-Burch, 2011 annotate MSA-DA news commentaries on Amazon Mechanical Turk and explore the use of a language-modeling based approach to perform sentence level dialect identification. They target three Arabic dialects; Egyptian, Levantine and Gulf, and develop different models to distinguish each of them against the others and against MSA. They achieve an accuracy of 80.9%, 79.6%, and 75.1% for the Egyptian-MSA, Levantine-MSA, and Gulf-MSA classification, respectively. These results support the common assumption that Egyptian, relative to the other Arabic dialectal variants, is the most distinct dialect variant of Arabic from MSA. Cotterell and Callison-Burch, 2014 extend the work of Zaidan and Callison-Burch, 2011 by handling two more dialects–Iraqi and Moroccan–and targeting a new genre, tweets. Their system outperforms ibid., achieving a classification accuracy of 89%, 79%, and 88% on the same Egyptian, Levantine and Gulf datasets.

Darwish, Sajjad, and Mubarak, 2014 use lexical, morphological, phonological, and syntactic features to perform sentence-level EDA identification. They test their system on tweets and achieve an accuracy of 94.6%.

For token-level Dialectal Arabic identification, as mentioned earlier, the shared task at the "First Workshop for Computational Approaches to Code-Switching" addressed token-level code-switch detection between several language pairs including MSA-EDA. Four systems participated in the task. The first system–IUCL–(King et al., 2014) used a language-independent approach that utilizes character n-gram probabilities, lexical probabilities, word label transition probabilities and existing named-entity recognition tools within a Markov model framework. Another system–IIIT–(Jain and Bhat, 2014) used a CRF based token-level language identification system that uses a set of easily computable features (ex. isNum, isPunc, etc.). The authors' analysis showed that the most important features are the word n-gram posterior probabilities and word morphology. MSR-India (Chittaranjan et al., 2014) built a system that uses character n-grams to train a maximum entropy classifier that identifies whether a word is MSA or EDA. The resultant labels were then used together with word length, existence of special characters in the word, current, previous and next words to train a CRF model that predicts the token-level classes of words in a given sentence/tweet. CMU (Lin et al., 2014) used a CRF model that relies on character n-grams probabilities (tri and quad grams), prefixes, suffixes, unicode-page of the first character, capitalization case, alphanumeric case, and tweet-level language ID predictions from two off-the-shelf language identifiers: cld2[7] and ldig.[8] They increased the size of the training data using a semi-supervised CRF autoencoder approach (Ammar, Dyer, and Smith, 2014) coupled with unsupervised word embeddings. While

[7]https://code.google.com/p/cld2/

[8]https://github.com/shuyo/ldig

there is no capitalization feature in Arabic, the system was targeted for other language pairs as well for which the capitalization feature can be useful. As previously shown in the experiments, our latest version of the system–AIDA-2–outperforms all of the previous systems.

More recently, a more language-independent approach–LILI–(Al-Badrashiny and Diab, 2016) was presented and evaluated on the same dataset. The authors relied on character-level and word-level n-grams to identify whether each word in a given input is *lang1/MSA* or *lang2/EDA*. The approach does not identify any of the other classes–*Ambig*, *Mixed*, *NE* or *Other*. The system achieved an $F_{\beta=1}$ score of 82% and 86.8% on *lang1/MSA* and *lang2/EDA* classes respectively resulting in a weighted average $F_{\beta=1}$ score of 85% on the two classes. AIDA-2, on the other hand, achieved an $F_{\beta=1}$ score of 92.9% and 82.9% on the two classes and an average $F_{\beta=1}$ score of 90.2%. So, while LILI outperforms AIDA-2 on *lang2/EDA* identification–by 2.1%, AIDA-2 outperforms LILI on *lang1/MSA* with 10.9% resulting in a much higher overall performance of AIDA-2.

Other closely related research efforts try to create multidialectal Arabic parallel corpora (Bouamor, Habash, and Oflazer, 2014), normalize the orthography of DA to a more standardized form (Eskander et al., 2013), cluster orthographic variants in DA (Dasigi and Diab, 2011), as well as convert romanized Arabic–Arabic written in Latin script–to Arabic script (Eskander et al., 2014). In our work, we do not model romanized Arabic and assume that any word written in a Latin script is a non-Arabic word.

Salloum and Habash, 2011, on the other hand, target the problem of DA to English Machine Translation (MT) by pivoting through MSA. The authors present a system that applies transfer rules from DA to MSA then uses state-of-the-art MSA to English MT system. In collaboration with the authors of this work (ibid.), we explored the use of

Dialect Identification on the performance of Arabic to English MT and found that it improves the results. Four different SMT systems were trained; (a) DA-to-English SMT, (b) MSA-to-English SMT, (c) DA + MSA-to-English SMT, and (d) DA-to-English hybrid MT system, treating the task of choosing which SMT system to invoke as a classification task. Using AIDA, various features–that indicate, among other things, how dialectal the input sentence is–were derived and used to train a classifier that learns to choose the correct MT system. Using this AIDA-based approach improved the performance by 0.9% BLEU points (Salloum et al., 2014).

3.7 Summary

In this chapter, we described AIDA, our system for token and sentence-levels code-switch detection between Egyptian Dialectal Arabic (EDA) and Modern Standard Arabic (MSA). Given a sentence, AIDA decides whether the sentence is predominantly MSA or EDA. Additionally, it identifies the class–whether MSA, EDA or some other class–of each word in the given sentence. We described the datasets used to train and evaluate both components of the system and compared the performance of these systems to other systems that perform the same task. Additionally, we introduced the newer version of AIDA–*AIDA-2*–which improves the performance on both tasks. In the next chapter, we will study whether or not utilizing code-switching features derived from AIDA can be used as an indicator of a given comment's perspective in Egyptian discussion fora.

Chapter 4

Perspective Identification in Arabic

As previously discussed, there are various elements governing the belief system of people that affect their stance on different ideological topics. In this chapter we describe our work on automatic perspective identification in Arabic discussion fora. We begin by explaining our process for building a taxonomy of the common community perspectives in Egypt and describe how we use this taxonomy in creating annotation guidelines and collecting large-scale annotations in Section 4.1. In Section 4.2 we formally describe our five classification tasks before describing our approach in Section 4.3. Next, we present the experiments along with their results in Section 4.4. Finally, we summarize the chapter in Section 4.5.

4.1 Egyptian Ideological Perspective

While collecting annotations of ideological perspectives is generally challenging due to the inherent subjectivity of the task, it is much more challenging in dynamic political settings such as the Egyptian one where the political stances themselves are emergent and shifting and where a clear taxonomy for the common community perspectives and ideologies is absent. In this case the problem becomes two-fold: (1) pinning down what the perspectives are; and, (2) gathering annotations on such perspectives while circumventing the subjectivity of the annotators themselves. Accordingly, we follow an iterative process for creating a robust and succinct set of guidelines for annotating "Egyptian Ideological Perspectives" that aim at decoupling the annotation process from possible subjective assessment of the annotators (Elfardy and Diab, 2016a). We build a list of major political events and sample

a set of discussion fora data that was posted within one week from the start of each of these events. We come up with a hypothesis on the most important elements governing the Ideological Perspective of most Egyptians and develop a set of guidelines and an annotation task to identify the perspective from which a given comment was written. Our hypothesis is that a person's perspective has two major underlying dimensions: (1) a person's stance on political reform versus stability; and, (2) a person's stance on the role Islam/religion should play in politics. We run our first annotation experiment where we ask annotators to identify the stance of a given comment towards several political entities such as January 25^{th} Revolution, Mubarak's Regime, Military Rule, Islamists and Secularists. Based on the feedback and error analysis of this pilot annotation, we note some interesting observations, the most impactful of which is the annotators having significant reservations in making a judgment on comments. Taking this feedback into consideration, we refine the guidelines for the annotation task. We have the same set of comments annotated based on the refined guidelines and note a significant increase in inter-annotator agreement measures from 75.7% to 92% overall agreement. Using this final set of guidelines, we annotate the final dataset of 5,000 comments that we use for building and evaluating supervised systems aimed at automatically identifying Egyptian Ideological Perspectives.

4.1.1 Data Collection

We select a set of public discussion fora pages of Egyptian activists and politicians of different political leanings: Revolutionists, Muslim Brotherhood leaders, Seculars, and Mubarak supporters. We curate posts and comments from these pages. The "*post*" refers to some piece of content shared on a page while the "*comment*" is a response to this original piece of content. We filter spam/repetitive comments that do not respond to the original post. Moreover, only comments with no Latin words and ones that have a length of at least ten words are preserved.

Event	Date Range
1. January 25^{th} Revolution	Jan. 25 - Jan. 31, 2011
2. Battle of the camel	Feb. 2 - Feb. 8, 2011
3. Mubarak Stepping Down	Feb. 11 - Feb. 17, 2011
4. Referendum on amendments to old constitution	Mar. 19 - Mar. 25, 2011
5. Mohamed Mahmoud Protests *(Clashes between Army & Rev.)*	Nov. 19 - Nov. 25, 2011
6. Announcement of presidential election results	Jun. 24 - Jun. 30, 2012
7. Presidential decree and associated protests	Nov. 22 - Nov. 28, 2012
8. Ousting of President Mohamed Morsi	Jun. 30 - Jul. 6, 2013
9. Army calls for mandate to crack down on terrorism	Jul. 24 - Jul. 30, 2013
10. Rabia (Pro-Muslim Brotherhood) camp dismantling	Aug. 14 - Aug. 20, 2013

Table 4.1: List of events covered by the Egyptian dataset

After the initial cleanup of the data, we use a list of major events such as January 25^{th} 2011 demonstrations, major protests, Presidential elections, etc. to select our final dataset. Table 4.1 shows the list of events and the dates covered by the selected data. We split the data into two groups based on whether it was curated from a page that supports (1) Reform [RFM] (Supporting January 25^{th} Revolution); or, (2) Old Guard Rule [OGR] (ex. Supporting the ousted Egyptian President Mubarak and his regime, or supporting the current Egyptian President–Sisi–who was the ex-Minister of Defense). We then select a sample of 31 comments per event for each one of the two groups. Since no comments were posted in the pro-OGR pages for the first event, we only have 31 pro-RFM comments for this event. This results in a total of 310 RFM and 279 OGR comments.

4.1.2 Taxonomy of Egyptian Ideological Perspectives

Prior to collecting the annotations, we come up with a high level taxonomy for the most common political leanings in Egypt for this timeframe. We base our taxonomy on the works of "The Hariri Center at the Atlantic Council" and "Carnegie Endowment

for International Peace" (Brown, 2013; *Carnegie Endowment for International Peace*; Carothers and Brown, 2012; *The Hariri Center at the Atlantic Council*).

As mentioned earlier, after January 25^{th} Revolution, the formation and breakup of alliances between different political entities resulted in a dynamic set of political leanings, and hence created a need for a dynamic classification. For the context of this thesis, we reduce the very rich perspective map of a person to two underlying dimensions: (1) stance towards democracy and political reform versus stability at the expense of loss of civil liberties; (2) stance towards the role played by Islam/religion in the public sphere or politics, namely Islamist vs. Secular. Accordingly, we assume that these two dimensions constitute a person's perspective. For example, a person can oppose involving Islam in politics and support political reform. Another person can focus on stability even if it ushers in an autocratic regime while either supporting or opposing Islamists. As stated by Converse (Converse, 1962), the dimension that is less central to a person's belief system is more likely to change over time.

Annotation Procedure

Noting how challenging the annotation will be, we wanted to get a sense of how to circumvent annotator bias. Accordingly, we devise an iterative feedback loop for the annotation process. We first have the sampled comments annotated by four trained Egyptian annotators. We ask the annotators to self-identify what their own positions are with respect to the two dimensions of interest. All annotators indicate that they support January 25^{th} Revolution. Additionally, three annotators–annotators 1-3–indicate that they are neutral towards the role of Islam in politics, while the fourth annotator indicates support towards the Army's leadership in ousting Islamists. An annotation lead managed the process of (1) training the annotators, (2) relaying their feedback about the clarity of the task to us. Based on the feedback and inter-annotator agreement (IAA) from this round, we refine the

guidelines and annotation task before having the same data annotated by the same set of annotators.

Pilot Annotation Experiment

For each task, we present annotators with a post and an associated comment. Except for one optional question that asks for feedback about the overall annotation task, all questions are formatted as multiple choice and require one answer to be provided. In order not to bias the judgments of the annotators, we do not reveal the leaning of the source page from which the comments were curated. Annotators were asked to answer the following questions for each task:

- Q1: Does the given *comment* discuss Egyptian politics? (Yes/No)
- Q2: Is there enough context to determine the political leaning of the *comment*? (Yes/No)

Does the given *comment* Support/Oppose/Not Sure/Not Applicable:

  - Q3: January 25^{th} 2011 Revolution?
  - Q4: Mubarak's regime?
  - Q5: Seculars?[1]
  - Q6: Islamists?
  - Q7: Military Rule?

- Q8: Do you have any feedback or suggestions?

Questions 3-7 aim to identify the two previously discussed dimensions that define a person's perspective. Questions 3, 4 and 7 attempt to uncover the first dimension–the person's position on political reform and democracy while questions 5 and 6 aim to identify the second dimension–the person's view on the role of Islam/religion in the public

[1]We choose to ask about "Seculars" despite the negative connotation of the term among some Egyptians. But we explain the intended meaning as being akin to Liberals and specifically that it is nothing more than the exclusion of religion from public governance.

political sphere/government.

Since the task is quite subjective, we tried to cover most possible scenarios and to provide examples in our guidelines in order not to rely on the annotators' subjective assessments of the different scenarios. Moreover, we attempted, to the best of our knowledge, to avoid any bias in the way the questions were phrased. Figure 4.1 shows the guidelines for this pilot annotation experiment.

Error Analysis: We calculate the pairwise and overall IAA for all questions. Table 4.2 shows the results. The average pairwise IAA for all questions is quite high, ranging from 84.1% to 88.4%. However, achieving a complete-row agreement (Row) by all annotators is quite challenging. The four annotators achieved a perfect row agreement–chose the same answers for all questions pertaining to a particular comment–on only 25.5% of the comments. We also note that Annotators 1 and 3 exhibit the most agreement.

In order to get better insights into the source of disagreement between annotators, we performed a manual error analysis by looking into the confusable comments and found that the examples we looked at fall under the following categories:

1. Comments that provide clues for both supporting and opposing the topic the question is addressing.

   ex. (Event 2)

   لازم نصبر ونشوف اللى حيحصل مفيش حاجة بتتغير بين يوم وليلة براحة علشان اللى عملناه ميتقلبش ضدنا يا جماعة خافوا على البلد شوية عايزين نعمرها تانى

   Translation: *We have to be patient and wait and see what will happen. Nothing changes in a day and night. Take it easy so what we did does not backfire on us.*

- All questions target the *comment*. (The *post* is meant to give you context)
- Please pay attention to the *post*'s and *comment*'s dates.
- Use your knowledge of the political events in Egypt when responding to the questions.

  ex. If a *comment* supports January 25^{th} Revolution and you know that this implies that it opposes Mubarak's regime then choose "*Oppose*" as an answer to Q4.
- If the answer to Q1 or Q2 is "*No*", then choose "*Does not apply*" as an answer to all other questions.
- Difference between "*Does not apply*" and "*Not Sure*":
  - "*Does not apply*" should be used when the *comment* does not discuss the subject of the question.

    ex. If a given *comment* does not discuss Mubarak's regime then you should choose "*Does not apply*" as an answer to Q4.

    If, on the other hand, the *comment* discusses Mubarak's regime but you are not sure whether it opposes it or supports it then choose "*Not Sure*".
- Q7 targets Military Rule at any point in time (not a specific Army leader).
- If a *comment* supports Islamists this does not necessarily mean that it opposes Seculars and vice versa. (Unless the author expresses anti-secular views)
- If you have any feedback, please respond to Q8.

Figure 4.1: Synopsis of annotation guidelines for Pilot annotation task

	Q1 Egy. Polit.	Q2 Context	Q3 Jan. 25^{th}	Q4 Mubarak	Q5 Seculars	Q6 Islamists	Q7 Military Rule	Avg	Row
Ann.1-2	95.6	81.8	80.3	76.1	96.8	80.6	79.5	84.4	44.3
Ann.1-3	96.8	87.6	81	83.4	97.8	86.1	86.1	88.4	55.5
Ann.1-4	97.1	86.1	81.2	81.5	97.5	83	78.8	86.4	48.6
Ann.2-3	95.8	82.3	77.2	72.8	97.3	82.7	80.3	84.1	42.3
Ann.2-4	96.4	83.5	87.3	79.8	98.8	82.7	78.8	86.8	47.5
Ann.3-4	98	84.9	80	81.2	97.3	84.9	81.3	86.8	49.4
All Ann.	93.5	71.1	66.9	64.3	95.9	72.2	65.9	75.7	25.5

Table 4.2: Inter-annotator agreement for the Pilot annotation experiment

	Pro-RFM Pages				Pro-OGR Pages			
	Yes		No		Yes		No	
Q1. Egy. Politics	97.6		2.4		97.8		2.2	
Q2. Context	84.4		15.6		81.5		18.5	
	Support	Oppose	Not Sure	NA	Support	Oppose	Not Sure	NA
Q3. Jan. 25^{th}	42.9	2.6	0.3	54.2	3.9	32.2	0.8	63.1
Q4. Mubarak	1.9	43.5	1.5	53.1	30.1	7.7	1.5	60.7
Q5. Seculars	0.2	2.6	0.2	96.9	0	1.3	0.3	98.4
Q6. Islamists	27.7	11.2	1.7	59.4	9.9	33.9	0.8	55.5
Q8. Military Rule	1.2	11.3	0.6	86.9	22.6	5.8	0.4	71.1

Table 4.3: Answer distribution (averaged over all annotators) for each question in the Pilot annotation split according to the leaning of the source page from which the data is curated

Care about the country. We need to rebuild it.

While above comment opposes the continued demonstrations, this does not necessarily mean that it opposes January 25^{th} Revolution since the author just prioritizes stability over immediate political reform.

2. Ambiguous pronouns.

   ex. (Event 9)

   وقيادتهم اللى دفعوا بهم للتجارة بدمهم اليسوا اول المسئولين؟ واول من يستطيعوا ان يوقفوا اهدار دمهم وعدم دفعهم للانتحار

   Translation: *And their leaders that pushed them in order to sell their blood, aren't they the responsible ones? They could have stopped their bloodshed if they didn't push them to commit suicide.*

   In this comment, although "their leaders" refers to leaders of the Muslim Brotherhood, it can be easily confused for the Army leaders.

3. Comments where the stance towards one entity is implied from the stance towards another entity.

   ex. (Event 6)

   يا ما انا فرحانة فيكوا يا وفي شفيق العتيه يا عبيييييد يا حرامية

   Translation: *I am gloating over the loss of the idiot Shafik, you slaves and thieves*

   In the above comment, the author gloats over the defeat of Ahmed Shafik (a key

figure of the OGR) in the 2012 presidential elections. While the comment clearly supports January 25^{th} Revolution and opposes Mubarak's regime, it is not clear whether or not the author actually supports the Muslim Brotherhood's candidate. Our guidelines did not address such case.

4. Authors that report the opinions of other people by quoting them instead of stating their own opinions in comments they post.

   ex. (Event 7)

   وكالات انباء عالمية: عدد المتظاهرين المعارضين لمرسي في التحرير يفوق عدد مؤيديه عند الاتحادية.

   Translation: *International News Agencies: "The number of anti-Morsi protestors in Tahrir exceeds the number of his supporters at the Heliopolis Palace"*

5. Sarcastic comments where the annotator judges the comment based on the literal and not the intended meaning;
6. Comments that oppose a certain group of Islamists (ex. Muslim Brotherhood) and support other ones (ex. Salafis). To handle these cases, the annotation task should provide a "Mixed Views" option to Q6 (a comment's stance on Islamists).

Qualitative Assessment: To perform a qualitative assessment of the annotations, we begin by calculating the distribution of the answers to all questions. We further split the comments according to whether the source pages they were collected from support OGR or RFM. One should note that even if a page supports democracy this does not necessarily mean that all people who comment on that page share the same views. However, we do expect a higher number of pro-RFM authors to comment on the pro-RFM pages and vice versa. Table 4.3 shows the distribution. By analyzing the responses, we find that the majority of the given comments (>97%) discuss Egyptian politics, which indicates that our filtration process works well in excluding spam and irrelevant comments. Moreover,

the majority of comments (>84%) provide enough context to determine their stance. Another observation is that annotators are very conservative in using the "*Not Sure*" category. As expected, we find a much higher percentage of comments that support Mubarak's regime and Military Rule and oppose January 25^{th} Revolution among the ones collected from pro-OGR pages. On the contrary, the majority of comments from pro-RFM pages that express a stance towards the different political entities support January 25^{th} Revolution and oppose both Military Rule and Mubarak's Regime. While pro-RFM pages have a higher percentage of comments that support Islamists (27.7%), and pro-OGR pages have a higher percentage of anti-Islamists comments (33.9%), a considerable number of comments in each of these pages follow the opposite trend–11.2% of comments in pro-RFM pages oppose Islamists and 9.9% of those in pro-OGR pages support them. This can be attributed to the constantly changing relation between Islamists and other political entities such as the Military, the Police and Revolutionists.

We analyze the answers per event and find that the distribution of the answers aligns with our knowledge of the political events in Egypt. For example, we expect and find a higher percentage of "*Does not apply*" for Q4 (Mubarak's Regime) as we move away from the start of January 25^{th} Revolution and more polarization on the stance towards Islamists for events 8 through 10. Almost all comments pertaining to the first three events do not convey any stance towards Islamists. In the days right after the start of January 25^{th} Revolution, most of the discussions addressed political reform versus stability and not the role of religion in politics. For "Event 6" (announcing the results of Presidential elections in which the Muslim Brotherhood's candidate was elected) the majority of comments sampled from pro-RFM pages support Islamists indicating acceptance of the election outcome, while the pro-OGR pages express negative stance towards Islamists indicating disappointment in election outcomes; namely, disappointment that the OGR candidate–former Prime Minister–Ahmed Shafik lost.

Pilot Annotation Weaknesses: Based on the feedback collected from the annotators and our manual error analysis, we notice the following problems with the way the task is formulated:

- The main point of confusion among annotators is deciding when they should infer the stance of the comment towards an entity based on the stance towards another entity. For example, if a person opposes the Army during Morsi's presidency term, does it imply that he/she supports Islamists?;
- The task does not model people whose main priority is stability regardless of political reform or the role of religion in politics;
- Even though the comments were collected from a specific set of events, we do not present the annotators with the event each comment is discussing; and rather; we rely on the comment's date and the annotators' knowledge of the timeline of political events in Egypt, which resulted in the task being more challenging and time consuming;
- Q7 (A comment's stance on Military Rule) relies to a great extent on each annotator's interpretation of the notion of Military Rule. A better way to phrase the question is to simply ask about the comment's stance towards the Military leaders and tap into our knowledge of the political timeline in Egypt in order to identify the periods where the Army/Military was actually in charge of governance;
- Most of the comments we looked at express the author's top priority, whether it is political reform, stability, supporting the army, opposing the intervention of religion in political governance, etc., but our task gives equal weight to all political entities and does not ask the annotators to identify the top priority that they think drives the author's stance on various issues;
- Annotators were tempted to choose "*Does not apply*" for many comments because they were trying to identify the reason behind a comment's stance. For example, a comment might support Islamists during Rabia camp dismantling because the author is against civil rights infringement but not necessarily because that person is pro-Islamists in gen-

eral. We made clear to the annotators that we are only interested in the stance of the given comment at the time of the event of interest, namely in the specific context of the comment, regardless of the reason behind this stance or the person's stance at other points in time. Hence, this changed the question from a potential confusable "why" question to a "what" question. As mentioned earlier, this might also reflect the annotators' own concern over expressing their opinion about the comments with such a contentious event, erring on the side of caution;

- Some annotators chose "Yes" as an answer to Q2 (Is there enough context to judge the comment) when they were able to identify the sentiment of the comment but not the target of the sentiment. We clarified that if knowing the target is needed to identify the leaning of the comment then they should choose "No" as the answer to Q2;
- The guidelines do not address the cases where a comment shows mixed views on different Islamist groups/parties;
- Finally, the task does not address how the cases of reported opinions should be handled.

4.1.3 Refined Annotation Experiment

In order to mitigate the sources of confusion in the original pilot guidelines, we came up with event-based guidelines where we clarify for each event whether or not the annotators should draw correlations between different entities. This is needed in order to rely less on each annotator's political leaning and more on the presented set of rules–within our guidelines–on how to draw these correlations. Additionally, we ask annotators to identify the priority expressed by the comment and change the questions and answer choices as follows:

- Q1: Does the given *comment* discuss Egyptian politics? (Yes/No)
- Q2: Is there enough context to identify the political leaning of the *comment*? (Yes/No)
- Q3: Does the *comment* report the opinion of another person/entity and not the opinion of

the author of the comment? (Yes/No/None)

- Q4: Which of the following do you think is the top priority for the comment: (1) Supporting January 25^{th} Revolution; (2) Stability; (3) Supporting Mubarak's Regime; (4) Supporting the Military; (5) Supporting Islamists; (6) Opposing Islamists; (7) Cannot determine the priority; (8) None.
- Q5: What is the *comment*'s stance (Support/Oppose/None) on January 25^{th} Revolution?
- Q6: What is the *comment*'s stance (Support/Oppose/None) on Mubarak and his regime?
- Q7: What is the *comment*'s stance (Support/Oppose/None) on the Military leaders during the period the *comment* was posted in?
- Q8: What is the *comment*'s stance Support/Oppose/Mixed/None) on Islamists? (

We split the comments according to the event they discuss and present the annotators with 10 sub-tasks for each one of the 10 events. Additionally, we clarify the following in the refined guidelines:

- When choosing "No" as an answer to Q1 or Q2, choose "None" for Q3-Q8;
- For Q4, choose "Can't determine the priority" when there is more than one priority in the *comment* and you cannot choose between them;
- For Q5-Q8, choose "None" if you cannot determine the leaning of the comment towards the entity in question;
- For all questions, if the *comment* expresses an opinion towards January 25^{th} Revolution or Mubarak's regime but not both of them, in most cases you can assume that supporting January 25^{th} Revolution implies opposing Mubarak's regime and vice versa;
- If a *comment* reports an opinion of another person/entity without opposing it, indicate in Q3 that it is a reported opinion then assume for all other questions that the reported opinion expresses the opinion of the author of the *comment*.
- For event 6:
  - Opposing the OGR candidate Ahmed Shafik does not imply supporting the Islamist

candidate Mohamed Morsi, while supporting Ahmed Shafik implies opposing Mohamed Morsi.

  - Similarly, opposing Mohamed Morsi does not imply supporting Ahmed Shafik while, supporting Mohamed Morsi implies opposing Ahmed Shafik.

- For events 9 and 10, if a *comment* expresses an opinion towards the Military or Islamists (not both of them), in most cases you can assume that supporting Islamists implies opposing the Military and vice versa.

It is worth mentioning that for Q4, except for opposing Islamists, we only address what a comment supports (not opposes). We did an exercise where we annotated 400 comments ourselves and found that for many comments the most central element to the belief systems of the authors is whether or not Islam/religion should be involved in politics. A person who supports RFM might temporarily support OGR if it guarantees ousting Islamists from the political scene and vice versa. Moreover, for all other entities (January 25^{th} Revolution, Mubarak, Army, etc.) one can infer what a person opposes based on what this person supports and the event that is being commented on.

Results of Refined Annotation: Table 4.4 shows the IAA for the second annotation experiment. As expected, Q4 has a lower IAA than all other questions. Overall, the new task yields a much higher agreement. The complete row agreement (Row) jumps from 25.5% to 76.9% and the average question agreement jumps from 75.7% to 92% comparing the pilot annotations to the refined annotations. Additionally, we get a much more positive feedback from the annotators on the clarity of the task. Tables 4.5 and 4.6 show the distribution to all answers in the second annotation experiment. While the distribution of the answers to Q1 almost remained the same, the distribution of Q2 answers changed. We attribute this to our emphasis on what constitutes enough context in the modified guidelines.

	Q1 Egy Polit.	Q2 Context	Q3 Reported Op.	Q4 Priority	Q5 Jan.25	Q6 Mubarak	Q7 Army	Q8 Islamists	Avg	Row
Ann. 1-2	99.3	95.1	95.1	89.5	92.9	92.9	95.4	94.9	94.4	82.7
Ann. 1-3	99.2	97.5	97.1	92.9	94.2	94.2	96.3	95.6	95.9	86.2
Ann. 1-4	99.2	94.6	94.1	88.6	91.7	91.9	94.6	94.4	93.6	80.8
Ann. 2-3	99.5	95.6	95.4	89.5	94.4	94.4	96.1	94.7	94.9	83.9
Ann. 2-4	99.5	97.1	96.6	92.0	95.9	95.8	97.1	95.6	96.2	86.9
Ann. 3-4	100.0	95.4	95.2	91.5	95.1	94.9	96.4	96.3	95.6	87.9
All-Ann.	99.0	93.2	92.9	85.2	90.5	90.5	93.5	92.4	92.1	76.9

Table 4.4: Inter-annotator agreement for the Refined annotation experiment

	Pro-RFM Pages				Pro-OGR Pages			
	Yes		No		Yes		No	
Q1. Egy Politics	97.7		2.3		97.9		2.1	
Q2. Context	85.9		14.1		87.7		12.3	
	Yes	No	None		Yes	No	None	
Q3. Rep. Opinion	2.3	83.6	14.1		0.4	87.3	12.3	
	Support	Oppose	None		Support	Oppose	None	
Q5. Jan. 25^{th} Rev.	46	3.6	50.4		12.2	44.2	43.6	
Q6. Mubarak	3.6	45.9	50.5		44	12.4	43.6	
Q7. Army	23.1	9.7	67.3		25	5.2	69.8	
	Support	Oppose	Mixed	None	Support	Oppose	Mixed	None
Q8. Islamists	29.4	12.9	57.3	0.3	12	37.7	0	50.3

Table 4.5: Answer distribution (averaged over all annotators) for questions Q1-3 and Q5-Q8 in the refined annotation experiment split according to the leaning of the source page from which the data is curated

	Pro-RFM	Pro-OGR
Jan. 25^{th} Rev.	33.5	3.3
Support Mubarak	0.6	31.5
Support Stability	9.4	7.8
Support Army	1.1	6.8
Support Islamists	28.5	11.5
Oppose Islamists	11.7	26.5
Can't Tell	1	0.4
None	14.1	12.3

Table 4.6: Answer distribution (averaged over all annotators) for Q4 (Identify the priority of the comment) in the Refined annotation experiment split according to the leaning of the source page from which the data is curated

	Comments	Tokens	Types	Tokens/Comment
Train	4,000	139,286	32,732	35
Dev.	500	17,326	7,097	35
Test	500	17,231	7,229	34

Table 4.7: Statistics of the training, development and test sets in the Final dataset

4.1.4 Egyptian Ideological Perspective Final Dataset

Using the refined version of the guidelines, we build the final dataset which we use in training and evaluating our Arabic "Perspective Identification" systems. We collect a total of 5,000 comments. Similar to the pilot annotation, the same list of events highlighted in Table 4.1 is used to select our data. We select a set of 500 comments for each event. For all events except the first one, which does not have any data from pro-OGR pages, half of the comments come from pro-OGR pages and the other half comes from pro-RFM pages. Then, for each event and each class of pages–pro-OGR and pro-RFM–we split the data into 80% training, 10% development and 10% testing before combining these smaller sets to form our final training, development and test sets. We choose this strategy

Training Set

	Yes		No					
Q1	97.3		2.6					
Q2	89.3		10.7					
	Yes		No		None			
Q3	1.7		87.4		10.9			
	Jan. 25^{th}	Mubarak	Stability	Army	Islamists	Opp. Islamists	Ambig.	None
Q4	24.9	16.4	4.6	7.5	14.2	18.1	3.3	10.9
	Support		Oppose		None			
Q5	29.3		22.1		48.6			
Q6	23.9		31.3		44.9			
Q7	20.7		15.6		63.8			
	Support		Oppose		Mixed		None	
Q8	16.3		25.9		0.2		57.6	

Table 4.8: Answer distribution for questions Q1-Q8 in the training set of the Final dataset

to split the data as opposed to a random split in order to have each leaning and each event equally represented in the training, tuning and test sets. Table 4.7 shows the statistics of the training, development and test sets while Tables 4.8-4.10 show the distribution of the answers to the different questions in these three sets.

Similar to the pilot annotation, most of the comments discuss Egyptian politics and provide enough context to judge their leaning. Additionally, the majority of comments (~85.8%) clearly express the issue of priority to the author. For each political entity addressed by the guidelines, almost half of the comments express a stance towards it. As mentioned earlier, the events cover a long time-frame, and not all entities were a subject of controversy and discussion throughout the whole timeline. Overall, we think that the dataset quite accurately represents the political spectrum in Egypt during the studied

	Development Set							
	Yes		No					
Q1	96.8		3.2					
Q2	88.8		11.2					
	Yes		No		None			
Q3	1.0		87.0		12.0			
	Jan. 25^{th}	Mubarak	Stability	Army	Islamists	Opp. Islamists	Ambig.	None
Q4	29.6	13.6	2.2	6.4	15.2	18.6	2.4	12.0
	Support		Oppose		None			
Q5	36.2		20.2		43.6			
Q6	20.4		39.0		40.6			
Q7	20.4		13.4		66.2			
	Support		Oppose		Mixed		None	
Q8	16.2		23.6		0.0		60.2	

Table 4.9: Answer distribution for questions Q1-Q8 in the development set of the Final dataset

timeline.

4.2 Computational Tasks

Based on the collected annotations, we define the following five classification tasks that try to uncover a person's perspective from a given written comment.

- **Task 1: Priority of a given comment:** In this task, we aim to identify the priority expressed by the given comment which can be thought of as the most central element to the person's perspective. The possible classes are: (1) Stability, (2) Supporting January 25^{th} Revolution, (3) Supporting Mubarak, (4) Supporting the Military Leaders, (5) Supporting Islamists, (6) Opposing Islamists, (7) Ambiguous and (8) None;

	Held-Out Test Set							
	Yes		No					
Q1	97.0		3.0					
Q2	87.0		13.0					
	Yes		No		None			
Q3	1.4		85.4		13.2			
	Jan. 25^{th}	Mubarak	Stability	Army	Islamists	Opp. Islamists	Ambig.	None
Q4	29.0	15.2	2.4	6.4	14.2	17.6	2.0	13.2
	Support		Oppose		None			
Q5	38.0		20.6		41.4			
Q6	20.2		39.8		40.0			
Q7	19.0		14.6		66.4			
	Support		Oppose		Mixed		None	
Q8	15.8		22.8		0.0		61.4	

Table 4.10: Answer distribution for questions Q1-Q8 in the held-out test set of the Final dataset

- **Task 2: Stance on January 25^{th} Revolution:** This task aims to identify whether a given comment (1) Supports, (2) Opposes, or is (3) Indifferent–None–towards January 25^{th} Revolution.
- **Task 3: Stance on Old Guard Rule (OGR):** This task aims to identify whether a given comment (1) Supports, (2) Opposes, or is (3) indifferent–None–towards Mubarak and his regime ;

- **Task 4: Stance on the Army's Leaders:** This task aims to identify whether a given comment (1) Supports, (2) Opposes, or is (3) indifferent–None–towards the Military leaders;

- **Task 5: Stance on Islamists:** The final task aims to identify a given comment's

stance on Islamists. The possible classes are (1) Support, (2) Oppose, (3) Mixed Views, and, (4) None. Mixed views category is targeted towards comments that support one group of Islamists (ex. Muslim Brotherhood) while opposing other ones (ex. Salafis). We only found nine instances of this case in the whole dataset.

For each one of the five tasks, we measure how well different lexical and semantic features can successfully predict the correct class.

4.3 Approach

In this Section, we present our approach for building computational systems for perspective identification in Arabic. We describe how we preprocess the data along with the features and the machine learning models employed.

4.3.1 Preprocessing

We preprocess the text by separating punctuation and numbers from words and restricting word-lengthening–speech–effects to three repeated letters instead of an unpredictable number of repetitions, in order to maintain the signal that there is a speech effect while reducing the sparsity of the feature space.

We then run the sentence-level component of AIDA on each given comment in order to identify whether the comment is predominantly MSA or EDA. Finally, depending on whether the comment was tagged as being MSA or EDA, we use the corresponding version of MADAMIRA (Pasha et al., 2014) toolkit to tokenize the text. We apply the most detailed level of tokenization (D3) (Habash and Sadat, 2006) to the text. As described in Chapter 4, D3 segments off conjunction and pronominal clitics and particles–including the definite article *Al*.

4.3.2 Lexical Features

Lexical features have been shown to perform well on most text categorization tasks. Perspective Identification is no exception, especially since the lexical choice often conveys a person's leaning. Accordingly, we use standard n-gram features and experiment with two pre-processing schemes.

- **Surface:** In this setting, we only separate punctuation and numbers from text. (This setting preserves more context but yields a more sparse feature space)
- **Tokenized:** In this setting, we apply the same preprocessing scheme highlighted in Section 4.3.1 where we apply D3 tokenization to the text.

For each one of the two setups, we experiment with n-grams having a maximum length between 1 and 7 and use the optimal n to train our systems. For each training/test instance, we create a count vector of size V, where V is the number of n-grams after excluding punctuation and n-grams that occur in only one training instance.

4.3.3 Code-Switching (CS) Features

In this set of features, we aim to identify whether the level of dialectalness and MSA-EDA code-switching in a given comment can be an indicator of the comment's leaning.

Prior to tagging the text with the CS classes, and, in addition to, applying D3 tokenization to the text, we use SPLIT toolkit (Al-Badrashiny et al., 2016) to identify whether each word in the given comment is a number, punctuation, emoticon, URL, other non Arabic word or a sound (e.x. hahaha) and replace each of these with NUM, PUNC, EMOT, URL, LAT and SOUND keywords respectively. We then use two different systems to tag each word with its code-switching class. The first is our system–AIDA–and the second is the Language Independent Language Identification (LILI) system (Al-Badrashiny and

Diab, 2016). Unlike AIDA, LILI relies on character- and word-level n-grams to identify the class of each word in a given piece of text.

Using AIDA, طيب حد يعمل مبادرة قوية ! اللهم احفظ مصر واحقن دمائهم *"Tyb Hd yEml mbAdrp qwyp ! Allhm AHfZ mSr wAHqn dmA}hm"* meaning "Someone should make a strong initiative! God bless Egypt and save their blood" is tagged as follows:

"Tyb/EDA Hd/EDA yEml/EDA mbAdrp/EDA qwyp/EDA Allhm/NE AHfZ/MSA mSr/NE w/MSA AHqn/MSA dmA}/MSA hm/MSA"

CS-Tagged N-grams

In this first subset of features, we tag each D3 tokenized word with its token-level CS class (MSA, EDA, NE, UNK, Mixed, Ambig, or, Other) and create a count vector for each training/test instance in a similar fashion to our standard n-gram features. Similar to the lexical features, we exclude punctuation and n-grams that occur in only one training instance. This first set serves as a word sense disambiguation tool where we identify whether, for a given word, the intended meaning is the MSA or the EDA one. In the previous example, the most common meaning for the word–*"Tyb"*–is *"kind"* but when used as the first word in an EDA context it acts as a discourse marker meaning "OK". By tagging it as EDA, we disambiguate the word.

CS N-grams

The second subset of this feature-set uses AIDA's class only, without the word-ID. For this feature-set, depending on the maximum n-gram length (n) it can either just yield the distribution of the different word classes (MSA, EDA, NE, Ambig, Mixed, URL, LAT,

MAJ-BL	54.0		37.9	
	AIDA-Acc.(%)	LILI Acc.	AIDA-$F_{\beta=1}$ score	LILI-$F_{\beta=1}$ score
t = 100	55.0	54.0	55.7	55.7
t = 90	72.0	69.0	72.8	70.4
t = 80	**79.0**	**72.0**	**78.7**	**73.1**
t = 70	**79.0**	70.0	78.1	69.8
t = 60	77.0	71.0	74.4	68.3

Table 4.11: Results of using AIDA and LILI and different threshold (*t*) values to decide whether a given comment is purely MSA, purely EDA, or Code-Switched

	AIDA			LILI		
	% MSA	% EDA	% CS	% MSA	% EDA	% CS
Train	21.6	58.8	19.6	23.3	45.4	31.3
Dev.	19.8	60.6	19.6	23.6	46.6	29.8
Test	19.4	58.6	22.0	23.4	47.0	29.6
All	21.2	59.0	19.8	23.3	45.7	31.0

Table 4.12: Distribution of MSA EDA and Code-Switched (CS) Comments in the training, development and test sets of our dataset calculated using AIDA and LILI

EMOT or SOUND) in the document–if we only use unigrams–or can provide insights on how often the author code-switches–when using higher-order n-grams. The previously shown example becomes *"EDA EDA EDA EDA EDA NE MSA NE MSA MSA MSA MSA"*

Comment Class

For this feature, we aim to identify whether a given comment exhibits code-switching or not. While the sentence-level component of our Dialectal Arabic code-switch detection–AIDA–decides whether a given text is predominantly MSA or EDA, a main caveat to it is that it does not identify whether the sentence purely belongs to one class or if the author code-switches between both language variants. In other words, it identifies the matrix language but does not identify whether or not there is an embedded language.

The most straightforward way to identify whether or not the author code-switches is to check whether the sentence has at least one word from each one of the two language variants. However, the drawback to this approach is that one word's misclassification can alter the sentence's decision. Instead, we rely on a thresholding approach to identify whether the sentence is (1) purely MSA, (2) purely EDA or (3) Code-switched. We calculate the following two percentages:

$$MSA_{\,percent} = \frac{Count\,(MSA)}{Count\,t(MSA) + Count\,(EDA)} \tag{4.1}$$

$$EDA_{\,percent} = \frac{Count\,(EDA)}{Count\,(MSA) + Count\,(EDA)} \tag{4.2}$$

where:

Count (MSA): the number of MSA words in the given comment,

Count (EDA): the number of EDA words in the given comment.

We then determine the class of the given comment as follows:

$$Sentence_{\,class} = \begin{cases} MSA & ;\ MSA_{\,percent} >= t \\ EDA & ;\ EDA_{\,percent} >= t \\ CS & ;\ otherwise \end{cases} \tag{4.3}$$

where *t*: is an empirical threshold.

In order to determine the optimal threshold, we select a sample of 100 comments and have it annotated by a trained linguist. We then experiment with different thresholds and measure the performance on this gold set. As Table 4.11 shows, setting the threshold (*t*) to 100% performs poorly while setting it to 80% yields best results. Moreover, we find

that AIDA outperforms LILI. We therefore set t to 80 and use equation 4.3 to decide upon the class of the comment. Table 4.12 shows the calculated MSA-EDA-CS distribution of the different subsets of the dataset. In the annotated sample, the annotator indicated that 23%, 54%, and 23% of the comments are MSA, EDA and Code-Switched respectively, which aligns more with the percentages estimated by AIDA than with those estimated by LILI.

4.3.4 Sentiment Features

As discussed in previous chapters, one's opinion towards different topics can often serve as an indicator for his/her perspective and identifying such sentiment can help us in uncovering the leaning from which a given comment was written. However, due to the lack of required resources, we do not use targeted-sentiment but rather rely on the comment-level sentiment. Using a publicly available generic sentiment analysis system (Badaro et al., 2014, 2015), we identify the overall sentiment expressed by a given comment and use it as a feature for our classifiers. Given a comment, the system uses an underlying sentiment lexicon to retrieve the (1) positive, (2) negative, and, (3) neutral scores for each word in the comment. The sum of each of these three scores is then used to identify the overall sentiment of this given comment. The main caveat of this system is that the underlying lexicon targets MSA only while our dataset is mostly EDA.

4.3.5 Weighted Matrix Factorization Features (WMF)

The next set of features relies on mapping text from the high-dimensional n-gram space to a low-dimensional topic space. We use the Weighted Textual Matrix Factorization (WMF) system (Guo and Diab, 2012) . WMF is a topic modeling approach that–unlike LDA(Blei, Ng, and Jordan, 2003)–is tailored for short texts. In addition to modeling observed words, WMF models missing ones, namely explicitly modeling what a given comment is not

about. Missing words are defined as the whole vocabulary of the training data minus the ones observed in the given document. However, since observed words are more informative than missing ones, WMF assigns missing words a very small weight (ex. 0.01) while setting the weight for observed ones to 1. This allows for modeling missing words at the right level of granularity–without being dominated by them. The main advantage of using WMF is that it leads to a denser vector space model, hence solving one of the major problems of bag-of-words models, namely, the sparse matrix representation. Moreover, since people sharing the same perspective often discuss similar topics, they tend to have close semantic textual similarity (STS). By using WMF, we capture this phenomenon, since similar comments yield high vector similarity. We use the default settings for the distributable version of WMF (Guo and Diab, 2012), which sets the number of topics (K) to 150 and sets the weight of the missing words to 0.01. We collect our training data from online discussion fora discussing Egyptian politics and preprocess it by applying the same cleaning process explained earlier in Section 4.3.1. The final data that we use to build our model has a total of ~17 million D3 preprocessed tokens corresponding to ~177 thousand types.

4.3.6 Machine Learning Model

For all experiments we use a Logistic-Regression classifier [2] within Scikit-Learn toolkit (Pedregosa et al., 2011).[3] Logistic Regression is a regression model for classification problems where a logistic function is used to model the probabilities of all possible class labels of a data instance. We use the One-vs-Rest implementation [4] in Scikit-Learn and apply L2 regularization while setting the inverse of regularization strength (C) to 10 for all datasets. One-vs-Rest trains a single binary classifier per class, treating the instances of that class as

[2]We experimented with SVM, Random Forest and Logistic Regression and achieved the best results using Logistic Regression.

[3]`http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`

[4]We experimented with both multinomial and One-vs-Rest Logistic Regression and the latter was much faster then the former while yielding comparable results to the former.

being positive samples and all other instances as being negative. We define the following feature categories:

- **N-grams:** This category includes both "Surface" and "Tokenized" n-grams;
- **Pragmatics:** This category includes code-switching and sentiment features;
- **WMF:** This category includes WMF features.

We first experiment with features in each feature category before combining feature sets from the three feature groups. For each feature set within the same feature group, we use the combined feature space to train a single classifier on the combined feature set. Figure 4.2 shows the feature combination approach for features within the same category.
After training a classifier for each category, we combine the classifiers for the different categories using a classifier ensemble approach (Figure 4.3) where the probabilities from the different classifiers are aggregated as follows:

$$score(y_c) = \sum_{k=1}^{K} prob_k(y_c) \tag{4.4}$$

$$y = max_{c \in C} \left(score(y_c)\right) \tag{4.5}$$

where:

C: is the number of classes,

K: is the number of classifiers,

y: is the predicted class label.

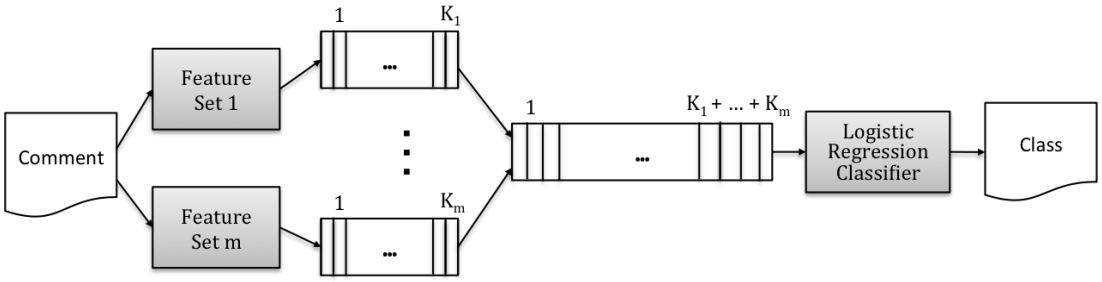

Figure 4.2: Combined-Features approach that uses all feature sets to train a single classifier to identify the class of a given post

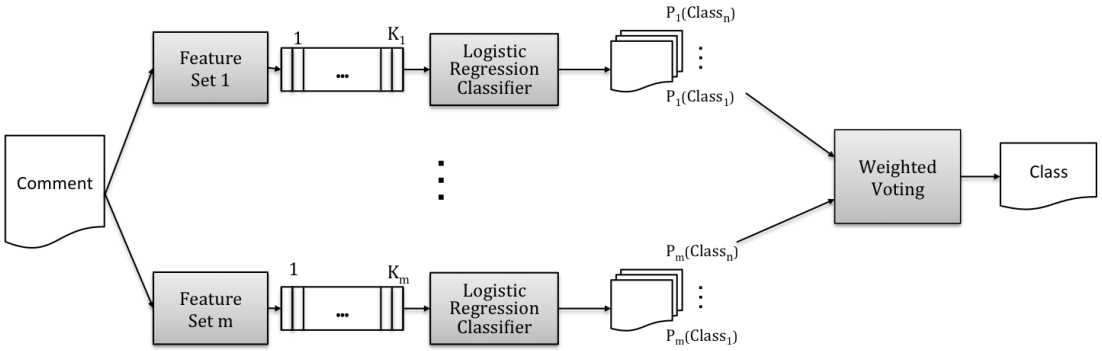

Figure 4.3: Classifier Ensemble Approach that uses weighted voting to combine the decisions from different classifiers. Each classifier yields a probability distribution over the possible class labels. The probability distributions from different classifiers are then summed in order to identify the final class label.

4.4 Experiments and Results

4.4.1 Evaluation Metric

We use weighted average $F_{\beta=1}$ score to evaluate the performance on each task. $F_{\beta=1}$ score for a given class represents the harmonic mean of precision and recall for that class.

$$F_{\beta=1}\ score(c) = 2 * \frac{Precision(c) * Recall(c)}{Precision(c) + Recall(c)} \tag{4.6}$$

$$F_{\beta=1}\ score\ (Task_i) = \sum_{c=1}^{C} \frac{n_c}{n} * F_{\beta=1}\ score(c) \tag{4.7}$$

Additionally, we calculate the average performance across all tasks in order to identify the best overall setup.

$$Average F_{\beta=1}\ score\ = \sum_{i=1}^{5} \frac{F_{\beta=1}\ score\ (Task_i)}{5} \tag{4.8}$$

where:

C: is the number of classes,

n_c: is the total number of test instances whose true class label is *c*,

n: is the total number of test instances,

K: is the number of classifiers,

y: is the predicted class label.

4.4.2 Baselines

We compare our approach to three baselines:

- **Majority Baseline *(MAJ-BL)***: which assigns all comments to the most frequent class-label;
- **Random Baseline *(RAND-BL)***: which randomly chooses the class-label;
- **Surface N-grams Baseline *(NGRM-BL)***: a strong baseline that uses standard n-gram features. In this baseline, we use words in the surface (non D3 tokenized) space. We choose the maximum n-gram length (n) by tuning.

4.4.3 Results

For all datasets, we tune on a development set and apply the setup that yields best results on the development set to the held-out test set. We begin by tuning for n-gram length and evaluating the two code-switch detection systems before exploring different configurations of features within each feature group. We then combine the best setups from the different feature groups in a classifier ensemble. For the held-out test set experiments, we train on both training and development sets. [5]

Tuning N-grams

We tune for the maximum n-gram by experimenting with a maximum n between 1 and 7 for all of our five tasks. Tables 4.13 and 4.14 show the results of using different values of (n) for both Surface and Tokenized setups on the development dataset. For surface-level n-grams, unigrams perform better than all higher order n-grams. In Arabic, several clitics and conjunctions can get attached to the word which results in unigrams capturing more context and higher order n-grams to be more sparse than the tokenized ones; this explains the lower performance for higher values of n. On the other hand, we find that for D3 tokenized n-grams, unigrams perform worst since they completely lose the context and word order. The average performance across the five tasks is almost the same when setting n to 3, 5 or 7. We therefore set n to 1 for all Surface (non D3 tokenized) experiments and n to 3 for D3 tokenized ones–since Arabic has a prefix+stem+suffix morphology for the most part.

Tuning CS Features

Since we perform the CS tagging on the D3 tokenized text, we set the maximum n-gram length to 3–similar to the non CS-tagged D3 tokenized text. We evaluate the performance

[5]We do not include statistical significance results due to the small size of the dataset.

Surface	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
n=1	43.3	**61.9**	**58.5**	**67.7**	**71**	**60.5**
n=3	**43.9**	60.8	**58.5**	66.5	69.9	59.9
n=5	42.9	60.5	**58.5**	66.2	69.8	59.6
n=7	43	60.5	58.3	66.4	69.6	59.6

Table 4.13: Surface N-grams: Tuning for maximum n

Tokenized	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
n=1	42.5	60.7	58.2	65.5	66.6	58.7
n=3	45.1	63.1	60.5	**67.8**	**69.7**	61.2
n=5	45	**63.3**	61	67.6	69.4	**61.3**
n=7	**45.2**	63.1	61.5	**67.8**	68.8	**61.3**

Table 4.14: D3 Tokenized N-grams: Tuning for maximum n

	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
AIDA	**44.7**	59.7	**58.6**	**68.0**	**70.5**	**60.3**
LILI	44.3	**59.9**	57.6	67.3	69.2	59.7

Table 4.15: Comparing the performance of the two Code-Switching Systems

of both CS-tagging systems–AIDA and LILI on all of the five tasks. As Table 4.15 shows, AIDA outperforms LILI, which is consistent with our previous evaluation of both systems in Section 4.3.3. Hence, we only use AIDA in our next experiments.

Single Feature Group Results

In this set of experiments, we evaluate the performance of features within each feature group separately. We explore how the level of preprocessing impacts performance by measuring the impact of adding D3-Tokenized n-grams to surface n-grams. Moreover, we evaluate the performance of AIDA-based code-switching features separately and when combined with sentiment features.

	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
MAJ-BL	14.0	26.0	23.0	53.0	45.0	32.2
RAND-BL	13.0	36.0	36.0	37.0	38.0	32.0
NGRM-BL/NgrmSur	43.3	61.9	58.5	67.7	71.0	60.5
NgrmSur+Event	48.4	68.5	66.5	73.7	**76.9**	66.8
NgrmSur+NgrmTok	43.9	59.3	59.7	69.1	70.5	60.5
NgrmSur+NgrmTok+Event	51.3	**70.4**	**67.6**	**74.8**	75.3	**67.9**
AIDA	44.7	59.7	58.6	68	70.5	60.3
AIDA+Event	**51.7**	69	66.4	72.8	75.8	67.1
Sentiment+AIDA+Event	51.5	69.3	67	72.4	75.8	*67.2*
WMF	42.4	54.9	54.6	61.6	63.4	55.4
WMF+Event	49.4	68.2	63.9	71.3	76.2	*65.8*

Table 4.16: Results of using different combinations in each feature group on the development set

In addition to these features, we explore whether adding an explicit feature indicating what event the comment is discussing yields better results. The events are the list of events used to curate our data. We selected comments posted within one week of the start of ten major political events. Hence, for each given comment, we knew which event was at the center of public attention when a comment was posted. Accordingly, even if the event is not explicitly mentioned in the comment, we can assume that the comment discusses this event with relatively high confidence. Moreover, annotators indicated that in the manual annotation, identifying the perspective of a comment was much easier when event information was available. Our results (Table 4.16) further confirms the annotators' feedback. We find, that adding event information results in a significant boost–6.3% when using surface n-grams and 7.4% when combining surface and tokenized n-grams–in the performance, on all tasks. Moreover, combining both surface and tokenized n-grams outperforms all other features, strongly suggesting that lexical features are crucial for the

task. For WMF features, they perform relatively well and beat the majority and random baselines but fail to beat the n-gram baseline.

Finally for pragmatic features, sentiment and code-switching, we find that adding sentiment features to code-switching features yields almost the same results as using code-switching features alone. This result–especially on Tasks 2-5–is surprising given the similarity of stance identification to sentiment analysis. This may be explained by two factors: the quality of the sentiment analysis system as it relies on MSA lexicons without taking negation nor context into consideration, and the manner in which we use sentiment; namely, the fact that we only utilize comment level sentiment, not targeted sentiment. For example, if the comment conveys a positive sentiment, we do not identify whether this sentiment is targeted towards Islamists, Military, OGR or even some other non-ideological entity, such as the results of a soccer game. We only use the signal that the comment is positive as a feature in our classifier.

Classifier Ensemble Results

Next, we combine the best setups–from each feature group–in a classifier ensemble. We define the best setup within a feature-group as one that outperforms all other setups within the same group on at least one task or on the average performance across the five tasks. For all feature-sets, we add the event-information. We compare the results to the baselines and the best set-ups from the previous set of experiments.

Combining code-switching features with lexical features, improves the performance of lexical features by 1.6% when using Surface n-grams only and 1.3% when using both Surface and Tokenized n-grams. Adding sentiment to code-switching features either slightly improves or slightly hurts the performance depending on the task, while adding WMF features slightly improves the performance. WMF improves the performance by

	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
MAJ-BL	14.0	26.0	23.0	53.0	45.0	32.2
RAND-BL	13.0	36.0	36.0	37.0	38.0	32.0
NGRM-BL	43.3	61.9	58.5	67.7	71.0	60.5
NgrmSur+Event	48.4	68.5	66.5	73.7	76.9	66.8
NgrmSur+NgrmTok+Event	51.3	70.4	67.6	74.8	75.3	67.9
Ens.(NgrmSur,AIDA)	51.3	71.2	68.2	74	**77.1**	68.4
Ens.(NgrmSur, AIDA, WMF)	50.9	71.5	69.3	**75.6**	76.9	68.8
Ens.(NgrmSur, AIDA+Sentiment)	50.8	70.9	68.7	74.1	76.2	68.1
Ens.(NgrmSur, AIDA+Sentiment, WMF)	51.5	71.3	69	74.7	76.6	68.6
Ens.(NgrmSur+NgrmTok,AIDA)	53.0	71.3	69.9	74.8	76.9	69.2
Ens.(NgrmSur+NgrmTok, AIDA, WMF)	**53.7**	72.1	70.2	75.4	77	**69.7**
Ens.(NgrmSur+NgrmTok, AIDA+Sentiment)	52.7	71.7	70.1	74.8	76.8	69.2
Ens.(NgrmSur+NgrmTok, AIDA+Sentiment, WMF)	**53.7**	**72.3**	**70.4**	75.2	76.3	69.6

Table 4.17: Results of combining features from different feature groups using a classifier ensemble on the development set against the baselines and against the best setups from each feature group. In addition to using the listed features, each classifier in the classifier ensemble uses the event information as a feature

0.4% when using Surface n-grams and code-switching features only, 0.5% when adding sentiment features, 0.5% and 0.4% when using surface and tokenized n-grams with and without sentiment respectively. Overall, using all of lexical, code-switching and WMF–with or without sentiment–features performs best.

Held-Out Test Set Results

For the held-out test, we experiment with the setup that resulted in the best average performance on the development task–using a classifier ensemble that utilizes lexical (both surface and tokenized), code-switching, WMF and event features–and compare against the

	Priority	Jan. 25^{th}	OGR	Military	Islamists	Avg
MAJ-BL	13.0	24.0	23.0	53.0	47.0	32.0
RAND-BL	14.0	38.0	36.0	37.0	37.0	32.4
NGRM-BL	42.3	60.2	57.6	70.8	69.0	60.0
NgrmSur+Event	51.5	66.6	64.1	74.9	75.9	66.6
NgrmTok+NgrmSur+Event	55	68.5	65.9	78	**77.9**	69.1
Ens.(NgrmSur+NgrmTok, AIDA, WMF)	**56.9**	71.4	**68.8**	77.9	77.4	**70.5**

Table 4.18: Results of using the best development setup on the held-out test set. In addition to using the listed features, the classifier ensemble uses the event information as a feature

baselines and the use of only lexical features. Table 4.18 shows the results. Similar to the development set, adding code-switching and WMF features improves over using only lexical features.

4.4.4 Error Analysis

We look at the confusion matrices and examples of both correctly classified and misclassified instances of the held-out test set. Tables 4.19 to 4.23 show the confusion matrices for all tasks. For the first task (identifying the priority of the comment), the classification errors are distributed across all classes. For tasks 2 to 5, identifying the stance on different political entities, the highest confusability is between each of the *None* class and the other two classes. This is not surprising given that it is the most dominant class in the data and also since each of the two classes–*Support* and *Oppose*–are closer to the *None* class than to each other. For identifying the stance towards Islamists, the held-out test set does not have any *Mixed views* instances, and the training data has only nine instances; hence, other classes are never confused as belonging to this class.

Tables 4.24 and 4.25 show some of the misclassified and correctly classified in-

	Stability	Jan. 25^{th}	Mubarak	Military	Islamists	Opp.Islamists	Ambig.	None
Stability	0.4	0.6	0.4	0.4	0.2	0.0	0.0	0.4
Jan. 25^{th}	0.2	22.6	3.4	0.2	0.4	0.8	0.2	1.2
Mubarak	0.6	2.6	8.8	0.2	0.4	2.0	0.0	0.6
Military	0.0	0.4	0.4	3.0	0.4	1.6	0.0	0.6
Islamists	0.0	0.6	0.2	0.2	7.0	4.6	0.2	1.4
Opp.Islamists	0.0	0.8	0.0	0.0	3.2	12.2	0.0	1.4
Ambig.	0.0	0.4	0.2	0.0	0.4	0.2	0.2	0.6
None	0.0	3.4	1.2	0.2	3.0	1.6	0.0	3.8

Table 4.19: Confusion Matrix for Task 1 (Identifying the priority of the comment).

	Support	Oppose	None
Support	24.4	5.2	8.4
Oppose	3.6	10.8	6.2
None	3.0	1.6	36.8

Table 4.20: Confusion Matrix for Task 2 (Identifying the stance of the comment on January 25^{th} Revolution).

stances. The examples highlight different reasons behind misclassfications. Example 1 was misclassified as supporting January 25^{th} Revolution when it actually opposes it. The stance was expressed only indirectly by mentioning the fact that Egyptians endured 30 years–in reference to Mubarak's presidency–and can wait for two more years in order not to negatively impact Egypt's economy. The author does not directly mention Mubarak or January 25^{th} Revolution, hence identifying the stance even for a human annotator is not straightforward since the reference is vague and requires world knowledge. The third example shows an instance of an annotation error. The annotator wrongfully identified the comment as supporting Islamists when the comment actually opposes them. The author expresses concern over the lack of security measures that resulted in the death of some soldiers. During the period this comment was posted–post Rabia camp dismantling–the

	Support	Oppose	None
Support	10.6	4.4	5.2
Oppose	5.4	24.4	10.0
None	2.0	3.6	34.4

Table 4.21: Confusion Matrix for Task 3 (Identifying the stance of the comment on Mubarak's Regime and the OGR).

clashes between the Army and the Islamists caused pro-Islamists to be only concerned about the death of their own supporters while pro-Military leaders were concerned about the death of the soldiers. Each side almost refused to acknowledge the deaths on the opposing side. The comment clearly expresses concern about the death of the soldiers, which indicates that the author belongs to the anti-Islamists camp. The fourth example, on the other hand, quotes a late Islamic Caliph. The quote is quite ambiguous, implying that being neutral is not the right choice at the given time. The system classified this comment as supporting Islamists probably because of the use of Standard Arabic and the religious reference. For the last example–example 5–the comment is sarcastic, and since we do not explicitly model sarcasm in our current approach, the system was not able to classify it correctly.

For the correctly classified examples, the lexical choice in the second and third examples are common among Islamist Supporters (Military Leaders opposers), hence they were classified correctly. Words such as "legitimacy", "coup" and "lynching" were only used by Islamist supporters during that period, whereas for event 4, the use of terrorists is a clear indication of anti-Islamists stance.

	Support	Oppose	None
Support	11.6	2.6	4.8
Oppose	2.6	5.6	6.4
None	3.4	1.0	62.0

Table 4.22: Confusion Matrix for Task 4 (Identifying the stance of the comment on Military Leaders).

	Support	Oppose	Mixed	None
Support	5.6	6.2	0.0	4.0
Oppose	3.4	16.6	0.0	2.8
Mixed	0.0	0.0	0.0	0.0
None	2.0	3.4	0.0	56.0

Table 4.23: Confusion Matrix for Task 5 (Identifying the stance of the comment on Islamists).

4.5 Summary

In this chapter, we addressed the problem of perspective identification in Egyptian social media. We presented our proposed taxonomy of the major community perspectives in Egypt and described how we used this taxonomy and an iterative process to collect large-scale linguistic annotations. We then explored the use of lexical and semantic features in building supervised systems that can identify several aspects of a given comment's underlying perspective. We found that adding the event the comment is discussing as a feature to our classifier helps us achieve better performance. This aligns with the annotators' feedback that adding the event information helped them in understanding the context and judging the leaning of each comment more easily. We also found that adding more sophisticated lexical features–by tokenizing the input to split off clitics and conjunctions–improves the performance for all tasks. We attribute this to how word segmentation increases the coverage and decreases the sparsity of the vocabulary. This is

	Comment	Stance on	Event	Gold	Pred.
1	لازم نزق سفينة الاقتصاد المصرية للامام عشان نرتاح بعد سنتين و الصبرنا ٣٠ سنة قادر يصبرنا سنتين و بعدين نعيش عيشة كريمة انشاء الله We have to support the Egyptian economy in order to live comfortably. We were patient for 30 years so we can be patient for two more years in order to live a better life if God wills.	Jan. 25^{th}	3	opp.	sup.
2	اة و الله عايزين نحدد يوم ننزل كلنا فية نعمل عمل خيرى انا بقول نتبرع بالدم الاول لان فية س محتاجة الدم دة Yeah seriously, we should specify a day where we all volunteer for a charity. I suggest donating blood first because people need it.	Mubarak	2	opp.	sup.
3	انا مضايقة منكم دلوقتي بسبب العساكر اللي ماتو دول ، دة تقصير منكم لانه مفيش تأمين كافي وانتو عارفين أنكم مستهدفين I am upset with you now because of the soldiers who died. You know you are targeted and yet you do not secure them enough.	Islamists	10	sup.	opp.
4	المحايد هو شخص لم ينصر الباطل ولكن من المؤكد أنه خذل الحق . الإمام علي بن أبي طالب “A neutral person is someone who did not support the immoral but who let the truthful down” Imam Ali Ibn Aby Taleb	Islamists	9	none	sup.
5	الريس هو المتهم الوحيد في حريق روما بتاع زمان ده .... و سمعني سلام مبارك يعني الحكمه The president is the one behind the fire in Rome from long ago. Chant with me Mubarak means wisdom.	mubarak	4	sup.	opp.

Table 4.24: Examples of Misclassified Instances

	Comment	Stance on	Event	Gold / Pred.
1	فى حد لسه عنده شك ان الثورة المضاده نجحت يوم ٣٠ - ٦ . ولا نسيتو يا ثوار ان كان فى ثورة مضاده ؟ Does anyone still doubt that the counter-revolution succeeded on June 30^{th} or did you revolutionists forget that there was a counter revolution?	Jan. 25^{th}	9	sup.
2	من برر السحل والقتل سيسحل ويقتل علي يد من برر له حتي ولو بعد حين The ones who justified killing and lynching will be killed and lynched even if after a while.	Islamists	9	sup.
3	انا مش من الاخوان لكن مع الشرعيه ... مبروك عليكم الانقلاب وضياع الحريه وعودة مبارك فى القريب العاجل I am not with the Brotherhood but rather with legitimacy. Congratulations on the coup and the loss of freedom and the return of Mubarak in the near future.	Military	8	opp.
4	انتوا تانى يالى انتخبتوا ارهابيين ومش عاجبكوا انهم بيموتوا المفروض نسبهم عايشين لحد مايخلصوا عليكم ساعتها مش هنسمع الكلام دا You who voted for terrorists and that denounce that they are dying. We should have let them kill you so that we do not listen to what you are saying.	Islamists	9	opp.
5	انا مش اسف ياريس علشان هزيت كرمتي ولا المجلس العسكري علشان برده عايز يهز كرمتي انا نزله مدان التحرير وهقول للفساد لا I am not sorry Mr. President because you humiliated me and I am also not sorry Military Council because you want to humiliate me. I am going to Tahrir square and I will say no to corruption.	Jan. 25^{th}	1	sup.

Table 4.25: Examples of Correctly Classified Instances

especially important in Arabic due to the fact that multiple clitics and conjunctions can be attached to the word causing a more sparse feature space. Counter to expectations, sentiment features yielded a very slight improvement to our results. This can be attributed to one of two factors: the quality of the sentiment analysis system itself and the fact that we only use comment-level and not targeted sentiment. The sentiment analysis system that we used only relies on an underlying MSA sentiment lexicon and does not handle negations or exploit syntactic information; hence, it ignores the contextual information. Moreover, due to the lack of resources for Arabic, we do not perform targeted sentiment but rather rely on comment-level sentiment. We expect targeted-sentiment to yield better results.

Code-switching features improve the performance on both the development and test sets. While this suggests that code-switching indeed serves as a signal for identifying a person's perspective, it would be interesting to study whether or not this signal is stronger in other datasets; specifically those that exhibit more code-switching.

Chapter 5

Perspective Identification in English

As previously discussed, there are various elements governing the belief system of people that affect their stance on different ideological topics.

In this chapter, we discuss our work on automatic identification of perspective in English. We begin by describing the task in Section 5.1, followed by the datasets we used in Section 5.2. We then describe our proposed approach, along with the experiments and results in Sections 5.3 and 5.4. Finally, we summarize the chapter in Section 5.5.

5.1 Task

As discussed earlier, our goal is to build computational systems that can uncover different elements governing a person's perspective from a given text discussing issues related to one's ideological perspective by identifying the person's stance towards these issues. In American politics, one's political party affiliation, as well as whether one self-identifies as being a conservative or a liberal, are often used as indicators for a person's leanings and perspectives. Such perspective is often expressed in one's stance on polarizing issues, such as legalization of abortion, feminist ideologies, climate change, etc. In this chapter, we address two very related tasks. The first task occurs when identifying a person's stance on a specific ideological topic expressed in a given online post or Tweet. The second task occurs when trying to identify people's candidate preference for the 2012 presidential race based on their answers to a set of open-ended, essay-style questions. While both the first and second tasks are related, the second is not only more abstract than the first but

dependent on it because a person's stance on different economic, political and social issues normally determines the party affiliation and his/her choice of presidential candidate.

5.2 Annotation and Datasets

We use one standard dataset "SemEval 2016 Task 6: Detecting Stance in Tweets" for the first task. Additionally, we create a new dataset "American National Election Studies dataset" to address the second task.

5.2.1 SemEval 2016 Stance Dataset

SemEval Task 6 "Detecting Stance in Tweets" (Mohammad et al., 2016) aims at evaluating how well an automated system can identify the stance of a Twitter user on several contentious targets. The first sub-task "Supervised Framework" of the task focuses on five targets: "Atheism", "Climate Change is a Real Concern", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". The training data has a total of 2,814 tweets and the test data has 1,249 tweets. We show sample tweets from each domain in Table 5.1. We use the shared task's training data to tune the system. Tables 5.2 and 5.3 show the statistics and class distribution of the training, development, and held-out test set for each one of the five targets. Since this dataset is based on tweets, the average post/tweet length is very short.

5.2.2 American National Election Studies (ANES) Dataset

We create this dataset–together with Chris Callison-Burch–by drawing a set of questions from the American National Election Studies (ANES) survey questions.[1] ANES conducts various surveys in order to provide better explanations and analysis of the outcomes of U.S. presidential elections. While the officially administered ANES survey contains both

[1] `http://www.electionstudies.org/studypages/2010_2012EGSS/2010_2012EGSS`

Domain	Stance	Tweet
Abortion	Favor	Yea, let's make a woman suffer 18 years and raise a child with an unfit parent who hates it because of "morality"
	Against	I was an accidental baby and as far as I know, my parents were happy to have me. #thanksmom
	None	The teen pregnancy rate has declined 51 percent, and the teen birth rate is down 57 percent.
Atheism	Favor	No, I'm not calling myself "agnostic" because my atheism scares you and u dont know what words mean.
	Against	Every life is a profession of faith, and exercises an inevitable and silent influence. ~Henri Frederic Amiel
	None	Should we take the #DNA at birth of every human being on Earth? #science #economics
Climate	Favor	Sea Level Rise above 6 meters - what does that mean? It means 20 ft above current heights.
	Against	We are not "killing the Earth". The Earth has been through worse and will be fine after all humans suffocate, drown or starve
	None	The Weather app keeps taunting us with rain. #PNW #drought
Clinton	Favor	Based on the long lines, I thought it was free burrito day at Pancheros but it was actually Hillary! #ReadyForHillary
	Against	I think that everything she says is a lie. I mean EVERYTHING. I don't even think her name is Hillary.
	None	Dad while watching the news: Politics is just show business for ugly people #HowAboutNo
Feminism	Favor	Dear parents, please don't tell your boys "not to be a girl" when they cry. Girls rock.
	Against	I like girls. They just need to know there place.
	None	Why would they have girl dragons in dragonvale?

Table 5.1: Sample tweets from the different domains of SemEval 2016 Stance dataset

		Tweets	Tokens	Types	Tokens/Tweet
Abortion	Train	544	10,385	3,033	19
	Dev.	59	1,072	589	18
	Test	280	5,521	1,944	20
Atheism	Train	461	9,325	2,845	20
	Dev.	51	1,010	537	20
	Test	220	4,580	1,707	21
Climate Change	Train	356	6,384	2,563	18
	Dev.	39	658	428	17
	Test	169	3,247	1,467	19
Hillary Clinton	Train	574	10,283	3,248	18
	Dev.	64	1,206	671	19
	Test	295	5,830	2,068	20
Feminist Movement	Train	597	11,458	3,429	19
	Dev.	67	1,258	645	19
	Test	285	5,719	2,026	20

Table 5.2: Statistics of SemEval 2016 Stance dataset

constrained multiple choice questions and open-ended (free form essay-style) questions, the answers to the open-ended questions, which are more interesting from an NLP perspective, are not made publicly available so as to protect respondents' privacy.

We run an Amazon Mechanical Turk survey where we ask Amazon Mechanical Turk annotators (aka Turkers) to answer a large set of constrained and open-ended questions drawn from ANES. The constrained questions may be considered a form of self-labeling annotation that indicate the respondent/Turker's background or perspective on specific issues. All Turkers participating in the experiment were required to be from the United States. Moreover, we added eight quality-control questions with a correct (and obvious) answer–such as asking them to write the number "thirty three" in digits and

	%	Favor	Against	None
	Train	17.5	55.3	27.2
Abortion	Dev.	16.9	55.9	27.1
	Test	16.4	67.5	16.1
	%	Favor	Against	None
	Train	18.0	59.4	22.6
Atheism	Dev.	17.6	58.8	23.5
	Test	14.5	72.7	12.7
	%	Favor	Against	None
	Train	53.7	3.9	42.4
Climate Change	Dev.	53.8	2.6	43.6
	Test	72.8	6.5	20.7
	%	Favor	Against	None
	Train	17.4	56.6	26.0
Hillary Clinton	Dev.	17.2	56.2	26.6
	Test	15.3	58.3	26.4
	%	Favor	Against	None
	Train	31.7	49.4	18.9
Feminist Movement	Dev.	31.3	49.3	19.4
	Test	20.4	64.2	15.4

Table 5.3: Class Distribution across the five domains of SemEval 2016 Stance dataset

asking them to select the sound a specific animal makes–in order to identify spam Turkers. All submissions that yielded more than two of these questions wrong were automatically rejected.

The first set of questions that required constrained answers–such as multiple choice or binary responses like *true* or *false*–can be placed into the following categories:

- Background Questions: A person's age, gender, educational level, income, marital-status, social-status, how often he/she follows the news, and what news sources he/she follows, among others;
- Opinion of Political Parties: Democratic and Republican parties and their respective public figure representatives;
- Opinion on Major Economic and Political Problems Facing the U.S.;
- Ideology Questions: Importance of religion, political party affiliation, presidential candidate choice, etc.;
- Opinion on Contentious Issues: White, Black, Asian and Hispanic Americans, same-sex marriage, gun control, universal healthcare, etc.

The second set of questions ask about a person's opinion on certain ideological topics. The responses are not constrained in any manner.

Since our main objective is to study whether a person's perspective can be automatically identified using NLP techniques applied to the written text, we choose to predict the answer to one of the constrained ideological questions, "Presidential Candidate Choice" (PCC), based on the answers to the following open-ended questions:

- **Q1**: Is there something that would make you vote for a Democratic presidential candidate?
- **Q2**: Is there something that would make you vote against a Democratic presidential candidate?

Q1	I approve of Obama's and the Democrats' position on abortion and gay marriage and their tendency to favor programs that help the poor and working class. They seem more compassionate and more socially progressive.
Q2	Neither Obama nor the Democrats seems able to get a hold on spending, the deficit or help the economy and unemployment. They seem to spend too much time criticizing their opponents rather than work toward viable solutions and seem to distort facts against the other party more.
Q3	I think Mitt Romney and the republicans in general would do a better job at lowering the deficit and stimulating the economy and reducing unemployment. I also agree with their position of less government involvement in some areas.
Q4	I dislike Mitt Romney's plans to eliminate funding for Planned Parenthood and the republicans stand on social issues such as abortion and gay rights, especially gay marriage. I feel Republicans have been taken over by the religious right and are socially regressive.

Table 5.4: Sample answers provided by one Turker to the first four essay questions in ANES dataset

- **Q3**: Is there something that would make you vote for a Republican presidential candidate?
- **Q4**: Is there something that would make you vote against a Republican presidential candidate?
- **Q5**: If you said there is something you like about the Democratic Party: What is that?
- **Q6**: If you said there is something you dislike about the Democratic Party: What is that?
- **Q7**: If you said there is something you like about the Republican Party: What is that?
- **Q8**: If you said there is something you dislike about the Republican Party: What is that?
- **Q9**: What has been the most important issue to you personally in this election?
- **Q10**: What has been the second most important issue to you personally in this election?
- **Q11**: What do you think is the most important political problem facing the United States today?
- **Q12**: What do you think is the second most important political problem facing the United States today?
- **Q13**: What do you think the terrorists were trying to accomplish by September 11^{th} attacks?

	Posts	Tokens	Types	Tokens/Post
Train	869	348,898	20,590	401
Dev.	96	50,135	7,088	522
Test	108	56,077	7,416	519

Table 5.5: Statistics of ANES dataset

%	Obama	Romney	Neither
Train	62.8	25.3	11.9
Dev.	63.5	25.0	11.5
Test	67.6	18.5	13.9

Table 5.6: Class Distribution of Presidential Candidate Choice (PCC) in ANES dataset

Table 5.4 shows the answers provided by a Turker to the first four of these questions.

In order to simulate user-generated content where people are not providing answers to a predefined set of questions but instead are discussing current events or topics, we decide to combine the answers to all of these questions in one document per Turker and to use this combined, resulting document to derive features (as opposed to deriving features from the answer to each question separately). The result of using this method of creating posts is a long average post length. In order to reduce ambiguity, we perform a quasi co-reference resolution step on pronouns. Prior to combining the answers to all 13 questions, we perform a "pronoun-rewriting" step where we replace the sentence's initial pronouns with the topic that the question is about. For example, for Q3, "Is there something that would make you vote for a Republican presidential candidate?" the answer provided is "They are against voting rights for illegal immigrants. They want to balance the budget and find a way to slowly reduce the national debt." In this case, we replace "they" with "Republicans". We split the data into 90% training and development and 10% held-out test set. We further split the first set into 90% training and 10% development. Tables 5.5 and 5.6 show the statistics and class distribution in the training, development and test sets. The majority of the Turkers chose Obama as their chosen 2012 presidential candidate, indicating a bias in

our data towards Democrats. [2]

5.3 Approach

Our goal is to determine how well lexical and semantic features can help in identifying a person's ideological perspective as determined by his/her answer to the 2012 Presidential Candidate Choice (PCC) question in the "ANES" dataset and his/her stance towards the ideological topics discussed in SemEval Stance dataset. In addition to using standard n-grams, we explore the use of word sense disambiguation, targeted-sentiment, as well as latent and frame semantics in identifying the leaning of a given post/tweet.

5.3.1 Lexical Features

Lexical features have been shown to perform well on most text categorization tasks. Perspective Identification is no exception, especially since the lexical choice often conveys a person's leaning. For example, when discussing abortion, supporters of the cause will often use words that convey their stance, such as "choice" and "women rights" while those who oppose the cause will focus on "life", "killing", and "baby". Similarly, when discussing gun rights, a supporter of the cause will highlight "self-defense" while an opponent might focus on "death". For lexical features, we use standard n-grams. We apply basic preprocessing to the text by removing all punctuation and converting all words to lowercase. Converting the text to lowercase is intended to reduce the sparseness of the data, while excluding punctuation is meant to avoid over-fitting the training data. For each n-gram, we create a binary feature indicating the presence/absence of this n-gram in a given post. We experiment with n-grams having a maximum length (n) between 1 and 7 and use the optimal n to train our systems.

[2]The dataset can be downloaded from: `https://github.com/helfardy/starsem-2015-perspective/tree/master/dataset`

5.3.2 Word Sense Disambiguation (WSD)

Our goal of using Word Sense Disambiguation is to group synonyms together in order to map them to the same form. This process is intended to reduce the sparsity of the vocabulary and allow for an abstract generalization. Using WN-Sense-Relate (Patwardhan, Banerjee, and Pedersen, 2005), we tag each word in the input with its most frequent part of speech tag and Sense-ID. WN-Sense-Relate relies on WordNet (*WordNet: An Electronic Lexical Database*) to identify the part of speech tags and sense IDs, as well as to identify compounds. The only parts of speech handled by WN-Sense-Relate are adjectives (a), adverbs (r), verbs (v), and nouns (n). After tagging each word with the part of speech and sense information, we use WN-QueryData (Pedersen, Patwardhan, and Michelizzi, 2004) to retrieve the list of synonyms (synset) for each tagged word. All synonyms are then mapped to the same form. For example, "*The Democratic Party supports women 's equality , including equal pay , access to health care and other issues .*" becomes: "*the#ND*[3] *democratic_party#n#1 supports#v#1 women#n#1 's#ND equality#n#1 including#v#1 equal#a#1 pay#v#1 access#n#1 to#ND health_care#n#1 and#ND other#a#1 issues#n#1*".

The synset for "support#v#1" is support#v#1', back_up#v#1 so any occurrence of support#v#1' or back_up#v#1 in the data is mapped to the same form (ex. back_up#v#1). Finally, after all synonyms are mapped to the same form, we remove the part of speech and sense information from all words. While this process maps synonyms to the same form, a main drawback is that we only rely on the most frequent sense, which might not always be the appropriate choice. [4]

[3] #ND indicates a non-defined word

[4] We experimented with the contextual variant of the this approach but the most frequent sense yielded better results.

5.3.3 Weighted Matrix Factorization Features

The next set of features relies on mapping text from the high-dimensional n-gram space to a low-dimensional topic space.

Similar to the Arabic Perspective Identification work, we use the Weighted Matrix Factorization (WMF) system (Guo and Diab, 2012) and apply the default settings for the distributable version of WMF, which sets the number of topics (K) to 150 and the weight of the missing words to 0.01. We collect our training data from online discussion fora discussing American politics. The data contains 972,100 posts corresponding to ~20 million tokens. We build two WMF models. The first model only applies basic preprocessing to the text by excluding punctuation and numbers, converting all text to lower case and stemming all words; the second model (WMF-WSD) uses the same process explained in Section 5.3.2 to map all synonyms to a single form prior to stemming the text.

5.3.4 Sentiment Features

Sentiment also provides another important clue through which a person's perspective can be expressed. A person's perspective normally influences–and is expressed in–his/her sentiment towards different social, economic and political topics. One's stance on topics such as "legalization of abortion", "climate change", and "feminism", and "abortion" is normally determined by one's ideological leaning. Thus, identifying a person's opinion on different issues discussed in a post can also help us uncover his/her leaning. In this feature set, we use a heuristic for identifying the topics discussed and the sentiment expressed by the author towards these topics, and we measure how well they can help us in identifying the leaning of the given post.

Sentiment Polarity Tagging

We use Stanford's Sentiment Analysis System (Socher et al., 2013) to identify the positive and negative words in a given post.

Target Identification

We identify two types of targets from our training data:

- **Named-Entities:** We use the Stanford Named-Entity Recognizer (Finkel, Grenager, and Manning, 2005) to identify three types of Named-Entities (NEs): (a) Organizations, (b) Locations and (c) Persons;
- **Noun Groups:** As an approximation for the discussed topics, we also extract all Noun-Groups where we define a noun-group as any phrase starting with a noun and followed by zero or more nouns and prepositions.[5]

For each extracted target, we generate a Target-ID and replace all occurrences of this target with the generated ID. For example, in the Legalization of Abortion dataset, "Partial Birth Abortion" is replaced by "Target1", "fetus" is replaced by "Target2", "woman" by "Target3", etc. We then create a list of the targets that appear in the training data and create a feature for each target that appears in more than one training instance. We start by matching the sentence against the longest targets first then move to the shorter ones. For all of the extracted nominal targets, we use the lemma instead of the word itself in order to avoid redundancy. [6] Tables 5.7 and 5.8 show some of the extracted targets for SemEval 2016 and ANES datasets, respectively. For each one of the domains of SemEval datasets, it is quite easy to infer the domain from the extracted targets. For example, the targets in SemEval "Feminist Movement" dataset include "feminism", "feminist" and "rape". For

[5]We choose this approach–as opposed to using a base-phrase chunker to identify noun-phrases–because we are interested in shorter targets.

[6]The longest target has a length of five words while the shortest one consists of just one word.

ANES dataset, the targets are more diverse due to the structure of the post where each post discusses the author's views on Democratic and Republican parties as well as major issues facing the United States.

Assigning Polarity to Targets

In order to pair opinion-words with the identified targets, we extract all dependency relations using the Stanford Dependency-Parser (De Marneffe, MacCartney, and Manning, 2006).[7] For each dependency relation, we detect explicit–and not implicit–negation and flip the polarity of the opinion word if it is negated. For example, if the given sentence is "I don't like target1.", we detect that the opinion word "like" is negated through the dependency relation "neg(like, n't)"; therefore, we assign target1 a negative sentiment instead of a positive one.

Calculating Final Sentiment Score

Finally, we calculate the final opinion scores. For each post, if a target is paired with more than one opinion word–i.e. the target is mentioned more than once in the sentence and appears in more than one dependency relation with opinion words–we sum the scores of all of these opinion words to calculate the final opinion-score and assign the target its polarity as follows:

$$Final\ Sentiment\text{-}Score = \begin{cases} -1 & ; score < 0 \\ 1 & ; score > 0 \\ 0 & ; score = 0 \end{cases} \tag{5.1}$$

[7] We experimented with Tweebo parser (Kong et al., 2014), but Stanford's parser yielded a better output on all of our datasets.

	Targets
Abortion	baby, life, birth, will, abortion, conception, child, right, people, nation, thing, matter, love, procedure, part, woman
Atheism	people, prayer, man, hand, faith, child, religion, love, matter, zealot, god, lord, jesus, sinner
Climate Change	kid, generation, glblwarmingnews, people, damage, tooth
Hillary Clinton	america, way, china, country, hillary, person, president, obama, hillary clinton, vote
Feminist Movement	politics, life, guy, mother, hate, gender, way, girl, people, narrative, woman, feminist, definition, feminism, rape, man

Table 5.7: Examples of the extracted opinion targets from SemEval 2016 Stance dataset

	Targets
ANES	abortion rights, barack obama, birth control, business owners, class citizens, class people, class warfare, climate change, defense spending, democratic party, dream act, education system, gas prices, government programs, government regulations, government spending, guantanamo bay, gun laws, health care, healthcare plans, health care reform, health care system, healthcare system, health insurance, immigration policy, insurance company, job growth, job situation, lip service, manufacturing jobs, middle east, mitt romney, opposition party, party system, planned parenthood, republican party, ron paul, sex marriage, stock market, tax breaks, tax cuts, tax dollars, tax loopholes, tax payer money, tax rates, tea party, terrorist country, troops home, unemployment rates, united states

Table 5.8: Examples of the extracted opinion targets from ANES dataset

5.3.5 Linguistic Inquiry & Word Count (LIWC)

Another method for inferring the discussed topics and opinions in a given text is to identify the percentage of words belonging to a set of interpretable categories. The Linguistic Inquiry and Word Count (LIWC) toolkit (Tausczik and Pennebaker, 2010) serves such a purpose. LIWC toolkit relies on a set of dictionaries to assign words in a given text to a set of psychologically meaningful categories such as death, science, emotion, space, family, swear words and many others. While exploring the topics inferred from WMF or the lexical features for each stance and perspective can help us understand the focus of different ideologies, the interpretability of LIWC features and the comprehensive set of categories it outputs provides better insights into the focus of people having different leanings. A main drawback to using LIWC toolkit is that it only performs a dictionary look-up on each word, hence judging words independent of their context.

We apply LIWC to each given post/tweet and estimate the percentage of words in each post belonging to each of the following categories: *swear words, social processes, family, friends, humans, affective processes, positive emotions, negative emotions, anxiety, anger, sadness, cognitive processes, insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusion exclusion, perceptual processes, seeing, hearing, feeling, biological processes, body, health, sexual, ingestion, relativity, motion, space, time, work, achievement, leisure, home, money, religion, death, assent, non-fluencies and fillers*. Out of all posts having a specific stance/perspective, we calculate the percentage of posts that have at least one word in each LIWC category in order to identify the focus of people according to their leanings. Tables 5.9 and 5.10 show these percentages for some of the LIWC categories in the two studied datasets. By analyzing the numbers in Table 5.9, we find that the "Atheism" dataset has the highest coverage from the "Religion" category, the "Legalization of Abortion" dataset has the highest coverage from the "Death" category,

while "Climate Change" has the highest coverage from the "Motion" category. Moreover, for "Legalization of Abortion", 22% of the tweets opposing it use words in the "death" category as opposed to only 10.6% of the tweets favoring the topic. The "Atheism" dataset has the highest coverage for the "religion" category, where 76.5% and 77.4% of the tweets favoring and opposing the topic, respectively, use words in this category. For ANES dataset, the "religion" category is more frequent in the pro-Romney posts (60.2%) as opposed to 56.2% for the pro-Obama posts. However, for most categories the percentages are quite close in both pro-Obama and pro-Romney posts. We attribute this to the nature of this dataset where each post discusses what the person both likes and dislikes about the Democratic and Republican parties, and their associated public figures, rather than focusing on just one side. Hence, we do not expect LIWC features to help in discriminating among the different classes of ANES.

5.3.6 Bag of Frames

Another way of identifying the focus of the given posts is by performing frame semantic parsing on each post. Frame semantics assemble the meanings of different elements in a given piece of text to model the meaning of the whole text (Baker, Fillmore, and Lowe, 1998). The basic semantic unit in frame semantics theory is the *"frame"*. A frame is a conceptual structure that refers to a group of related concepts (or elements) where understanding one of these concepts requires an understanding of its whole structure. When any of these structures is present, it automatically triggers all of the other ones in the reader's mind (Fillmore, 2006). The word (or phrase) that triggers the frame is called the *"frame target"*. For example, the target for the frame "Killing" can be any kill verb, and the frame elements will include the *killer* and the *victim*.

We use SEMAFOR (Chen et al., 2010; Das et al., 2010), a publicly available frame-semantic parser, to identify all the semantic frames in each given post/tweet. For example, in the tweet *"Because I want young American women to be able to be proud of the 1st*

	Abortion		Atheism		Climate Change		Hillary Clinton		Feminism	
	Favor	Against	Favor	Against	Favor	Against	Favor	Against	Favor	Against
swear words	4.6	3.1	6.5	0.9	2.7	3.8	1.3	4.1	9.7	9.6
social	88.7	83.2	72.6	79.7	57.6	57.7	78.2	75.4	88.1	86.9
family	4	8.6	8.1	6.2	2.1	0	5.8	3.2	5.6	4.7
friends	0.7	1.5	1.6	0.9	1.5	0	3.2	0.8	2.6	3.1
humans	53	45.3	16.1	15.7	12.2	26.9	16.7	13.9	58.2	47.6
affective	60.9	63.1	73.4	69.4	52.8	57.7	69.9	65.5	68.7	70.8
pos. emot.	39.7	41.5	54.8	56.9	32.5	30.8	62.2	43.7	41.4	45.4
neg. emot.	37.1	36.9	37.9	26.7	29.3	30.8	19.9	34.1	42.9	47.2
anxiety	4	4.2	5.6	5	6.6	3.8	2.6	3.8	6	4.5
anger	23.8	25	24.2	8.2	11.3	7.7	9.6	22	28.4	31.5
sadness	7.9	4.8	5.6	5.4	6.6	19.2	4.5	4.1	2.2	8
cognitive	89.4	89.5	92.7	90.7	80	88.5	76.3	84.2	86.2	87.3
certainty	14.6	24.9	28.2	30	15.5	15.4	18.6	19.9	21.6	23.3
body	23.2	11.7	9.7	8	6	23.1	3.8	5.8	14.6	13.5
health	48.3	43.6	7.3	10.6	6.3	3.8	5.1	4.3	7.1	5.1
sexual	41.7	31	8.1	8.8	1.8	3.8	4.5	5.4	22.8	19.8
work	11.9	17.2	24.2	21.6	24.5	11.5	37.2	35.3	20.1	19
achieve	27.8	17.8	25	33.6	23.3	26.9	41	28.5	19	21.5
leisure	4	8.6	12.1	12.5	8.7	15.4	12.8	10.9	13.1	25.2
home	1.3	2.7	2.4	3	8.7	3.8	3.8	3.6	5.6	5.7
money	6.6	6.7	5.6	5.2	13.4	7.7	7.1	11.4	9	8.4
religion	7.3	18	77.4	76.5	3.9	0	3.8	4.3	4.9	5.7
death	10.6	22	6.5	3.4	3.9	3.8	3.2	3.9	3.4	3.5
assent	5.3	4	4.8	2.2	3	3.8	8.3	6.9	4.9	5.7

Table 5.9: Percentage of posts that use words in each of the shown LIWC categories in Semeval 2016 Stance dataset

ANES	Obama	Romney	Neither
swear words	11.3	13.3	10.9
social	99.7	100	100
family	34	35.6	25.6
friends	9.9	9.5	10.1
humans	91	91.3	89.1
affective	100	100	100
pos. emot.	99.4	99.6	97.7
neg. emot.	99.6	99.6	100
anxiety	95.7	93.2	93
anger	89.6	88.3	86.8
sadness	66.3	59.8	69.8
cognitive	100	100	99.2
certainty	91.3	92.4	87.6
body	43.1	38.3	34.1
health	87.1	88.3	89.1
sexual	59.9	60.6	55.8
work	99.7	100	99.2
leisure	78.5	81.1	74.4
home	41.9	42.4	39.5
money	97.8	97	98.4
religion	56.2	60.2	54.3
death	49.7	49.6	46.5
assent	39.9	34.8	34.1

Table 5.10: Percentage of posts that use words in each of the shown LIWC categories in ANES dataset

woman president.", SEMAFOR identifies the following frames: *"Leadership Target: president", "Capability Target: able", "Origin Target: American","Desiring Target: want", "People Target: women" and "Age Target: young"* . We create a list of all the frames that occur in the training data and use binary features to indicate the presence/absence of each of them in each given post. This set of features provides yet another abstraction in order to infer the topics discussed in the given text.

5.3.7 Machine Learning Model

Similar to the work on Arabic Perspective Identification, we divide the feature sets into three categories, train a single Logistic Regression classifier for each feature category, and combine the best set-ups from each of these categories in classifier ensemble.
We use the following categories of features:

- **N-grams:** This category includes both Basic and WSD n-grams;
- **WMF:** This category includes WMF and WMF-WSD features;
- **Pragmatics:** This category includes sentiment, LIWC and bag of frames features.

For the model's parameters, we use the "One vs Rest" implementation in Scikit-Learn toolkit (Pedregosa et al., 2011) and apply L2 regularization while setting the inverse of regularization strength (C) to 10.

5.4 Experiments and Results

5.4.1 Baselines

We compare our approach to three baselines;

- **Majority Baseline *(MAJ-BL)***: which assigns all posts to the most frequent class-label;

- **Random Baseline *(RAND-BL)***: which randomly chooses the class-label;
- **N-gram Baseline *(NGRM-BL)***: a strong baseline that uses standard n-gram features that are preprocessed using basic preprocessing scheme.

Additionally, for SemEval's held-out test set we compare the performance to the best participating system in the task.

5.4.2 Evaluation Metrics

We use the official metric of the shared task to evaluate the performance of our approach on SemEval 2016 Stance dataset and use weighted average $F_{\beta=1}$ score to evaluate the performance on ANES dataset. Both metrics rely on the $F_{\beta=1}$ of individual classes but differ in the way they combine these scores.

SemEval 2016 Stance Evaluation Metric

For SemEval dataset, we use the official evaluation script for the task to evaluate systems. While each tweet can have one of three class labels–"Favor", "Against" or "None"–the official metric calculates the performance of each system as the non-weighted average $F_{\beta=1}$ score of the first two class labels only.

$$Final\ \ F_{\beta=1}\ score = \frac{1}{2} * [F_{\beta=1}(Favor) + F_{\beta=1}(Against)] \tag{5.2}$$

ANES Evaluation Metric

For ANES datasets, we combine the scores from the different classes by performing weighted averaging. We calculate the weighted average $F_{\beta=1}$ of all classes as follows:

$$Weighted\ Average\ F_{\beta=1}\ score = \sum_{c=1}^{C} \frac{n_c}{n} * F_{\beta=1}\ score(c) \tag{5.3}$$

where:

n_c: is the total number of test instances whose true class label is c,

n: is the total number of test instances,

C: is the total number of classes.

5.4.3 Results

For all datasets, we tune on a development set and apply the setup that yields best results on the development set to the held-out test set. For the held-out test set experiments, we train on both training and development sets. [8]

Tuning N-grams

We experiment with n-grams having a maximum length between 1 and 7 in order to identify the best (n). Tables 5.11 and 5.12 show the results. On SemEval dataset, while the performance varies across different domains, the best overall value of *n* is also 3. However, for ANES, unigrams perform best. This is not surprising given that the length of posts (~420 words) in this dataset is much longer than SemEval dataset. This results in higher n-grams making the n-gram feature space very sparse for ANES and therefore degrading the performance. Accordingly we set *n*=1 for ANES dataset and *n*=3 for SemEval dataset.

[8] We do not include statistical significance results due to the small size of the datasets.

Max. (n)	Abortion	Atheism	Climate	Clinton	Feminism	Avg
1	**56.6**	54.7	**41.5**	**63.4**	49	60.5
3	47.6	55.6	40	59.5	**54.1**	**62.1**
5	50.1	**56.8**	40.5	52.6	51.5	61.7
7	45	47.7	40.5	52.6	51.1	60.5

Table 5.11: Tuning N-grams for SemEval 2016 Stance dataset

Max. (n)	PCC
1	**70.1**
3	66.7
5	62.1
7	53.9

Table 5.12: Tuning N-grams for ANES dataset

SemEval 2016 Stance Results

We evaluate the performance of each of the different setups within each of the three feature groups: (1) N-grams, (2) WMF, and, (3) Pragmatics Features before combining the best setups from different feature groups in a classifier ensemble. For N-grams and WMF, we experiment with Basic and WSD setups as well as the combination of both. Table 5.13 shows the results of using the feature sets within each feature group on SemEval 2016 Stance dataset. Applying WSD to N-grams improves the performance on three out of the five domains but overall performs slightly less than the Basic setup. However, for WMF features the WSD setup performs much better than the Basic setup. Combining both Basic and WSD features degrades the performance of the best setup (whether Basic or WSD) on all domains. Besides being based on tweets and having very short post/tweet length, the size of this dataset in terms of number of instances is very small (~500 and ~56 tweets per domain in the training and development sets, respectively), so doubling the number

	Abortion	Atheism	Climate	Clinton	Feminism	Avg
MAJ-BL	35.9	37	35	36	33	54.2
RAND-BL	48.3	27.2	15.6	31.8	39.4	36.3
Ngrm (Basic) (NGRM-BL)	47.6	55.6	40	**59.5**	**54.1**	**62.1**
Ngrm (WSD)	57.4	53.4	**42.9**	46.1	51.1	61.6
Ngrm (Basic+WSD)	53.3	54.8	41.5	53.9	49.5	61.3
WMF	53.4	36	25.6	40.4	49.2	50.8
WMF (WSD)	**63.4**	**60.5**	38.5	39.6	48.8	58.6
WMF (Basic+WSD)	60.6	54.6	34.3	36.9	45.6	54.2
Sentiment	34.4	37	35	36	37.9	54.7
Sentiment+LIWC	60	36.9	32.6	50.4	48.5	56.2
Sentiment+LIWC+BOF	39	52.7	30	36.9	39.8	46

Table 5.13: Results of using each feature-set in each feature category separately on SemEval 2016 Stance development set

of features by combining both Basic and WSD setups without increasing the training instances hurts the performance. Accordingly, we only include either Basic or WSD setups in the next experiments.

For pragmatic features, adding LIWC features to sentiment features results in either similar or better performance than using sentiment features only, except on the "Climate Change" domain where LIWC features cause the performance to drop. Generally, the performance on "Climate Change" domain is much lower than all other domains. We believe that this can be attributed to the small size of this set and to its very skewed distribution. 43.6% of the tweets belong to "None" class whose $F_{\beta=1}$ score is not taken into consideration in the task's official evaluation metric. Overall, using Pragmatic features separately fails to beat the N-gram baseline. This is not surprising given how

	Abortion	Atheism	Climate	Clinton	Feminism	Avg
MAJ-BL	35.9	37	35	36	33	54.2
RAND-BL	48.3	27.2	15.6	31.8	39.4	36.3
NGRM-BL	47.6	55.6	40	**59.5**	54.1	62.1
Basic						
Ens.(Ngrm,WMF)	51.9	55.5	42.9	51.5	**56.1**	62.8
Ens.(Ngrm,Sentiment)	52.6	55.8	39.5	52.9	51.9	62
Ens.(Ngrm,Sentiment+LIWC)	51.6	47.4	40	**59.5**	48	60.6
Ens.(Ngrm,Sentiment+LIWC+BOF)	47.4	58	40.5	57.1	40.7	56.9
Ens.(Ngrm,WMF,Sentiment)	62.3	47.5	**43.2**	51.7	54.9	**64.7**
Ens.(Ngrm,WMF,Sentiment+LIWC)	55.2	48.7	40	52.2	53.3	62.1
Ens.(Ngrm,WMF,Sentiment+LIWC+BOF)	55.6	**62.2**	40.9	41.8	49.3	60.2
WSD						
Ens.(Ngrm,WMF)	61.1	57.9	41.5	43.6	52.4	61.5
Ens.(Ngrm,Sentiment)	**67.6**	56.8	41.9	45.7	49.6	63.3
Ens.(Ngrm,Sentiment+LIWC)	58.3	48.5	41.9	50.3	47.1	60.6
Ens.(Ngrm,Sentiment+LIWC+BOF)	57.8	53.6	40.5	48.4	41.3	57.2
Ens.(Ngrm,WMF,Sentiment)	65.9	46.4	41.5	37.4	48.2	60.1
Ens.(Ngrm,WMF,Sentiment+LIWC)	58.3	43.4	40.5	44.7	50.8	59.3
Ens.(Ngrm,WMF,Sentiment+LIWC+BOF)	62.7	57.9	41.9	32.9	49.3	60

Table 5.14: Results of combining feature groups using a classifier ensemble on SemEval development set

	Abortion	Atheism	Climate	Clinton	Feminism	Avg
MAJ-BL	40.3	42.1	42.1	36.8	39.1	65.2
RAND-BL	27.8	28.7	33.3	30.1	28.7	31.9
NGRM-BL	56.2	50.6	38.7	46	56.2	64.2
Best Task System (MITRE)	**57.3**	**61.5**	41.6	**57.7**	**62.1**	**67.8**
Ens.(BasicNgrm,WMF,Sentiment)	48.4	53.8	42.9	41.3	56.7	66.6

Table 5.15: Results of the best development setup on SemEval held-out test set

well N-grams generally perform on all text categorization tasks. Adding Bag of Frames features to Sentiment and LIWC features hurts the performance except on the "Atheism" tweets. Overall, for pragmatic features, combining Sentiment and LIWC performs best, or close to best, on all domains except "Atheism" where combining Sentiment, LIWC and Bag of Frames performs best and "Climate Change" where using only Sentiment features performs best.

Table 5.14 shows the results of combining the best setups from each group using a classifier ensemble approach on SemEval dataset. The best configuration differs across all five domains. This size of the development set is quite small (~59 tweets per domain), so the results are not very robust. A difference in classification of just one tweet between two setups can impact the performance greatly. Overall, combining N-gram, WMF and Sentiment features without WSD performs best.

Held-Out Test Results: We apply the performance of the setup that resulted in best overall performance–combining Ngram, WMF and Sentiment–on the development set to the held-out test set. Table 5.15 shows the results. Unlike the development set, the held-out test set is much larger–~250 tweets per domain–which is approximately half the size of the training data. Overall, the best development setup yields an overall performance of 66.7%, 1.2% below the best participating system (MITRE).

			Predicted		
			Favor	Against	None
Abortion	**Gold**	Favor	1.8	12.1	2.5
		Against	1.4	60.4	5.7
		None	0.4	12.9	2.9
			Favor	Against	None
Atheism	**Gold**	Favor	2.3	11.4	0.9
		Against	2.3	68.6	1.8
		None	0.5	10.0	2.3
			Favor	Against	None
Climate Change	**Gold**	Favor	67.5	0.0	5.3
		Against	4.1	0.0	2.4
		None	13.0	0.0	7.7
			Favor	Against	None
Clinton	**Gold**	Favor	0.7	14.2	0.3
		Against	0.0	56.6	1.7
		None	0.0	23.7	2.7
			Favor	Against	None
Feminism	**Gold**	Favor	8.8	11.2	0.4
		Against	12.3	50.9	1.1
		None	2.5	12.3	0.7

Table 5.16: Confusion matrices for all domains of SemEval 2016 Stance test set

	PCC
MAJ-BL	49.4
RAND-BL	37.4
Ngrm-Basic (NGRM-BL)	70.1
Ngrm-Wsd	65.5
Ngrm-All	**74.2**
WMF	60.6
WMF-Wsd	66.6
WMF-All	69.1
Sentiment	54.6
Sentiment+LIWC	55.8
Sentiment+LIWC+BOF	59

Table 5.17: Results of using different combinations in each feature group on ANES development set

We look into the confusion matrices of the best overall held-out test setup for all domains in order to identify the sources of confusion and how the systems can perform better. Table 5.16 shows the confusion matrices for all five domains. For all domains except "Climate Change", the highest confusability is between the "Against" class (the most frequent class) and both of the other classes. For "Climate Change" dataset none of the tweets gets assigned "Against" class. The reason behind this is that even though "Against" class is the most frequent class in all other four domains, for "Climate Change" only 3.8% of the training data belongs to this class, so the classifier is biased against it.

ANES Results

Finally, we analyze the performance of our systems on the second task, identifying the presidential candidate choice (PCC) of a person based on his/her responses to the open-ended ANES questions. We begin by assessing the difficulty of the task and the feasibility

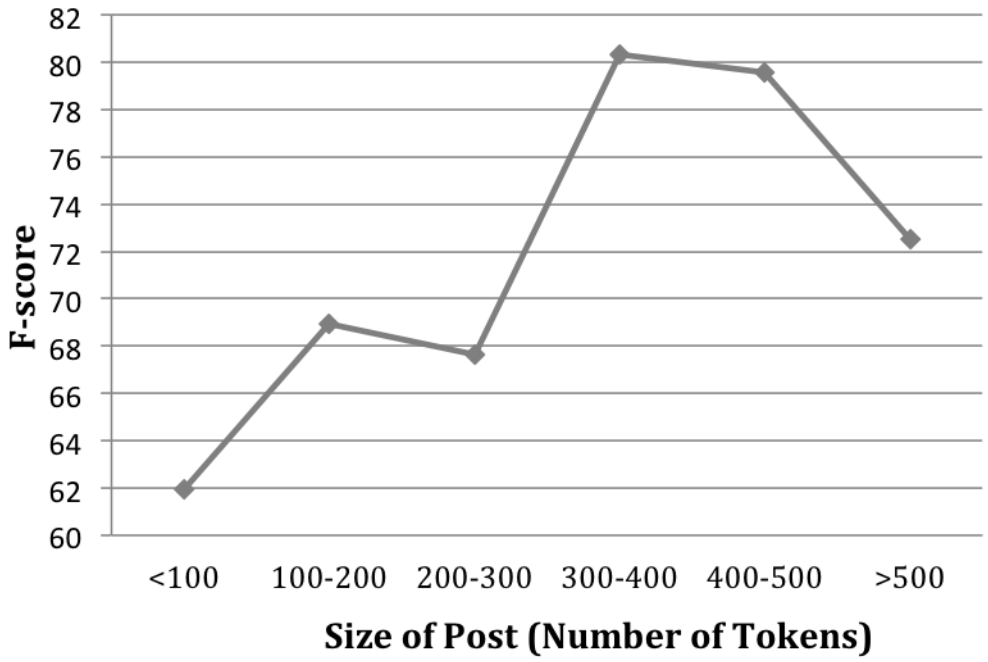

Figure 5.1: $F_{\beta=1}$ score of human judgments in predicting "PCC" from the answers to the essay questions in ANES dataset across different post-sizes

of our approach for constructing posts based on the answers to different questions. We conduct a human evaluation by running an Amazon Mechanical Turk experiment on all ANES posts where we ask Turkers to read each post and guess the PCC of the person who wrote the post. Additionally, we ask them to write the reasons behind their choice and reject all entries that do not specify at least one reason. We found that Turkers were able to predict the PCC with an average $F_{\beta=1}$ score of ~75%. We also found that the task is particularly difficult for very short ($<$100 words) posts and that the posts that have a length of 300-500 words yield the best performance. Figure 5.1 shows the results of this qualitative assessment.

Next, we evaluate the performance of different feature sets on automatically identifying the PCC. Tables 5.17 and 5.18 show the results of using feature-sets in a single feature category and combining the best setups from each feature category in a classifier

	PCC
MAJ-BL	49.4
RAND-BL	37.4
NGRM-BL	70.1
Basic+WSD	
Ens.(Ngrm,WMF)	**75**
Ens.(Ngrm,Sentiment+LIWC+BOF)	72.7
Ens.(Ngrm,WMF,Sentiment+LIWC+BOF)	70.8

Table 5.18: Results of combining feature groups using a classifier ensemble on ANES development set

	PCC
Maj-BL	54.5
Rand-BL	29.8
NGRM-BL	69.4
Ensemble (Ngrm (Basic+WSD), WMF (Basic+WSD))	**72.9**

Table 5.19: Results of the best development setups on ANES held-out test set

ensemble, respectively. Unlike the stance datasets, combining both Basic and WSD features outperforms using either feature set. Similarly, combining Sentiment, LIWC and Frame features performs better than only using Sentiment or Sentiment and LIWC. Similar to the previous task, N-grams outperform all other features. For the classifier ensemble, combining only N-grams and WMF performs best, beating the N-gram baseline with 4.9%. Adding Sentiment, LIWC and Frame features hurts the performance but still outperforms all baselines.

Held-Out Test Results: On the held-out test set (Table 5.19), using the same best set-up–combining Basic and WSD N-grams and WMF features–outperforms all baselines yielding a 3.5% improvement over the strongest baseline and only ~2% below the human perfor-

		Predicted		
		Obama	Romney	Neither
Gold	Obama	64.8	2.8	0.0
	Romney	12.0	6.5	0.0
	Neither	12.0	0.9	0.9

Table 5.20: Confusion matrix for ANES held-out test set

mance.We look at the confusion matrix of this best setup (Table 5.20) and find that the majority of the confusability in the test data is between the most frequent class, "Obama", and both of "Romney" and "Neither" classes. Since the training data is highly biased– where around 63% of the posts chose "Obama" for PCC– the system is biased towards this class.

5.4.4 Discussion

We perform a manual error analysis in order to gain better insights into how the systems can perform better (Table 5.21 show some misclassification examples for SemEval dataset), and find the following:

1. In ANES dataset, due to the structure of the questions, some Turkers were trying to be objective towards both political parties, which makes it difficult even for a human evaluator to identify the political leaning of the person who wrote the text;

2. The use of sarcasm, which can be easily detected by human evaluators but not by our system. For example, in Abortion dataset, a participant who supports legalization of abortion wrote *"Why should people use reason and logic to discover right and wrong when a priest can decide for them?"*. A possible solution is to use a sarcasm detection system such as the one presented by Ghosh, Guo, and Muresan, 2015 and use the output of that as a feature within our model;

3. Misspelled words such as writing *"Romeny"* instead of *"Romney"*;

4. In SemEval dataset, due to the very short length and the limited context of each tweet, identifying the correct label is not straightforward. Even though the annotation was performed by humans and not automatically, the annotators misjudged some of the tweets. For example, the following "Hillary Clinton" tweet *"@HillaryClinton: Here's to fearless women chasing their goals. Congratulations, Team #USA!"* was labeled as being "Against" Hillary Clinton when it is actually implicitly praised Clinton's pursuit of her goals.Philip Related work

   Similarly for the *"Feminist Movement" tweet: "Just want gender politics to be over. My brain hurts. #genderequity #patriarchy #MRA"*, even though it was labeled as being "Against" feminism, it can be argued that the stance is not very clear since the author highlighted both "patriarchy" and "gender equity".

While both datasets are informal, they vary in their levels of informality as well as in the length of the posts. These variations affect the performance of each one of them. Nevertheless, on both datasets, lexical features have the most impact. Moreover, latent semantics–WMF–is the second most important feature-set. In most domains of "SemEval" and on "ANES" datasets, WMF improved over using only lexical features on both the development and held-out test sets. Finally, pragmatic features do not seem to pattern per dataset. Hence, tuning per dataset and per domain is important in order to identify the best set of pragmatic features.

Domain		Gold	Pred.	Tweet
Abortion	1	None	Against	Yes. Antis just don't make sense. #WarOnWomen
	2	Against	Favor	I have a right to identify as pregnant and have an abortion. Having an abortion confirms my power of choice. #PregnancyForAll
Atheism	3	Against	Favor	These days, the cool kids are atheists. #freethinker
	4	Favor	Against	#Religions can't all be right, but they can all be wrong.
Climate Change	5	Against	None	ONE Volcano emits more pollution than man has in our HISTORY!
	6	Against	Favor	The only thing "man made" about global warming is the false narrative. #WakeUpAmerica #boycottSanFrancisco #Election2016
Clinton	7	Against	Favor	@HillaryClinton: Here's to fearless women chasing their goals. Congratulations, Team #USA!
	8	Favor	Against	You know, when you talk bad about Hillary, in a sense, you're talking bad about me.
Feminism	9	Favor	Against	Rather be an "ugly" feminist then be these sad people that throws hat on people that believes in equality!
	10	Against	Favor	Just want gender politics to be over. My brain hurts. #genderequity #patriarchy #MRA

Table 5.21: Misclassfication Examples in SemEval 2016 Stance held-out test set

5.5 Summary

In this chapter, we explored the use of lexical and semantic features in performing automatic identification of ideological perspective from written text. We addressed two tasks. The first task is identifying the stance of a person towards topics that are influenced by one's belief system while the second is identifying a person's 2012 Presidential Candidate Choice. For the first task–identifying the stance of the person–we evaluate the performance of the proposed approach on the dataset used for SemEval 2016 Task 6 "Detecting Stance in Tweets", which aims at identifying the stance of a given tweet towards five targets; "Abortion", "Atheism", "Climate Change", "Hillary Clinton" and "Feminist Movement".

For both tasks, we explore the use of standard N-gram features, Word Sense Disambiguation (WSD), targeted sentiment, a latent semantics model tailored for short texts (WMF), frame semantics and Linguistic Inquiry and Count Features that assign words in a given text to a set of psychologically meaningful categories. For both tasks, using lexical features whether with or without WSD performs best among all features. Adding WMF features improves the results over using only N-grams. Pragmatic features–Sentiment, LIWC and frame semantics–yield mixed results.

Chapter 6

Related Computational Work

In this chapter, we review the related computational work and compare it to the work presented in this thesis.

6.1 Perspective Identification in English

Current computational linguistics research on automatic perspective identification uses both supervised and unsupervised techniques. The main task handled by supervised approaches is to perform document/post-level perspective or stance classification, whether binary or multiway. Unsupervised approaches, on the other hand, mainly try to cluster users in a discussion.

One of the early works on binary perspective identification is that of Lin et al., 2006 (Lin06), which uses articles from the Bitter-Lemons website–a website that discusses the Palestinian-Israeli conflict from each side's point of view–to train a system for performing automatic perspective identification on the sentence and document levels. On the website, an Israeli editor and a Palestinian editor, together with invited guests, contribute articles to the website on a weekly basis. Lin06 uses bag-of-words features, assuming once a binomial and once a multinomial distribution for words in each article, and use a balanced dataset to evaluate their approach. The authors run different experiments in which they vary the training and test sets between: (a) editors' articles, and (b) guests' articles. The accuracies of the different experimental conditions vary between 86% and 99%. As one

might expect, the highest accuracy (99%) is achieved by the system that is trained and tested on the editors' articles. For this system, the classifier is not only capturing the perspective but also the editors' writing styles. Overall, Bitter-Lemons's corpus is much more formal than the ones we studied in this thesis since it is based on edited articles and not spontaneously occurring informal text, such as the language used in blogs, discussion fora and tweets. Moreover, the average length of an article is much longer than that of a tweet or a discussion forum comment, hence most of the challenges imposed in the genre we target are absent in this genre. It is very likely that when moving to less formal genres, the systems' performance yielded by Lin06 will drop. In Klebanov, Beigman, and Diermeier, 2010, the authors tackle the same problem of binary-perspective identification and experiment with four corpora from different genres corresponding to different levels of formality. The first corpus, Bitter-Lemons, is the same one used by Lin06 while the second one, Bitter-Lemons International, contains articles from the same website that discuss other Middle Eastern issues. The third corpus comprises posts collected from several blogs discussing the "Death Penalty", while the last corpus contains transcripts of U.S. House and Senate debates on "Partial Birth Abortion" (PBA). The authors show that using term-frequencies does not improve over using binary bag-of-words and that using only the best 1-4.9% features is sufficient to achieve high accuracy. They achieve the highest accuracy (97%) on the PBA dataset and the lowest accuracy (65%) on the Bitter-Lemons International dataset. For the PBA dataset, the language used in House of Representatives debate is more formal than that used in blogs or discussion fora or other informal genres. Unlike our work, both of the previous works use only lexical features and do not explore the use of any semantic or pragmatic features.

Somasundaran and Wiebe, 2010 employ the notion of *"arguing"* to identify a person's stance (supporting or opposing) towards a topic. Arguing can utilize either positive lexical cues, such as *"actually"*, or negative ones such as *"certainly not"*. They construct

an "arguing" lexicon and use it to derive features for their classifier. They experiment with both arguing and sentiment features on ideological debates pertaining to four domains; "Abortion", "Creationism", "Gay Rights" and "Gun Rights". They show that combining arguing and sentiment features outperforms a unigram baseline on "abortion", "gay rights" and "gun rights" datasets while the unigram system performs best on the "creationism" dataset. Their results go along with our findings that there is no one setup that fits all datasets but rather that tuning the system for each task and domain is necessary to achieve optimal performance. Overall, they achieve a cross validation accuracy of 63.93%.

Another work that addresses the same problem on binary stance identification is that of Anand et al., 2011. The authors use n-grams, document statistics, percentage of words in different LIWC (Tausczik and Pennebaker, 2010) categories, punctuation and syntactic dependencies as features for post-stance classification. They evaluate their approach on a variety of both ideological (ex. abortion, climate change and death penalty) and non-ideological (ex. Mac versus PC, Firefox versus IDE) topics. While the proposed approach beats the n-gram baseline on most domains, they find that–similar to Somasundaran and Wiebe, 2010–the best setup varies across different domains, which is again consistent with our findings. The best cross-validation accuracy of their approach varies between 53.75% and 62.31% for non-ideological domains and 53.42% and 69.23% for ideological ones.

Hasan and Ng, 2012 extend the previous work by using Integer Linear Programming (ILP) to perform joint inference over the predictions made by a post-stance classifier and several topic-stance classifiers, and by extending the features used. In addition to the features proposed in (Anand et al., 2011), the authors use sentence-type, topic features–where they define a topic as a word sequence starting with zero or more adjectives followed by one or more nouns and topic-opinion features. They create topic-opinion features by first identifying both the topics and the sentiment words in each sentence and associating each

topic with a positive (or negative) sentiment if a dependency relation exists between the topic and a sentiment word. They collect debate posts discussing abortion and gun-rights and achieve an $F_{\beta=1}$ score of 57.8% on the abortion dataset, and 61.1% on the gun-rights dataset. It is worth mentioning that the topic-opinion features used in this work are slightly different than the targeted-sentiment features we use in our English work; instead of manually labeling the topics with the stance, we rely completely on the automatically inferred stances. In Hasan and Ng, 2013, they extend their previous work by incorporating two soft-constraints that treat the task of post-stance classification as a sequence-labeling problem and ensure that the topic-stance of each author is consistent across all posts.

More recently, 19 systems participated in *SemEval-2016 Task 6: Detecting Stance in Tweets* (Mohammad et al., 2016). The systems use a variety of features and machine learning approaches (Augenstein, Vlachos, and Bontcheva, 2016; Bøhler et al., 2016; Boltuzic et al., 2016; Dias and Becker, 2016; Elfardy and Diab, 2016b; Igarashi et al., 2016; Krejzl and Steinberger, 2016; Liu et al., 2016; Misra et al., 2016; Patra, Das, and Bandyopadhyay, 2016; Vijayaraghavan et al., 2016; Wei et al., 2016; Wojatzki and Zesch, 2016; Zarrella and Marsh, 2016; Zhang and Lan, 2016). The best system (Zarrella and Marsh, 2016) achieved an average $F_{\beta=1}$ score of 67.82%. It used transfer learning and two Recurrent Neural Networks (RNNs) to infer the stance of the given tweets. The first RNN uses a very large unlabeled set of tweets and learns to predict the hashtags in these tweets. The second RNN is trained to predict the stance labels using the task's dataset and is initialized using the parameters learned from the first RNN. While our systems achieved an $F_{\beta=1}$ score of 66.6% (1.2% below this system), we did not use RNNs or any external resources.

Other work that uses Recursive Neural Networks (RNNs) for a quite related task is that of Iyyer et al., 2014. In this work, the authors explore detecting *Liberal* versus

Conservative bias through the use of RNNs. The authors collect annotations for different phrases in a given parse-tree and utilize this information to build a RNN that can model compositionality in a given sentence. They find that RNNs outperform standard approaches that only rely on bag-of-words as well as stronger baselines. The authors evaluate their approach on two datasets. The first one is a dataset of U.S. Congressional debates where each sentence is annotated as having either a *Conservative* or *Liberal* bias. The dataset has a total of ~8,000 sentences balanced across the two classes. The second dataset comprises a set of ~12,000 books and magazine articles. For both datasets, the authors extract different phrases and have them annotated for the *Conservative-Liberal* bias. They achieve an accuracy of 70.2% and 69.3% on both datasets and beat all baselines. As opposed to this work, we are interested in identifying different elements governing the perspective on the post/document–not sentence–level.

Yano, Resnik, and Smith, 2010 also look at the conservative-liberal bias by studying the linguistic cues for bias in political blogs. The authors draw sentences from American political blogs and annotate them for bias on Amazon Mechanical Turk. They explore whether the Turkers' decisions are influenced by their perspectives: for example, whether a self-proclaimed liberal Turker is more likely to view sentences written by a conservative as biased and vice versa. Since in our Arabic annotation experiment, the data was annotated by only four annotators, and since all of these four annotators self-identified as being pro-reform, we could not conduct a similar analysis.

Similar to (Iyyer et al., 2014) and (Yano, Resnik, and Smith, 2010), we focus not only on identifying the stance of a given post/tweet towards a specific topic of interest where the dataset only discusses that very specific topic. Our work identifies more general notions of perspective by looking at datasets that discuss broader topics, but unlike both works that only focus on English, we studied the problem from a multilingual standpoint.

Another quite related task to our work is attempting to subgroup discussants in an ideological discussion (Abu-Jbara et al., 2012; Dasigi, Guo, and Diab, 2012). In Abu-Jbara et al., 2012 the authors perform subgroup detection by clustering authors according to their sentiment towards topics, named-entities as well as other discussants. Dasigi, Guo, and Diab, 2012 extend the previous work by introducing the notion of implicit attitude, which models the similarity between the topics discussed by a pair of people. They note that people who share the same opinion tend to discuss similar topics, thus having a high semantic similarity. By explicitly modeling latent sentential semantics as a stand in for implicit attitude, they achieve an $F_{\beta=1}$ score improvement of 3.83%, and 2.12% on the task of subgroup detection within "Wikipedia-Discussions" and "Online-Debates" datasets respectively.

Volkova, Coppersmith, and Van Durme, 2014 use a dynamic Baysian model that relies on different notions of similarity between twitter users to predict political preferences of users even in the absence of self-authored tweets. The authors find that exploring the content of the neighbors of a Twitter users are as–and sometimes more– helpful than the user's self-authored text and that the most helpful neighborhood measure for predicting political preferences among Twitter users involves friends, user mentions and retweets. Moreover, they find that dynamic models are capable of achieving better performance than batch models for predicting political preferences. As opposed to this work, our datasets do not provide such helpful neighborhood meta-data that can help in predicting people's leanings. Accordingly we only rely on the self-authored textual content in identifying those leanings.

6.2 Perspective Identification in Arabic

Most of the current research on Perspective Identification targets English. To the best of our knowledge, the only research efforts that target Arabic Ideological Perspective from a computational point of view are those of Abu-Jbara et al., 2013, Siegel, 2014, Al Khatib, Schütze, and Kantner, 2012 and Borge-Holthoefer et al., 2015.

In Abu-Jbara et al., 2013, the authors use the same approach of clustering users in a discussion that was previously applied to English datasets (Abu-Jbara et al., 2012). The authors rely on targeted-sentiment and LDA topic model features and experiment with different mechanisms to cluster participants. The approach was evaluated on a dataset of Modern Standard Arabic discussion fora discussing political parties. The devised system achieved an $F_{\beta=1}$ score of 76%. Unlike our Egyptian dataset, Abu-Jbara et al., 2013's dataset is self annotated, and the annotations are more abstract–only providing the binary stance of each post towards the debate question. Moreover, their approach only handles Modern Standard Arabic and does not address any of the challenges posed by Dialectal and Code-Switched Arabic.

In Al Khatib, Schütze, and Kantner, 2012, the authors use a set of parallel Arabic and English Wikipedia articles about Arab and Israeli public figures to explore the differences in points of view between Arabic and English articles about each figure. They employ n-gram features within a Maximum Entropy classification framework to classify whether each sentence in the given set of articles is positive, negative, or neutral. They evaluate their results against the majority baseline that yields an $F_{\beta=1}$ score of 20.6% and 21.4% on the English and Arabic datasets, respectively, and achieve a classification $F_{\beta=1}$ score of 53.3% and 47.4% on the two sets. The authors then use the estimated classes to assign a point of view score to each article and compare the differences in points of view

across both languages. Similar to Abu-Jbara et al., 2013's work, this work addresses only Standard–and not Dialectal–Arabic.

Borge-Holthoefer et al., 2015 use a set of heuristics to curate tweets that are relevant to Egyptian politics and that were posted between June and September 2013. The authors manually annotate a subset of 1,000 tweets as either supporting, opposing or being neutral towards the Military rule. The dataset is then utilized within an SVM classification framework that relies on n-gram and hashtag features to classify each given tweet as belonging to one of the three class labels. The classifier achieves a classification accuracy of 87% beating the 54% majority baseline. As opposed to our work, this work focuses on a much shorter time-frame during which most people were polarized between either supporting the Military or supporting the Muslim Brotherhood (Islamists). Other political entities such as January 25^{th} Revolution, Mubarak's regime, etc. were not a subject of discussion during this time-frame, which made the task and the annotation process less challenging.

Siegel, 2014 studies a related topic. The author comes up with a set of hypotheses that aim to identify whether Egyptian twitter users who are exposed to a more diverse twitter network become more tolerant towards people having different political and ideological leanings. Based on a set of heuristics, the author classifies Egyptian twitter users as either supporting Secularists or supporting Islamists and studies the impact of the diversity of each users' network on a person's position towards the civil liberties of the opposing group. The findings from the study suggest that there is a direct relation between the diversity of one's twitter network and political tolerance among Egyptian twitter users.

To the best of our knowledge, the work presented in this thesis is the first one to address the problem of automatic perspective identification from a multilingual point of

view in such level of detail while addressing the challenges imposed by both the level of informality present in the studied genres and the absence of a formal taxonomy of the community perspectives in the Arab world.

Chapter 7

Discussion, Conclusions and Future Directions

In this chapter we summarize our findings from our work on both Arabic and English, summarize our research contributions, and identify the limitations and directions for future research.

7.1 Discussion

In this thesis, we addressed the problem of automatically identifying a person's perspective from written text in both English and Arabic. In English, we developed systems for inferring the stance of a person on various ideological issues. We evaluated our approach on different informal genres including discussion fora, tweets, as well as a newly created corpus based on American National Election Studies. In Arabic, we specifically targeted Egyptian discussion fora. In doing so, we developed a system for automatically handling code-switching between variants of Arabic. Additionally, we developed a taxonomy for the major community perspectives in Egypt, collected large-scale linguistic annotations and explored how code-switching can be utilized with other linguistic cues to help us identify the perspective from which a given comment was written. To gain a better understanding of both the similarities and differences between studied languages, as well as how one can extend the work to new languages, we summarized our main findings and insights from this thesis.

- Our first finding is that there is a difference in the notion of perspective in both studied languages. This seems to be attributed to cultural differences. Identifying

ideological perspectives in politically established environments tends to be easier as the correlation and association between the underlying perspective and the stance on issues is less dynamic/more stable. For the Arab world and Egypt in particular, prior to Jan. 25^{th} Rev., most of the populace were politically apathetic, hence there was no established association between perspectives and positions on various political issues leading to a more dynamic situation where not only are the stances shifting but even perspectives are emerging. This presented a significant challenge for annotating the data and characterizing the phenomena under study. Accordingly, in the English datasets, people's perspectives are less dynamic. Hence, despite the long time-frame (2009 to 2015) covered by the datasets we used in this thesis, and apart from the change in politicians and public figures, the discussed polarizing topics did not change across time. Partisanship and issues of importance to the *Conservative-Liberal* continuum remained the same. The situation is very different in Arabic, where not only the topics but how various ideologies align with these topics was constantly changing. This made the task of identifying people's perspective in Arabic–whether through manual annotation or through computational systems–more challenging. The focus of the public changed over time, which was apparent in the abundance of comments that are neutral towards the different political entities depending on whether each entity was in focus as the subject of current public attention, i.e. at the time the comment was posted. Nevertheless, we think that our taxonomy of political leanings and the annotation guidelines created in this thesis represent a step in the right direction towards defining a more general notion of perspective in such a dynamic setting. The very high inter-annotator agreement, as well as the high performance of our computational systems, indicate the robustness of our taxonomy and annotation guidelines.

- While the main focus of people changed across time, more than >96% of the

comments that discuss Egyptian politics express one priority, such as stability, political reform, and the role of Islam in politics among others. This further confirms Converse's idea that within a belief system idea-elements vary in centrality (Converse, 1962). The priority expressed by each comment in our Egyptian dataset can be thought of as the most central element in the belief system of the author.

- Another finding is that regardless of the language, with the correct level of tuning, lexical features are the most important cue for successfully identifying ideological perspective. In all of our experiments, removing lexical features resulted in a significant drop in performance of our systems. This shows how well perspective is expressed in one's lexical choice. However, using the correct level of preprocessing–especially in Arabic–is crucial for these features to be fully exploited.

- We also find that while semantic features help in identifying a person's perspective, there is no "one setup fits all" set of semantic features that works on all datasets. The optimal set of features varies across domains, genres, and levels of formality, among other factors.

- In informal Arabic, code-switch detection can help in identifying a person's perspective. Due to the lexical overlap between Standard and Dialectal Arabic, identifying the correct class of the word becomes a sense disambiguation process where, for a given word, tagging a word with its intended language class specifies whether the intended meaning is the Dialectal or the Standard one.

- Utilizing non-linguistic cues can simplify the task of identifying the leaning of a comment for both computational systems and human annotators. This was apparent in our Egyptian dataset, where we found that specifying what event a comment

discusses resulted in a more positive feedback from the annotators, and in a much better performance across all computational tasks.

- Finally, having a good understanding of the political and ideological settings in a given society, and of the different challenges posed by the studied language, is invaluable when trying to identify the different perspectives, whether manually or computationally.

7.2 Summary of Contributions

In this thesis, we addressed a number of challenges. Following is a summary of the research contributions of this thesis along with the limitations and directions for future research.

The first challenge we addressed in our research is automatically handling code-switching between Standard and Dialectal Arabic. The following is a list of contributions to solve this challenge:

- **Creating Annotation Guidelines:** We created annotation guidelines to detect token-level code-switching in Arabic. For a given sentence, the guidelines address the problem of how to identify the class of each word in that sentence.

- **Collecting Token-Level Annotations:** Using the proposed guidelines to annotate a corpus that is rich in Dialectal Arabic (DA) with frequent code-switching to Modern Standard Arabic.

- **Computationally Handling Token-Level Dialectal Arabic Identification:** We built a state-of-the-art system, AIDA, for performing automatic token-level identification of Dialectal Arabic in a given Arabic text. AIDA was evaluated intrinsically on the dataset released for the shared task at EMNLP 2014 code-switching workshop

(Solorio et al., 2014) and outperformed all participating systems on the token-level classification task (Elfardy, Al-Badrashiny, and Diab, 2014b). To the best of our knowledge, AIDA is the first system to solve token-level code-switching between Modern Standard and Dialectal Arabic in a given text.

- **Computationally Handling Sentence-Level Dialectal Arabic Identification:** We extended our token-level Dialectal Arabic identification system, AIDA, to identify whether a given sentence is predominately MSA or EDA. The sentence level component was evaluated intrinsically on a standard dataset (Al-Badrashiny, Elfardy, and Diab, 2015; Elfardy and Diab, 2013) and it outperforms all baselines. Both components of the system were used to improve the quality of Dialectal Arabic to English statistical machine translation (Aminian, Ghoneim, and Diab, 2014; Salloum et al., 2014), as well as to help in identifying uncertainty cues in Arabic tweets (Al-Sabbagh, Diesner, and Girju, 2013).

The second challenge is the absence of a taxonomy for the most common perspectives among Egyptians, and the lack of annotated corpora for Egyptian Ideological Perspective. Our contributions for solving this challenge are:

- **Developing a taxonomy and annotation guidelines:** We developed a taxonomy for the most common community perspectives among Egyptians based on current political science efforts in classifying the political movements in Egypt after the Arab Spring (*Carnegie Endowment for International Peace*; Carothers and Brown, 2012; *The Hariri Center at the Atlantic Council*). We then used the proposed taxonomy and an iterative feedback-loop process to devise guidelines on how to successfully annotate a given online discussion forum post with different elements of a person's perspective (Elfardy and Diab, 2016a). To the best of our knowledge, this is the first effort at creating guidelines for collecting fine-grained multidimensional

annotations of Egyptian Ideological Perspectives that try to uncover the different underlying elements of a person's belief system. Moreover, we believe that, while the presented guidelines specifically target Egyptian politics, the annotation process and insights are applicable to any emergent and shifting political setting where there is no agreed-upon taxonomy for the common community perspectives.

- **Collecting large-scale annotations:** Using the proposed taxonomy and annotation guidelines that aim to identify different elements of the comment's ideological leaning, we annotated a large set of Egyptian discussion fora posts. We found that the annotations quite accurately depict the political spectrum in Egypt after January 25^{th} Revolution. More specifically, the polarization on different political entities across the studied timeline agrees with our prior knowledge of the studied events.

The last challenge is building computational systems that can successfully identify the perspective from which a given informal text is written. The contributions for solving this challenge are:

- **Building Computational "Perspective Identification" Systems for English:** Using lexical and semantic features we built supervised computational systems that can successfully identify the stance of a person in English informal text on different topics that are determined by one's perspective such as legalization of abortion, feminist movement, gay and gun rights, in addition to being able to identify a more general notion of perspective–namely, the 2012 choice of presidential candidate. We explored the use of standard n-gram features along with other semantic features, including weighted matrix factorization, to convert the high-dimensional n-gram space into a low-dimensional topic space. We evaluated our approach on different genres including tweets, discussion fora, as well as a new dataset based on questions

drawn from the American National Election Studies surveys. We found that lexical features perform best and that the performance of different semantic features varies across the studied datasets and domains.

- **Building Computational "Perspective Identification" Systems for Arabic:** We built supervised systems for automatically identifying different elements of a person's perspective given an Egyptian discussion forum comment. We explored different levels of linguistic preprocessing to the text and found that performing morphological preprocessing by separating clitics, determiners and conjunctions from words improves the performance. This result agrees with the common knowledge that word-segmentation reduces the sparsity of the feature space hence improving the performance of Arabic Natural Language Processing systems. Moreover, we found that code-switching can be utilized as a signal in identifying a comment's perspective since it improved the performance of our system on both tuning and held-out test sets. Counter to expectations, and unlike our results on English datasets, sentiment features did not help–or hurt–the performance. Our last finding was that, similar to the annotators' feedback, adding information about what political event the comment discusses improves the quality of our results.

- **Drawing insights from our work on both English and Arabic:** We discussed the similarities and differences between both languages by analyzing the difference in the notion of perspective elements in both studied languages as well as the analyzing which linguistic devices help in identifying the perspective in each language.

7.3 Limitations and Future Directions

Despite the progress and contributions presented in this thesis, there remains some very interesting directions that we did not explore. We identify the following limitations and areas for future direction of our work:

- **Handling Other Arabic Dialects:** As previously mentioned, we are interested in exploring the problem of perspective identification in both English and Arabic. Currently, for Arabic we only handle Modern Standard Arabic and Egyptian Dialectal Arabic (EDA) and do not handle other dialects. However, most of the methods proposed can be ported to other dialects;

- **Joint Modeling of Different Perspective Elements:** It would be interesting to jointly model different perspective elements–such as the position on the role of Islam/religion in politics and Stance on political reform versus stability in the Arabic dataset–to study whether this results in a performance boost;

- **Handling Sarcasm:** People often use irony and sarcasm in informal genre, especially when discussing a polarizing topic; since we do not model sarcasm in our current work, our systems do not capture these cases. We plan on modeling it in our future work;

- **Using Neural Networks:** Neural networks have been shown to be very powerful in solving a variety of NLP tasks, such as part of speech tagging, named-entity recognition, semantic role labeling (Collobert et al., 2011), sentiment analysis (Socher et al., 2013), stance identification (Zarrella and Marsh, 2016), among others. However, in this thesis we do not use neural-network-based methods but identify it as an interesting area to expand our work in the future.

Bibliography

Abu-Jbara, Amjad, Mona Diab, Pradeep Dasigi, and Dragomir Radev (2012). "Subgroup Detection in Ideological Discussions." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Abu-Jbara, Amjad, Ben King, Mona Diab, and Dragomir R Radev (2013). "Identifying Opinion Subgroups in Arabic Online Discussions." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Al-Badrashiny, Mohamed and Mona Diab (2016). "LILI: A Simple Language Independent Approach for Language Identification." In: *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Al-Badrashiny, Mohamed, Heba Elfardy, and Mona Diab (2015). "AIDA2: A Hybrid Approach for Token and Sentence Level Dialect Identification in Arabic." In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*.

Al-Badrashiny, Mohamed, Arfath Pasha, Mona Diab, Nizar Habash, Owen Rambow, Wael Salloum, and Ramy Eskander (2016). "Split: Smart Preprocessing (quasi) Language Independent Tool." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Al Khatib, Khalid, Hinrich Schütze, and Cathleen Kantner (2012). "Automatic Detection of Point of View Differences in Wikipedia." In: *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Al-Sabbagh, Rania, Jana Diesner, and Roxana Girju (2013). "Using the Semantic-Syntactic Interface for Reliable Arabic Modality Annotation." In: *Proceedings of the Sixth Inter-*

national Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP).

Aminian, Maryam, Mahmoud Ghoneim, and Mona Diab (2014). "Handling OOV Words in Dialectal Arabic to English Machine Translation." In: *Proceedings of the Workshop on Language Technology for Closely Related Languages and Language Variants (LT4CloseLang) at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ammar, Waleed, Chris Dyer, and Noah A Smith (2014). "Conditional Random Field Autoencoders for Unsupervised Structured Prediction." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger. Curran Associates, Inc.

Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor (2011). "Cats Rule and Dogs Drool!: Classifying Stance in Online Debate." In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*. Association for Computational Linguistics.

Augenstein, Isabelle, Andreas Vlachos, and Kalina Bontcheva (2016). "USFD at SemEval-2016 Task 6: Any-Target Stance Detection on Twitter with Autoencoders." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Badaro, Gilbert, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj (2014). "A Large Scale Arabic Sentiment Lexicon for Arabic Opinion Mining." In: *Proceedings of the Workshop on Arabic Natural Language Processing (WANLP) at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Badaro, Gilbert, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban (2015). "A Light Lexicon-based Mobile Application for Sentiment Mining of Arabic Tweets." In: *Proceedings of the Workshop on Arabic Natural Language Processing (WANLP) at the Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*.

Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The Berkeley Framenet Project." In: *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Barberá, Pablo, Amber Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler (2016). "Methodological Challenges in Estimating Tone: Application to News Coverage of the US Economy." In: *Meeting of the Midwest Political Science Association, Chicago, IL*.

Benajiba, Yassine, Paolo Rosso, and Jos Miguel Benedruiz (2007). "ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy." In: *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*.

Biadsy, Fadi, Julia Hirschberg, and Nizar Habash (2009). "Spoken Arabic Dialect Identification Using Phonotactic Modeling." In: *Proceedings of the Workshop on Computational Approaches to Semitic Languages at the meeting of the European Association for Computational Linguistics (EACL)*.

Blei, David M., Andrew Y Ng, and Michael I Jordan (2003). "Latent Dirichlet Allocation." In: *the Journal of machine Learning research* 3.

Bøhler, Henrik, Petter Fagerlund Asla, Erwin Marsi, and Rune Sætre (2016). "IDI@ NTNU at SemEval-2016 Task 6: Detecting Stance in Tweets Using Shallow Features and GloVe Vectors for Word Representation." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Boltuzic, Filip, Mladen Karan, Domagoj Alagic, and Jan Šnajder (2016). "TakeLab at SemEval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Borge-Holthoefer, Javier, Walid Magdy, Kareem Darwish, and Ingmar Weber (2015). "Content and Network Dynamics behind Egyptian Political Polarization on Twitter." In:

Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, pp. 700–711.

Bouamor, Houda, Nizar Habash, and Kemal Oflazer (2014). "A Multidialectal Parallel Corpus of Arabic." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Boydstun, Amber E, Justin H Gross, Philip Resnik, and Noah A Smith (2013). "Identifying Media Frames and Frame Dynamics Within and Across Policy Issues." In: *New Directions in Analyzing Text as Data Workshop, London*.

Boydstun, Amber and Justin Gross (2014). *Policy Frames Codebook*.

Brown, Nathan J. (2013). "Egypt's Failed Transition." In: *Journal of Democracy* 24.4, pp. 45–58.

Carnegie Endowment for International Peace. `http://carnegieendowment.org/2015/01/22/2012-egyptian-parliamentary-elections/`.

Carothers, Thomas and Nathan J. Brown (2012). "The Real danger for Egyptian democracy." In: *Carnegie Article*.

Chan, Joyce Y. C., P. C. Ching, Tan LEE, and Helen M. Meng (2004). "Detection of Language Boundary in Code-Switching Utterances by Bi-Phone Probabilities." In: *Proceedings of the International Symposium on Chinese Spoken Language Processing*.

Chen, Desai, Nathan Schneider, Dipanjan Das, and Noah A Smith (2010). "SEMAFOR: Frame Argument Resolution with Log-Linear Models." In: *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Chittaranjan, Gokul, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury (2014). "Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System." In: *Proceedings of the First Workshop on Computational Approaches to Code Switching at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chong, Dennis and James N. Druckman (2007). "Framing Theory." In: *Annual Review of Political Science* 10.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011). "Natural Language Processing (almost) From Scratch." In: *Journal of Machine Learning Research* 12.Aug, pp. 2493–2537.

Converse, Philip E (1962). *The Nature of Belief Systems in Mass Publics*. Survey Research Center, University of Michigan.

Cotterell, Ryan and Chris Callison-Burch (2014). "A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Darwish, Kareem, Hassan Sajjad, and Hamdy Mubarak (2014). "Verifiably Effective Arabic Dialect Identification." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1465–1468.

Das, Dipanjan, Nathan Schneider, Desai Chen, and Noah A Smith (2010). "Probabilistic Frame-Semantic Parsing." In: *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Dasigi, Pradeep and Mona Diab (2011). "CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic." In: *Proceedings of the 5th International Joint Conference on Natural Language Processing (ICJNLP), Chiangmai, Thailand*.

Dasigi, Pradeep, Weiwei Guo, and Mona Diab (2012). "Genre Independent Subgroup Detection in Online Discussion Threads: A Pilot Study of Implicit Attitude Using Latent Textual Semantics." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D Manning (2006). "Generating Typed Dependency Parses from Phrase Structure Parses." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Diab, Mona and Ankit Kamboj (2011). "Feasibility of Leveraging Crowd Sourcing for the Creation of a Large Scale Annotated Resource for Hindi English Code Switched Data: A Pilot Annotation." In: *Proceedings of the 9th Workshop on Asian Language Resources*.

Diab, Mona, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Pradeep Dasigi, Heba Elfardy, Ramy Eskander, Nizar Habash, Abdelati Hawwari, and Wael Salloum (2014). "Tharwa: A Multi-Dialectal Multi-Lingual Machine Readable Dictionary." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Dias, Marcelo and Karin Becker (2016). "INF-UFRGS-OPINION-MINING at SemEval-2016 Task 6: Automatic Generation of a Training Corpus for Unsupervised Identification of Stance in Tweets." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Elfardy, Heba, Mohamed Al-Badrashiny, and Mona Diab (2013). "Code Switch Point Detection in Arabic." In: *Proceedings of the International Conference on Application of Natural Language to Information Systems (NLDB2013)*.

— (2014a). "A Hybrid System for Code Switch Point Detection in Informal Arabic Text." In: *XRDS: Crossroads, The ACM Magazine for Students* 21.1.

— (2014b). "AIDA: Identifying Code Switching in Informal Arabic Text." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Elfardy, Heba and Mona Diab (2012a). "AIDA: Automatic Identification and Glossing of Dialectal Arabic." In: *Proceedings of the European Association for Machine Translation Conference: System Demonstrations*.

Elfardy, Heba and Mona Diab (2012b). "Simplified Guidelines for the Creation of Large Scale Dialectal Arabic Annotations." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

— (2012c). "Token Level Identification of Linguistic Code Switching." In: *Proceedings of the International Conference on Computational Linguistics (COLING)*.

— (2013). "Sentence-Level Dialect Identification in Arabic." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

— (2016a). "Addressing Annotation Complexity: The Case of Annotating Ideological Perspective in Egyptian Social Media." In: *Proceedings of the 10th Linguistic Annotation Workshop (LAW-X) at the Annual Meeting of the Association for Computational Linguistics (ACL)*.

— (2016b). "CU-GWU Perspective at SemEval-2016 Task 6: Ideological Stance Detection in Informal Text." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Elfardy, Heba, Mona Diab, and Chris Callison-Burch (2015). "Ideological Perspective Detection Using Semantic Features." In: *Proceedings of the Conference of Lexical and Computational Semantics (*SEM 2015)*.

Elmasry, Mohamad Hamas, Alaa El Shamy, Peter Manning, Andrew Mills, and Philip J Auter (2013). "Al-Jazeera and Al-Arabiya Framing of the Israel–Palestine conflict during War and Calm Periods." In: *International Communication Gazette*.

Elmasry, Mohamad (2009). "Death in the Middle East: An analysis of How the New York Times and Chicago Tribune Framed Killings in the Second Palestinian Intifada." In: *Journal of Middle East Media* 5.1, pp. 1–46.

Entman, Robert M. (1993). "Framing: Toward Clarification of a Fractured Paradigm." In: vol. 43. 4. Wiley Online Library.

Eskander, Ramy, Nizar Habash, Owen Rambow, and Nadi Tomeh (2013). "Processing Spontaneous Orthography." In: *Proceedings of the Meeting of the the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Eskander, Ramy, Mohamed Al-Badrashiny, Nizar Habash, and Owen Rambow (2014). "Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script." In: *Proceedings of the First Workshop on Computational Approaches to Code-Switching at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Wiley Online Library.

Ferguson, Charles A. (1959). *Diglossia. Word 15. 325340*. DOI: 10.1080/00437956.1959.11659702.

Fillmore, Charles J (2006). "Frame Semantics." In: *Cognitive Linguistics: Basic Readings* 34.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). "Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling." In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*.

Forney G.D., Jr. (1973). "The Viterbi Algorithm." In: *IEEE* 61.3, pp. 268–278. ISSN: 0018-9219. DOI: 10.1109/PROC.1973.9030.

Gentzkow, Matthew and Jesse M Shapiro (2006). "Media Bias and Reputation." In: *Journal of political Economy* 114.2, pp. 280–316.

Ghosh, Debanjan, Weiwei Guo, and Smaranda Muresan (2015). "Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Greene, Stephan and Philip Resnik (2009). "More than Words: Syntactic Packaging and Implicit Sentiment." In: *Proceedings of the Meeting of the North American Chapter of*

the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT).

Guo, Weiwei and Mona Diab (2012). "Modeling Sentences in the Latent Space." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Habash, Nizar (2010). "Introduction to Arabic Natural Language Processing." In: *Synthesis Lectures on Human Language Technologies* 3.1, pp. 1–187.

Habash, Nizar, Ramy Eskander, and AbdelAti Hawwari (2012). "A Morphological Analyzer for Egyptian Arabic." In: *Proceedings of the Workshop on Computational Morphology and Phonology (SIGMORPHON2012) at the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Habash, Nizar and Fatiha Sadat (2006). "Arabic Preprocessing Schemes for Statistical Machine Translation." In: *Proceedings of the Meeting of the the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten (2009). "The WEKA Data Mining Software: An Update." In: *ACM SIGKDD Explorations Newsletter* 11.1.

Hasan, Kazi Saidul and Vincent Ng (2012). "Predicting Stance in Ideological Debate with Rich Linguistic Knowledge." In: *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

— (2013). "Extra-Linguistic Constraints on Stance Recognition in Ideological Debates." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Igarashi, Yuki, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui (2016). "Tohoku at SemEval-2016 Task 6: Feature-based Model versus Convolutional Neural Network for Stance Detection." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Iyyer, Mohit, Peter Enns, Jordan Boyd-Graber, and Philip Resnik (2014). "Political Ideology Detection Using Recursive Neural Networks." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jain, Naman and Ahmad Riyaz Bhat (2014). "Proceedings of the First Workshop on Computational Approaches to Code Switching at the Conference on Empirical Methods in Natural Language Processing (EMNLP)." In: chap. Language Identification in Code-Switching Scenario.

Joshi, Aravind K (1982). "Processing of Sentences with Intra-Sentential Code-Switching." In: *Proceedings of the 9th Conference on Computational Linguistics (COLING)*.

King, Levi, Eric Baucom, Timur Gilmanov, Sandra Kübler, Daniel Whyatt, Wolfgang Maier, and Paul Rodrigues (2014). "The IUCL+ System: Word-Level Language Identification via Extended Markov Models." In: *Proceedings of the First Workshop on Computational Approaches to Code-Switching at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Klebanov, Beata Beigman, Eyal Beigman, and Daniel Diermeier (2010). "Vocabulary choice as an indicator of perspective." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith (2014). "A Dependency Parser for Tweets." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Krejzl, Peter and Josef Steinberger (2016). "UWB at SemEval-2016 Task 6: Stance Detection." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Lin, Chu-Cheng, Waleed Ammar, Lori Levin, and Chris Dyer (2014). "The CMU Submission for the Shared Task on Language Identification in Code-Switched Data." In: *Proceedings of the First Workshop on Computational Approaches to Code-Switching at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lin, Wei-Hao, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann (2006). "Which Side are You on?: Identifying Perspectives at the Document and Sentence Levels." In: *Proceedings of the Conference on Computational Natural Language Learning*.

Liu, Can, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler (2016). "IUCL at SemEval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Maamouri, Mohamed, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick (2010). *LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1*.

Manandise, Esm and Claudia Gdaniec (2011). "Morphology to the Rescue Redux: Resolving Borrowings and Code-Mixing in Machine Translation." In: *SFCM'11*, pp. 86–97.

Mansour, Essam (2012). "The Role of Social Networking Sites (SNSs) in the January 25th Revolution in Egypt." In: *Library Review* 61.2, pp. 128–159.

Misra, Amita, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker (2016). "NLDS-UCSC at SemEval-2016 Task 6: A Semi-Supervised Approach to Detecting Stance in Tweets." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Mohammad, Saif M., Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry (2016). "SemEval-2016 Task 6: Detecting Stance in Tweets." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Molina, Giovanni, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio (2016). "Overview for the Second Shared Task on Language Identification in Code-Switched Data." In: *Proceedings of the Second Workshop on Computational Approaches to Linguistic Code Switching (CALCS) at the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn (2008). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." In: *Political Analysis* 16.4, pp. 372–403.

Nguyen, Viet-An, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler (2015). "Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in t." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth (2014). "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic." In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Patra, Braja Gopal, Dipankar Das, and Sivaji Bandyopadhyay (2016). "JU NLP at SemEval-2016 Task 6: Detecting Stance in Tweets using Support Vector Machines." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen (2005). "SenseRelate:: TargetWord: A Generalized Framework for Word Sense Disambiguation." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi (2004). "WordNet:: Similarity: Measuring the Relatedness of Concepts." In: *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Poole, Keith T and Howard Rosenthal (1985). "A Spatial Model for Legislative Roll Call Analysis." In: *American Journal of Political Science*, pp. 357–384.

Rashwan, Mohsen, Mohamed Al-Badrashiny, Mohamed Attia, Sherif M Abdou, and Ahmed Rafea (2011). "A Stochastic Arabic Diacritizer Based on a Hybrid of Factorized and Unfactorized Textual Features." In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.1, pp. 166–175. ISSN: 1558-7916.

Salloum, Wael and Nizar Habash (2011). "Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation." In: *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*.

Salloum, Wael, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab (2014). "Sentence Level Dialect Identification for Machine Translation System Selection." In: *Proceedings of the annual meeting of the Association for Computational Linguistics (ACL)*.

Sha, Fei and Fernando Pereira (2003). "Shallow Parsing with Conditional Random Fields." In: *the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Siegel, Alexandra (2014). "Tweeting Beyond Tahrir: Ideological Diversity and Political Tolerance in Egyptian Twitter Networks." In: *Unpublished working paper, New York University*.

Socher, Richard, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts (2013). "Recursive Deep Models for Semantic Com-

positionality over a Sentiment Treebank." In: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vol. 1631. Citeseer.

Solorio, Thamar and Yang Liu (2008a). "Learning to predict code-switching points." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

— (2008b). "Part-of-Speech Tagging for English-Spanish Code-Switched Text." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steve Bethard, Mona Diab, Mahmoud Gonheim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirshberg, Alison Chang, and Pascale Fung (2014). "Overview for the First Shared Task on Language Identification in Code-Switched Data." In: *Proceedings of the First Workshop on Computational Approaches to Code-Switching at the Conference on Empirical Methods in Natural Language Processing. (EMNLP)*.

Somasundaran, Swapna and Janyce Wiebe (2010). "Recognizing Stances in Ideological Online Debates." In: *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text at the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Stolcke, Andreas (2002). "SRILM an Extensible Language Modeling Toolkit." In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Tausczik, Yla R and James W Pennebaker (2010). "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." In: *Journal of language and social psychology* 29.1.

The Hariri Center at the Atlantic Council. `http://www.atlanticcouncil.org/blogs/egyptsource/egyptian-politics`.

Van Dijk, Teun A (1998). *Ideology: A Multidisciplinary Approach*. Sage.

Vijayaraghavan, Prashanth, Ivan Sysoev, Soroush Vosoughi, and Deb Roy (2016). "DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Volkova, Svitlana, Glen Coppersmith, and Benjamin Van Durme (2014). "Inferring User Political Preferences from Streaming Communications." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wei, Wan, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang (2016). "pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Wojatzki, Michael and Torsten Zesch (2016). "ltl. uni-due at SemEval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Yano, Tae, Philip Resnik, and Noah A Smith (2010). "Shedding (a thousand points of) Light on Biased Language." In: *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk at the Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (NAACL-HLT)*.

Zaidan, Omar F and Chris Callison-Burch (2011). "The Arabic Online Commentary Dataset: An Annotated Dataset of Informal Arabic with High Dialectal Content." In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zarrella, Guido and Amy Marsh (2016). "MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection." In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval '16.

Zhang, Zhihua and Man Lan (2016). “ECNU at SemEval-2016 Task 6: Relevant or Not? Supportive or Not? A Two-step Learning System for Automatic Detecting Stance in Tweets.” In: *Proceedings of the International Workshop on Semantic Evaluation*. SemEval ’16.